

Hruschka, Harald

Article — Published Version

Relevance of dynamic variables in multicategory choice models

OR Spectrum

Provided in Cooperation with:

Springer Nature

Suggested Citation: Hruschka, Harald (2022) : Relevance of dynamic variables in multicategory choice models, OR Spectrum, ISSN 1436-6304, Springer, Berlin, Heidelberg, Vol. 46, Iss. 1, pp. 109-133, <https://doi.org/10.1007/s00291-022-00690-z>

This Version is available at:

<https://hdl.handle.net/10419/307272>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



<https://creativecommons.org/licenses/by/4.0/>



Relevance of dynamic variables in multicategory choice models

Harald Hruschka¹ 

Received: 30 November 2021 / Accepted: 18 August 2022 / Published online: 11 September 2022
© The Author(s) 2022

Abstract

We investigate the relevance of dynamic variables that reflect the purchase history of a household as independent variables in multicategory choice models. To this end, we estimate both homogeneous and finite mixture variants of the multivariate logit model. We consider two types of dynamic variables. Variables of the first type, which previous publications on multicategory choice models have ignored, are exponentially smoothed category purchases, which we simply call category loyalties. Variables of the second type are log-transformed times since the last purchase of any category. Our results clearly show that adding dynamic variables improves statistical model performance with category loyalties being more important than log-transformed times. The majority of coefficients of marketing variables (features, displays, and price reductions), pairwise category interactions, and cross-category relations differ between models either including or excluding dynamic variables. We also measure the effect of marketing variables on purchase probabilities of the same category (own effects) and on purchase probabilities of other categories (cross effects). This exercise demonstrates that the model without dynamic variables tends to overestimate own effects of marketing variables in many product categories. This positive omitted variable bias provides another explanation for the well-known problem of “overpromotion” in retailing.

Keywords Multicategory choice · Market basket analysis · Multivariate logit model · Purchase history

✉ Harald Hruschka
harald.hruschka@ur.de

¹ Faculty of Business, Economics and Management Information Systems, University of Regensburg, 93040 Regensburg, Germany

1 Introduction

It might seem obvious that the probability to purchase a product or brand depends on previous purchases. Appropriate econometric approaches start from a static model that they enlarge by dynamic variables that reflect the purchase history of a household (Meyer et al. 2017). The best-known example of such a dynamic variable is the exponentially smoothed measure introduced by Guadagni and Little Guadagni and Little (1983), which these authors call brand loyalty, added to a static multinomial logit brand choice model. Most brand choice models include this dynamic variable (Chiang 1991; Chintagunta 1993).

The situation for multicategory choice models, of which multivariate logit (MVL) and multivariate probit (MVP) models are the dominant functional forms, turns out to be completely different. MVP models as a rule do not include any dynamic variable at all. Several MVL models consider one dynamic purchase variable, log-transformed time since the last purchase of a category. On the other hand, these MVL models exclude exponentially smoothed category loyalties.

In contrast to the previous literature, we add exponentially smoothed category purchases, which we simply call category loyalties in the following, to the predictors. We investigate how category loyalties improve statistical performance compared to log-transformed time since the previous purchase. Our results clearly show that category loyalties improve performance more than log-transformed times. Nonetheless, keeping log-transformed times as predictors in addition to category loyalties leads to further improvements.

Homogeneous models may overestimate the effects of dynamic variables because they ignore that households may have different category preferences that are unrelated to previous purchases (Keane 1997). Therefore we also investigate finite mixture extensions of the MVL model (FM-MVL), which by allowing for heterogeneous preferences avoid this weakness.

We do not apply the MVL model with continuous heterogeneity because of its higher computational complexity, which explains why publications using this type of MVL model do not consider more than six categories (Gentzkow 2007; Kwak et al. 2015; Richards et al. 2018). In addition, related econometric models (multinomial logit and Tobit) with finite heterogeneity have been shown to outperform their continuous counterparts (Andrews et al. 2002; Ansari and Mela 2003; Schröder and Hruschka 2017).

Multicategory choice models include independent variables, e.g., marketing variables. Our research focuses on measuring the effects of marketing variables. We do not consider machine learning algorithms such as associations rules (Agrawal and Srikant 1994; Hahsler et al. 2006) or topic models (Hruschka 2014), because they usually exclude independent variables.

Based on their global estimation performance, we compare two FM-MVL models in more detail. The basic FM-MVL model includes marketing variables as independent variables. Independent variables of the enlarged FM-MVL model consist of both marketing variables and dynamic variables. We investigate whether these two models differ with respect to coefficients and cross-category dependences. We also

show that managerial implications for these two models differ. To this end, we measure the effect of marketing variables on purchase probabilities of the same category as well as on purchase probabilities of other categories. Our results demonstrate that the model without dynamic variables tends to overestimate own effects of marketing variables in many product categories. This positive omitted variable bias provides another explanation for the well-known problem of “overpromotion” in retailing.

2 Investigated dynamic purchase variables

This section is based on a thorough search of the literature in which probabilistic models, to which multivariate probit or logit models belong, serve to analyze multicategory choices, up to and including 2022. As already mentioned in the introduction, papers applying multivariate probit models do not include any dynamic purchase variable (Chib et al. 2002; Duvvuri et al. 2007; Manchanda et al. 1999; Hruschka 2013, 2017a, b, c).

Several recent relevant publications use probabilistic models with latent variables. Topic models replace choices of several product categories by a lower number of latent variables. Hruschka investigates two topic models, latent Dirichlet allocation and the correlated topic model, which do not include independent variables (Hruschka 2014). Jacobs et al. extend latent Dirichlet allocation to consider one independent variable (time of the first order at the retailer’s website), which is constant for each household (Jacobs et al. 2016). The probabilistic model of Ruiz et al. comprises three types of latent variables that compress purchases, prices and seasonal effects of products, respectively (Ruiz et al. 2020). These authors specify latent variables without dynamic effects. To summarize, publications on probabilistic models with latent variables ignore dynamic purchase variables just like publications on the multivariate probit model.

We can distinguish two groups of papers that apply variants of the MVL model. One group ignores dynamic purchase variables (Dippold and Hruschka 2013; Kwak et al. 2015; Richards et al. 2018). Papers of the other group consider one dynamic purchase variable, log-transformed time since the last purchase of the corresponding category, as one of the independent variables (Russell and Petersen 2000; Boztuğ and Hildebrandt 2008; Boztuğ and Reutterer 2008; Solnet et al. 2016). However, papers of this group do not investigate the importance of this dynamic purchase variable relative to other independent variables, e.g., marketing variables.

To log-transformed times since the last purchase, we add exponentially smoothed category loyalties in analogy to the exponentially smoothed brand loyalties, which are widespread in brand choice models. What authors of several MVL papers call loyalty in fact measures the long run propensity of a household to make a category purchase (Russell and Petersen (2000); Boztuğ and Hildebrandt (2008); Boztuğ and Reutterer (2008); Solnet et al. (2016)). Consequently, this variable is not dynamic and does not change across purchases of the same household.

We compute the loyalty of household m for category j at shopping visit t as follows:

$$\text{loy}_{jmt} = \alpha y_{jmt-1} + (1 - \alpha) \text{loy}_{jmt-1} \quad (1)$$

$0 \leq \alpha \leq 1$ denotes the smoothing constant. The binary purchase incidence y_{jmt-1} equals one, if household m purchases category j at the previous shopping trip $t - 1$. 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_{jmt0} 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 smoothing constant α is, the more it smooths purchases of the past. This smoothing distinguishes category loyalties from log-transformed times, which may largely fluctuate between shopping. We measure the importance of these two dynamic purchase variables relative to each other and to the other independent variables.

As higher category loyalties increase purchase probabilities, we expect positive response coefficients. The situation is less clear-cut for log-transformed times, though coefficients are positive according to the majority of studies (Russell and Petersen 2000; Boztuğ and Hildebrandt 2008; Solnet et al. 2016).

We also investigate whether the effects of marketing variables suffer from an omitted variable bias (Wooldridge 2013) due to ignoring dynamic purchasing variables. If the correlation between a dynamic variable and a marketing variable is positive and the effect of the dynamic variable on purchase probability is positive, a positive bias results, i.e., a model that excludes the dynamic variable overestimates the effect of the marketing variable.

3 Investigated model variants

In this section, we present the two investigated variants of the MVL model, the homogeneous MVL model and its finite mixture extension. 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 at purchase occasion t , the respective element y_{jmt} equals one. Vector x_{mt} consists of independent variables relevant for market basket t of household m .

3.1 Homogeneous multivariate logit model

In the homogeneous MVL model, each coefficient is constant across households. Extending the expression for the homogeneous MVL model without independent variables (also known as auto-logistic model) given in Besag (1972) we define the probability of market basket y_{mt} conditional on independent variables x_{mt} as follows:

$$\begin{aligned} & \exp(y'_{mt}a + x'_{mt}b y_{mt} + 1/2 y'_{mt} V y_{mt})/C \\ & \text{with } C = \sum_{v \in \{0,1\}^J} \exp(v' a + x'_{mt} b v + 1/2 v' V v) \end{aligned} \quad (2)$$

Expression (2) shows that computation of this probability requires division by the so-called normalization constant C that is obtained by summing over all possible market baskets defined by different binary vectors v . Coefficients contained in (J, J) matrix V measure pairwise interactions between categories. As a pairwise interaction of a category with itself does not make sense, all diagonal elements of V are zero. Off-diagonal elements are symmetric, i.e., $V_{j1,j2} = V_{j2,j1}$. Column vector a consists of J constants. The (L, J) matrix b holds the effect of L independent variables on purchase probabilities. The homogeneous MVL model has been applied to market basket data by Russell and Petersen (2000) building upon earlier publications in statistics (Cox 1972; Besag 1974).

For the homogeneous MVL model, 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_{-jmt} and the independent variables x_{mt} as:

$$P(y_{jmt} = 1 | y_{-jmt}, x_{mt}) = \varphi \left(a_j + b_j x_{mt} + \sum_{l \neq j} V_{jl} y_{lmt} \right) \quad (3)$$

φ denotes the binomial logistic function $1/(1 + \exp(-Z))$. We obtain the independent logit model that excludes interactions between categories by setting all coefficients in V equal to zero.

3.2 Finite mixture multivariate logit model

We also investigate the finite mixture extension of the MVL model (FM-MVL). Coefficients of the FM-MVL model differ between household segments. The purchase probability of category j in market basket t of household m conditional on purchases of the other categories and the independent variables x_{mt} is:

$$\begin{aligned} P(y_{jmt} = 1 | y_{-jmt}, x_{mt}) &= \sum_{s=1}^S u_{sm} P_s(y_{jmt} = 1 | y_{-jmt}) \\ &= \sum_{s=1}^S u_{sm} \varphi \left(a_{sj} + b_{s,j} x_{mt} + \sum_{l \neq j} V_{sj,l} y_{lmt} \right) \end{aligned} \quad (4)$$

S denotes the number of segments. u_{sm} is a binary membership indicator set to one if household m is assigned to segment s . P_s is the segment-specific conditional probability function.

4 Model estimation and evaluation

We exclude the null basket for which all purchase indicators y_j equal zero in accordance with previous related publications (Russell and Petersen 2000; Boztuğ and Reutterer 2008; Kwak et al. 2015). This way we model purchases conditional on the

purchase of at least one category. Therefore, the number of possible market baskets is $2^J - 1$.

Maximum likelihood estimation of the MVL model requires computation of the so-called normalization constant obtained by summing over all possible market baskets (see expression (2)) in each iteration. For 31 categories we would have to deal with more than 2.14×10^9 possible market baskets. Because of the impracticality of this approach, we resort to maximum pseudo-likelihood (MPL) estimation. In a simulation study Bel et al. (2018) compare MPL to maximum likelihood estimation for a maximum number of 12 alternatives. These authors conclude that MPL estimation leads to negligible efficiency losses only.

The pseudo-probability \tilde{P}_{jmt} of category j in market basket t of household m can be written for both the homogeneous MVL model and the FM-MVL model as:

$$\tilde{P}_{jmt} = P(y_{jmt} = 1 | y_{-jmt}, x_{mt})^{y_{jmt}} (1 - P(y_{jmt} = 1 | y_{-jmt}, x_{mt}))^{1-y_{jmt}} \quad (5)$$

Expressions (3) and (4) show how to compute the conditional probability $P(y_{jmt} = 1 | y_{-jmt}, x_{mt})$ for the homogeneous MVL model and the FM-MVL model, respectively. 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 equation (5) that its first part is relevant if category j is purchased and its second part if category j is not purchased.

MPL estimation consists in maximizing the log pseudo-likelihood LPL across households, market baskets and categories:

$$LPL = \sum_{m=1}^M \sum_{t=1}^{T_m} \sum_{j=1}^J \log(\tilde{P}_{jmt}) \quad (6)$$

T_m symbolizes the number of market baskets of household m . Due to the binary membership indicators given in expression (4) the same segment-specific conditional probability function is used for all market baskets of any household m . Equation (6) shows that the computation of log pseudo-likelihood values requires to sum across J logarithmic conditional probabilities. Summing across product categories makes MPL estimation feasible because it replaces summing across all possible baskets that would be necessary in ML estimation. In the case of a MVL model with pairwise interactions each of the J logarithmic conditional probabilities is related to the purchase incidences of the other $J - 1$ categories.

Estimation of the homogeneous MVL model turns out to be straightforward because LPL has only one local maximum. On the contrary, for the FM-MVL model LPL may have multiple local maxima. That is why we start estimation of the FM-MVL models ten times by randomly assigning each household to one of S segments. Our estimation approach for the FM-MVL model is akin to maximizing the classification likelihood (McLachlan and Basford 1988; Ngatchou-Wandji and Bulla 2013) replacing the intractable likelihood by segment-specific pseudo-probabilities.

We evaluate models by their log pseudo-likelihood on holdout data. This way we consider the complexity of models. A model, whose complexity is too high, leads to a worse (lower) log pseudo-likelihood value for the holdout data. In contrast to

information criteria such as AIC or BIC, holdout validation has the advantage to do without assumptions about the true underlying model.

We randomly form two groups with about 2/3 of the households in the first group. We use data (estimation data) of the first group to estimate models. Data of the second group (holdout data) serve to evaluate models whose coefficients we estimate on data from the first group. We also use holdout log pseudo-likelihood values to decide on the number of segments S for each of the considered FM-MVL models. We select the model with S segments if both the model with a lower number of segments $S - 1$ and the model with a higher number of segments $S + 1$ attain lower holdout log pseudo-likelihood values. For our data this procedure leads to an unambiguous determination of the number of segments.

To make comparison of model performances easier, we also compute $IAPP$, the increase of the average pseudo-probability of the respective model over the average pseudo-probability of the least complex model, which is homogeneous and excludes both interactions and independent variables. Using the log pseudo-likelihood of the respective model LPL , the log pseudo-likelihood of the least complex model LPL_0 and the total number of purchase visits across households n_v , we determine this increase as follows:

$$IAPP = \exp(LPL/n_v) / \exp(LPL_0/n_v) - 1 = \exp((LPL - LPL_0)/n_v) - 1. \quad (7)$$

This expression shows that we define average pseudo-probabilities as geometric means, i.e., as $\exp(LPL/n_v)$ and $\exp(LPL_0/n_v)$, respectively. $IAPP$ can be seen as measure of relative model performance. $IAPP$ values are positive, if the average pseudo-probability of the respective model is greater than the average pseudo-probability of the least complex model. $IAPP$ values are zero (negative), if the average pseudo-probability of the respective model equals (is lower than) the average pseudo-probability of the least complex model.

5 Model comparisons

We compare the best performing model without dynamic variables $M0$ to the best performing model with dynamic variables $M1$. To this end, we test whether average category constants and average coefficients differ between these two models. We also examine whether models $M0$ and $M1$ lead to different results on the dependencies between categories.

Average category constants and average coefficients are determined by weighting segment-specific constants or coefficients by relative segment sizes. We determine the significance of a difference by means of the following t-statistic:

$$t_stat = \frac{\sum_{s=1}^{S(M0)} \pi_{s(M0)} \gamma_{s(M0)} - \sum_{s=1}^{S(M1)} \pi_{s(M1)} \gamma_{s(M1)}}{\sqrt{\sum_{s=1}^{S(M1)} \pi_{s(M1)} \sigma_{s(M1)}^2}} \quad (8)$$

$S(M0)$ and $S(M1)$ denote the number of segments according to $M0$ and $M1$, respectively. $\pi_{s(M0)}$, $\pi_{s(M1)}$ symbolize the relative size of segment s for $M0$ and $M1$,

respectively. $\gamma_{s(M0)}, \gamma_{s(M1)}$ is a category constant or coefficient of segment s for $M0$ and $M1$, respectively. $\sigma_{s(M1)}$ denotes the standard error of the constant or coefficient of segment s for $M1$.

We measure the relation of a category j conditional on another category j' by the average marginal effect with respect to the purchase pseudo-probability of category j . We classify two categories as purchase complements if the average marginal effect is positive and as purchase substitutes if the average marginal effect is negative. This definition is analogous to the one put forward by Betancourt and Gautschi (1990), who consider two products as purchase complements (purchase substitutes) if they are purchased jointly more (less) frequently than expected under stochastic independence.

The average marginal effect corresponds to the difference of the average pseudo-probability of a purchase of category j given a purchase of category j' and the average pseudo-probability of a purchase of category j given a non-purchase of category j' (Greene 2003). Note that we average across baskets by keeping the observed values of independent variables and the observed purchase incidences of categories other than j and j' .

We can write the average marginal effect for segment s of model $M0$ or $M1$ as:

$$\begin{aligned} \text{ame}(j, j')_{s(M0)} &= \tilde{P}_{s(M0)}(y_j = 1 | y_{j'} = 1) - \tilde{P}_{s(M0)}(y_j = 1 | y_{j'} = 0) \\ \text{or} \\ \text{ame}(j, j')_{s(M1)} &= \tilde{P}_{s(M1)}(y_j = 1 | y_{j'} = 1) - \tilde{P}_{s(M1)}(y_j = 1 | y_{j'} = 0) \end{aligned} \quad (9)$$

$\tilde{P}_{s(M0)}, \tilde{P}_{s(M1)}$ denote pseudo-probabilities averaged across baskets, for segment s of model $M0$ and $M1$, respectively.

The standard error of the average marginal effect for segment s of model $M1$ is Greene (2003):

$$\begin{aligned} \sigma(j, j')_{s(M1)} &= \tilde{P}_{s, M1}(y_j = 1 | y_{j'} = 1)(1 - \tilde{P}_{s, M1}(y_j = 1 | y_{j'} = 1)) \\ &\quad - \tilde{P}_{s, M1}(y_j = 1 | y_{j'} = 0)(1 - \tilde{P}_{s, M1}(y_j = 1 | y_{j'} = 0)) \end{aligned} \quad (10)$$

Finally, we compute the t-statistic of the difference of an average marginal effect between models $M0$ and $M1$ as follows:

$$t_stat = \frac{\sum_{s=1}^{S(M0)} \pi_{s(M0)} \text{ame}(j, j')_{s(M0)} - \sum_{s=1}^{S(M1)} \pi_{s(M1)} \text{ame}(j, j')_{s(M1)}}{\sqrt{\sum_{s=1}^{S(M1)} \pi_{s(M1)} \sigma(j, j')_{s(M1)}^2}} \quad (11)$$

6 Derivation of managerial implications

We investigate whether the two models $M0$ and $M1$ lead to different managerial implications. We consider the decision problem of choosing the category to be promoted by, e.g., a price cut, a feature, or a display. This decision depends on the effect a promotion has on purchases of the promoted category itself, the so-called

own effect, as well as on cross effects, i.e., the effects on purchases of other categories. Note that we use a broad definition of promotion that includes display and feature advertising activities besides price reductions (Gedenk et al. 2010).

In a first run of our simulation approach, we set all marketing variables for all categories to zero. This way we estimate total purchase probabilities if no category is promoted. Then we estimate total purchase probabilities by setting one of the three marketing variables in one category j to one. The differences of total probabilities for this constellation and the total probabilities for no promotion measure the effects of the respective marketing variable. Such a difference represents the own effect of the marketing variable if it refers to the same category j . If a difference refers to another category, we get a cross effect of the marketing variable. We apply this procedure both for model $M0$ and model $M1$, which in the following enables us to compute differences of both own and cross effects between the two models.

As we base our estimation approach on pseudo-probabilities, we cannot directly determine purchase probabilities and have to resort to simulation. For each segment s of models $M0$ and $M1$, we generate simulated purchases by iterated Gibbs-sampling from the conditional distribution (Besag 2004) given as:

$$\varphi\left(a_{sj} + b_{s,j}x + \sum_{l \neq j} V_{sj,l}y_l\right) \quad (12)$$

For model $M0$ we obtain segment-specific purchase probabilities by averaging simulated purchases and compute total purchase probabilities as averages of segment specific probabilities weighted by relative segment size. For model $M1$ the dynamic variables vary across baskets.

As computation times of Gibbs sampling for each observed market basket are prohibitively high, we cluster market baskets by K-means based on the dynamic variables. We use the averages of each cluster as values of the dynamic variables. In a first step, we obtain cluster-specific and segment-specific purchase probabilities by averaging simulated purchases. Averaging these probabilities weighted by cluster size (i.e., the number of baskets assigned to a cluster) in the next step gives segment-specific purchase probabilities. Finally, we obtain total purchase probabilities as averages of segment specific probabilities weighted by relative segment size.

7 Empirical study

7.1 Data

Our data refer to 24,047 shopping visits to one specific grocery store over a one-year period made by a random sample of 1500 households. 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).

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		

The average number of shopping visits per household amounts to 16.031, its standard deviation to 13.464. The average basket size (i.e., the number of purchased categories) is 3.852, its standard deviation 2.654.

Table 2 shows relative marginal purchase frequencies for the 31 categories, and Table 3 the highest 20 pairwise relative frequencies. Milk is the category

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			

The 20 highest relative pairwise frequencies

most frequently purchased. Carbonated beverage and milk are the two categories most frequently purchased together.

The variables household size (number of persons) and household income with three categories are constant across baskets of the same household. Average household size amounts to 1.415, its standard deviation to 0.493. Low, medium and high income have relative frequencies of 0.507, 0.332, and 0.161, respectively.

Three binary marketing variables, feature, display, and price reductions indicate whether any brand of the respective category is on feature, display, and has its price reduced, respectively. Table 4 shows average values of these marketing

Table 4 Average values of marketing and dynamic variables

Category	Features	Displays	Price reductions	Time	Loyalty
beer	0.061	0.080	0.147	24.547	0.058
blades	0.040	0.090	0.100	34.209	0.014
carbbev	0.175	0.283	0.258	7.166	0.307
cigets	0.000	0.000	0.000	36.480	0.026
coffee	0.124	0.080	0.213	16.327	0.103
coldcer	0.151	0.114	0.170	9.127	0.218
deod	0.083	0.034	0.110	26.404	0.032
diapers	0.171	0.010	0.220	38.493	0.015
factiss	0.119	0.048	0.114	20.032	0.065
fzdin	0.187	0.007	0.199	22.568	0.070
fzpizza	0.174	0.121	0.178	19.318	0.084
hhclean	0.041	0.016	0.051	30.018	0.023
hotdog	0.094	0.034	0.153	17.661	0.081
laundet	0.106	0.081	0.130	16.006	0.092
margbutr	0.130	0.026	0.132	14.215	0.119
mayo	0.100	0.054	0.126	14.868	0.084
milk	0.129	0.009	0.186	5.838	0.359
mustketc	0.041	0.054	0.051	15.657	0.081
paptowl	0.067	0.071	0.084	15.445	0.109
peanbutr	0.133	0.053	0.150	19.628	0.062
photo	0.039	0.196	0.066	39.488	0.004
razors	0.093	0.206	0.209	41.772	0.001
saltsnck	0.154	0.267	0.152	7.818	0.274
shamp	0.094	0.077	0.137	24.280	0.041
soup	0.112	0.061	0.100	12.430	0.149
spagsauc	0.169	0.072	0.157	12.176	0.142
sugarsub	0.008	0.000	0.014	38.274	0.009
toitisu	0.095	0.081	0.116	12.876	0.133
toothbr	0.017	0.031	0.055	36.064	0.010
toothpa	0.089	0.045	0.096	22.118	0.046
yogurt	0.179	0.020	0.168	14.383	0.161

variables for each category. Frozen dinner is the most frequently featured category. We see that carbonated beverage has the highest number of both displays and price reductions.

We consider two dynamic variables, time since the last purchase of a category and category loyalty. Table 4 also contains the average time in days since the last purchase and the average loyalty for each category using a smoothing constant $\alpha = 0.2$, which puts more weight on the loyalty of the previous shopping visit. This value of the smoothing constant leads to the best performing MVL models with category loyalty as additional independent variable according to a grid search over $[0.1, 0.2, 0.3, \dots, 0.9]$. Given such a value, previous purchases are strongly smoothed.

Milk attains both the lowest average time and the highest category loyalty. We obtain a negative correlation between the two dynamic variables average time and loyalty across all categories amounting to -0.542 , which indicates an unproblematic degree of collinearity.

7.2 Model evaluation results

Tables 5 and 6 contain the evaluation results for independent logit models and multivariate logit models, respectively. We do not show results for models with household attributes (household size, income) because adding these variables does not improve model performance.

By looking at both holdout log pseudo-likelihood values (LPL) and increases of the average pseudo-probability of the respective model (*IAPP*) over the average pseudo-probability of the least complex model (homogeneous, no interactions, no independent variables) we see that:

- multivariate logit models that include pairwise interactions between categories are better than independent logit models no matter which independent variables (if any) are considered.
- models with marketing variables are better than the corresponding models without independent variables;
- features appear to be more important than price reductions, the latter appear to be more important than displays;
- models with marketing and dynamic variables are better than models with marketing variables only;
- category loyalties are more important than log-transformed times since the last category purchase specified as $\log(1 + \text{time})$ like in Boztuğ and Reutterer (2008);
- FM-MVL models perform better than their homogeneous counterparts except for models which include only features as independent variables.

The average pseudo-probability of the least complex model for the holdout data amounts to about 79% of the corresponding value for the estimation data. Consequently, more complex models as a rule have more room to improve performances in the holdout data. This fact is reflected by *IAPP* values of the same model, which are often higher for the holdout data compared to those for the estimation data.

Table 5 Evaluation of independent logit models

Number of segments	Estimation data		Holdout data	
	LPL	IAPP	LPL	IAPP
<i>No independent variables</i>				
1	– 157,513	0.00	– 80,503	0.00
2	– 157,337	0.01	– 80,499	0.00
3	– 157,273	0.02	– 80,488	0.00
4	– 157,178	0.02	– 80,478	0.00
5	– 157,014	0.03	– 80,558	– 0.01
<i>Marketing variables</i>				
fea				
1	– 137,815	2.41	– 69,384	3.00
2	– 137,599	2.46	– 69,402	2.99
dis				
1	– 141,186	1.76	– 71,628	2.03
2	– 140,946	1.81	– 71,594	2.04
3	– 140,864	1.82	– 71,585	2.04
4	– 140,745	1.84	– 71,560	2.05
5	– 140,636	1.86	– 71,529	2.06
6	– 140,525	1.88	– 71,555	2.05
red				
1	– 139,944	1.99	– 70,571	2.45
2	– 139,702	2.03	– 70,523	2.47
3	– 139,503	2.07	– 70,527	2.47
fea, dis, red				
1	– 132,320	3.80	– 66,575	4.68
2	– 131,967	3.91	– 66,419	4.79
3	– 131,675	4.00	– 66,423	4.79
<i>Marketing and dynamic variables</i>				
fea, dis, red, loy				
1	– 120,803	8.84	– 60,278	11.47
2	– 120,517	9.01	– 60,157	11.66
3	– 120,370	9.11	– 60,111	11.73
4	– 120,245	9.18	– 60,100	11.75
5	– 120,042	9.31	– 60,059	11.81
6	– 119,762	9.50	– 59,902	12.07
7	– 119,647	9.57	– 59,903	12.06
fea, dis, red, log(time + 1)				
1	– 124,283	6.92	– 62,297	8.69
2	– 124,013	7.05	– 62,251	8.75
3	– 123,771	7.18	– 62,149	8.87
4	– 123,614	7.26	– 62,102	8.93
5	– 123,403	7.37	– 62,023	9.03
6	– 123,257	7.44	– 62,090	8.94

Table 5 (continued)

Number of segments	Estimation data		Holdout data	
	LPL	IAPP	LPL	IAPP
fea, dis, red, loy, log(time + 1)				
1	– 120,368	9.11	– 60,089	11.76
2	– 120,084	9.29	– 59,969	11.96
3	– 119,902	9.40	– 59,891	12.08
4	– 119,738	9.51	– 59,929	12.02

Independent logit models: all pairwise interaction coefficients are set to zero. Log pseudo-likelihood values are rounded to the nearest integer. Rows with the number of segments equal to 1 show results for a homogeneous model, other rows show results for a mixture model with the given number of segments. The maximum number of segments given for each model variant is the number of segments for which the holdout log pseudo-likelihood decreases

fea, features; dis, displays; red, price reductions; loy, loyalties; LPL, Log pseudo-likelihood; IAPP, Increase of the average pseudo-probability

The best performing model without dynamic variables, *M0*, is a FM-MVL model with two segments, includes interactions and considers the three marketing variables features, price reductions, and displays, as independent variable. Its IAPP value amounts to 7.44. The overall best performing model *M1* includes interactions and distinguishes three segments. *M1* considers the two dynamic variables loyalties and time in addition to the three marketing variables as independent variables. *M1* doubles IAPP compared to *M0* to a value of 15.28, which constitutes a quite impressive performance improvement.

We now discuss the average coefficients of the two dynamic variables in the best performing model *M1*. Averages are determined by weighting segment coefficients by relative segment sizes. For $\log(1 + \text{time})$ we obtain only three significant coefficients, which are all positive (mayonnaise 0.151, peanut butter 0.075, toilet tissue 0.05). On the other hand, coefficients of loyalties are positive and significant for all categories. We obtain the lowest coefficient for razors (0.528), the highest for household cleaners (4.891).

7.3 Model comparison results

In this section, we compare the best performing model without dynamic variables *M0* to the overall best performing model *M1* in more detail. We start by investigating whether average category constants, average coefficients of marketing variables and average pairwise interaction coefficients differ between these two models. Table 7 shows average category constants and average coefficients of marketing variables for each model and their difference if the latter is significant. 28 of 31 category constants differ significantly, 25 of these are higher for *M0*. All category constants are negative.

Table 6 Evaluation of multivariate logit models

Number of segments	Estimation data		Holdout data	
	LPL	IAPP	LPL	IAPP
<i>No independent variables</i>				
1	– 147,341	0.88	– 76,620	0.62
2	– 146,383	1.00	– 76,608	0.63
3	– 145,637	1.10	– 76,613	0.62
<i>Marketing variables</i>				
fea				
1	– 128,830	4.97	– 66,145	5.00
2	– 127,960	5.30	– 66,177	4.97
dis				
1	– 131,926	3.92	– 68,220	3.63
2	– 130,994	4.21	– 68,204	3.64
3	– 130,213	4.47	– 68,215	3.63
red				
1	– 130,902	4.24	– 67,294	4.20
2	– 130,004	4.55	– 67,293	4.20
3	– 129,206	4.83	– 67,296	4.19
fea, dis, red				
1	– 123,624	7.25	– 63,461	7.38
<i>M0: best model without dynamic variables</i>				
2	– 122,593	7.80	– 63,405	7.44
3	– 121,793	8.25	– 63,444	7.40
<i>Marketing and dynamic variables</i>				
fea, dis, red, loy				
1	– 115,015	13.11	– 58,395	14.77
2	– 114,254	13.79	– 58,336	14.88
3	– 113,587	14.42	– 58,298	14.96
4	– 113,014	14.98	– 58,297	14.96
5	– 112,366	15.63	– 58,299	14.96
fea, dis, red, log (time + 1)				
1	– 117,584	11.02	– 59,974	11.95
2	– 116,774	11.64	– 59,924	12.03
3	– 116,098	12.19	– 59,909	12.05
4	– 115,411	12.76	– 59,867	12.12
5	– 114,693	13.39	– 59,894	12.08
fea, dis, red, loy, log (time + 1)				
1	– 114,624	13.45	– 58,215	15.13
2	– 113,874	14.14	– 58,154	15.25
<i>M1: best model with dynamic variables</i>				
3	– 113,229	14.76	– 58,140	15.28
4	– 112,714	15.28	– 58,536	14.49

Multivariate logit models: all pairwise interaction coefficients are free. Log pseudo-likelihood values are rounded to the nearest inte-

Table 6 (continued)

ger. Rows with the number of segments equal to 1 show results for a homogeneous model, other rows show results for a finite mixture model with the given number of segments. The maximum number of segments given for each model variant is the number of segments for which the holdout log pseudo-likelihood decreases

fea, features; dis, displays; red, price reductions; loy, loyalties; LPL, Log pseudo-likelihood; IAPP, Increase of the average pseudo-probability

All coefficients of marketing variables are positive, i.e., a feature (display, price reduction) increases the pseudo-probability of a purchase of the respective category. 23 of 31 feature coefficients differ significantly, 14 are higher for $M0$. 22 of 31 display coefficients differ significantly, 13 of these coefficients are higher for $M0$. 20 of 31 price reduction coefficients differ significantly. 16 of these coefficients are higher for $M0$.

About 6%, 14%, 15%, and 65% of the 465 pairwise interactions differ significantly between models at p-values of 0.10, 0.05, 0.01 and 0.005, respectively. Therefore, about 60% of pairwise interactions coefficients differ significantly between models at p-values ≤ 0.05 . 204 of these interactions are higher for $M0$, 211 of these interactions are positive for $M1$. Table 8 contains the 20 average interaction coefficients with highest absolute differences between the two models. The lowest absolute t-statistic of these differences amounts to 12.91. Sixteen of these interactions are higher for $M0$, and 13 are positive for $M1$.

We also examine whether models $M0$ and $M1$ lead to different results on the relations between categories. We measure the relation of a category j conditional on another category j' by the average marginal effect with respect to the purchase pseudo-probability of category j . We average across baskets by keeping the observed values of independent variables and the observed purchase incidences of categories other j and j' . 94% of marginal effects differ significantly between models, 72% of these are higher for model $M0$.

As 76% of marginal effects are positive, most category pairs can be seen as purchase complements. Nonetheless, our results hint at a considerable number of substitutive relations between category pairs. Examples of substitutive relations indicated by model $M1$ are razors and milk, photo and yogurt as well as cigarettes and soup with average marginal effects of -0.294, -0.059, and -0.039, respectively.

Table 9 shows the 20 highest marginal effects in absolute size. These marginal effects are all positive and significantly lower for $M1$. The minimum absolute t-statistic of differences between the two models is 46.40.

7.4 Managerial implications

In this section, we answer the question whether the two models $M0$ and $M1$ entail different managerial implications. Based on K-means for 62 dynamic variables (two variables in each of 31 product categories), we choose six clusters. This procedure drastically reduces computation time as we only have to sample purchases for each cluster using cluster-specific arithmetic means of the dynamic variables followed by

Table 7 Average category constants and average coefficients of marketing variables

Category	M1	M0	M0–M1	Category	M1	M0	M0–M1
<i>Category constants</i>							
carbbev	– 2.789	– 1.967	0.822	milk	– 1.637	– 0.856	0.781
yogurt	– 3.462	– 2.689	0.773	toitisu	– 3.645	– 3.082	0.563
saltsnck	– 3.079	– 2.518	0.561	mayo	– 4.012	– 3.487	0.525
fzdin	– 4.134	– 3.671	0.463	peanbutr	– 4.769	– 4.340	0.429
soup	– 3.227	– 2.807	0.420	fzpizza	– 4.865	– 4.463	0.402
coldcer	– 2.981	– 2.582	0.400	margbutr	– 3.338	– 2.953	0.385
coffee	– 4.328	– 3.963	0.365	spagsauc	– 3.542	– 3.188	0.354
paptowl	– 3.867	– 3.513	0.354	beer	– 4.066	– 3.760	0.306
factiss	– 4.250	– 3.946	0.304	photo	– 6.931	– 7.198	– 0.266
hotdog	– 4.201	– 3.966	0.235	cigets	– 3.905	– 3.672	0.233
<i>Features</i>							
diapers	2.435	2.968	0.533	peanbutr	0.454	0.196	– 0.257
hhclean	3.063	2.873	– 0.190	razors	3.859	4.038	0.179
coffee	3.776	3.613	– 0.163	spagsauc	2.774	2.636	– 0.138
hotdog	2.529	2.392	– 0.137	toitisu	2.287	2.157	– 0.130
carbbev	0.680	0.804	0.124	soup	2.719	2.598	– 0.121
laundet	2.992	2.878	– 0.114	milk	1.159	1.046	– 0.113
deod	3.765	3.659	– 0.105	mayo	0.304	0.394	0.091
photo	3.236	3.146	– 0.090	fzdin	2.032	2.120	0.087
yogurt	2.760	2.676	– 0.083	coldcer	2.062	2.141	0.079
sugarsub	0.839	0.912	0.072	mustketc	0.498	0.427	– 0.070
<i>Displays</i>							
beer	3.561	3.838	0.277	fzdin	2.950	2.676	– 0.274
yogurt	0.691	0.958	0.267	toitisu	2.923	2.679	– 0.243
hhclean	2.424	2.601	0.176	paptowl	3.416	3.245	– 0.171
laundet	2.746	2.891	0.145	toothpa	3.472	3.378	– 0.093
fzpizza	3.685	3.761	0.076	hotdog	3.348	3.273	– 0.074
soup	2.766	2.710	– 0.055	factiss	2.575	2.521	– 0.054
spagsauc	1.755	1.714	– 0.042	saltsnck	2.170	2.129	– 0.041
peanbutr	3.246	3.206	– 0.039	mustketc	4.396	4.435	0.039
coffee	2.909	2.946	0.037	blades	4.940	4.903	– 0.037
shamp	4.119	4.154	0.035	carbbev	1.960	1.928	– 0.032
<i>Price reductions</i>							
hotdog	2.283	2.509	0.226	carbbev	1.297	1.072	– 0.225
peanbutr	2.934	3.157	0.223	coffee	1.787	1.991	0.204
beer	2.486	2.658	0.172	hhclean	0.389	0.553	0.164
soup	0.171	0.320	0.149	spagsauc	0.734	0.865	0.131
milk	1.207	1.315	0.108	fzpizza	0.853	0.949	0.096
diapers	2.051	2.140	0.089	mayo	2.873	2.804	– 0.069
factiss	1.636	1.569	– 0.067	fzdin	0.303	0.364	0.061
saltsnck	0.094	0.150	0.056	toothbr	1.643	1.690	0.047
coldcer	0.551	0.515	– 0.036	mustketc	2.933	2.961	0.028
sugarsub	0.606	0.632	0.026	razors	2.338	2.363	0.026

Table 7 (continued)

20 average category constants and coefficients of each marketing variable with highest absolute differences between models and a minimum absolute t -statistic of 2.0

Table 8 Average pairwise interaction coefficients

		$M1$	$M0$	$M0-M1$
cigets	razors	- 0.231	0.139	0.370
cigets	coffee	0.392	0.657	0.265
razors	toothbr	0.940	1.203	0.263
diapers	yogurt	0.416	0.667	0.251
milk	razors	- 1.809	- 1.584	0.225
fzdin	razors	- 2.081	- 2.286	- 0.205
paptowl	toitisu	1.167	1.365	0.198
peanbutr	razors	- 2.173	- 2.352	- 0.179
milk	photo	0.312	0.484	0.172
fzdin	sugarsub	0.198	0.361	0.163
peanbutr	yogurt	0.237	0.399	0.162
laundet	paptowl	0.492	0.654	0.162
photo	yogurt	- 0.596	- 0.437	0.159
carbbev	cigets	0.150	0.302	0.153
beer	blades	0.455	0.607	0.152
beer	diapers	- 0.383	- 0.533	- 0.150
carbbev	diapers	0.175	0.324	0.149
margbutr	razors	- 2.274	- 2.420	- 0.147
laundet	toitisu	0.433	0.579	0.145
coldcer	yogurt	0.508	0.652	0.144

The 20 interactions with highest absolute differences between models (minimum absolute t -statistic 12.91)

weighting according to cluster sizes. Without clustering sampling for each of the 24,047 market baskets using observed values of dynamic variables would be necessary (please also see Sect. 6).

Table 10 shows the own effects of marketing variables that differ between the two models by at least 0.005 (i.e., a half percentage point) in absolute size. For features, such higher differences occur in 55% of the categories. For displays and price reduction, we see higher differences in 39% and 32% of the categories. Most of these higher differences are positive (71%, 92%, and 90% for features, displays, and price reductions, respectively) and therefore indicate positive omitted variable bias, whose principle we have explained in Sect. 2.

Model $M0$ without dynamic variables frequently overestimates own effects by falsely attributing the effects of the omitted dynamic variables to the marketing variables. These positive omitted variable biases can be traced back to both positive effects of the more important dynamic variable category loyalty on purchase

Table 9 Category relations measured by average marginal effects

Conditional on		<i>M1</i>	<i>M0</i>	<i>M0–M1</i>
diapers	yogurt	0.051	0.108	0.057
paptowl	toitisu	0.146	0.198	0.051
razors	yogurt	0.127	0.177	0.050
photo	milk	0.062	0.108	0.046
saltsnck	carbbev	0.108	0.150	0.041
toitisu	paptowl	0.119	0.159	0.040
coldcer	yogurt	0.060	0.099	0.038
yogurt	milk	0.053	0.091	0.038
sugarsub	yogurt	0.073	0.106	0.033
diapers	carbbev	0.029	0.062	0.033
cigets	carbbev	0.024	0.057	0.033
peanbutr	yogurt	0.028	0.060	0.033
carbbev	saltsnck	0.101	0.133	0.032
fzpizza	saltsnck	0.053	0.081	0.028
yogurt	coldcer	0.074	0.102	0.028
milk	yogurt	0.030	0.057	0.027
cigets	coffee	0.026	0.052	0.025
beer	milk	0.018	0.044	0.025
laundet	toitisu	0.046	0.071	0.024
photo	carbbev	0.043	0.067	0.024

20 category relations with highest absolute differences between models (minimum absolute *t*-statistic 46.40)

probabilities and positive correlations between loyalties with marketing variables. Across all categories, these correlations amount to 0.150, 0.130, 0.143 for features, displays, and price reductions, respectively.

Results for cross effects stand in marked contrast to those for own effects (see Table 11). We get only a few absolute differences between the two models of at least 0.005 (4.6%, 6.9%, and 0.8% of the $930 = 31 \times 30$ cross effects for features, displays, and price reductions, respectively). For these differences, all coefficients of *M0* are higher. In other words, models *M0* and *M1* agree on the size of a clear majority of cross effects.

We consider the number of purchases of any product category as managerial objective to demonstrate implications of positive biases of own effects. Purchases equal the sum of purchase probabilities inferred for a model across households. If managers use the basic model *M0* in spite of its worse statistical performance, they would set more sales promotion activities in many categories due to overestimating purchase increases. We assess the importance of a positive bias by expressing it as percentage of the marginal purchase frequency of the respective category (see Table 12). These percentages measure how much managers overestimate purchase increases in a category in relative terms if they ignore dynamic variables by relying on model *M0*. On average, these percentages amount to 10.64, 14.66 and 10.26 for features, displays and price reduction, respectively. Percentages higher than ten

Table 10 Own effects of marketing variables on purchase probabilities

Category	$M1$	$M0$	$M0-M1$	Category	$M1$	$M0$	$M0-M1$
<i>Features</i>							
beer	0.029	0.039	0.010	carbbev	0.074	0.104	0.029
coldcer	0.195	0.220	0.024	diapers	0.021	0.026	0.006
factiss	0.089	0.096	0.006	fzdin	0.070	0.086	0.016
fzpizza	0.114	0.124	0.010	hhclean	0.058	0.051	0.007
hotdog	0.114	0.106	0.008	laundet	0.158	0.151	0.007
margbutr	0.198	0.214	0.015	milk	0.135	0.127	0.008
peanbutr	0.014	0.005	0.009	saltsnck	0.256	0.288	0.033
soup	0.219	0.226	0.007	toitisu	0.164	0.172	0.007
yogurt	0.206	0.229	0.023				
<i>Displays</i>							
beer	0.154	0.199	0.045	carbbev	0.213	0.229	0.016
coffee	0.167	0.184	0.017	coldcer	0.217	0.239	0.022
fzpizza	0.167	0.191	0.024	laundet	0.146	0.162	0.015
margbutr	0.123	0.133	0.011	mustketc	0.297	0.304	0.007
saltsnck	0.212	0.235	0.023	shamp	0.137	0.142	0.005
toothpa	0.137	0.129	0.007	yogurt	0.052	0.082	0.030
<i>Price reductions</i>							
beer	0.079	0.102	0.023	coffee	0.076	0.088	0.013
fzpizza	0.026	0.032	0.006	hotdog	0.089	0.099	0.010
mayo	0.148	0.141	0.007	milk	0.142	0.158	0.016
peanbutr	0.105	0.117	0.012	saltsnck	0.009	0.017	0.008
soup	0.013	0.025	- 0.012	spagsauc	0.046	0.053	- 0.007

Own effects with a minimum absolute difference between models of 0.005

(minimum absolute t -statistic 58.49)

occur in nine categories. We even notice relative overestimations of at least 20% for features of diapers and household cleaners, for displays of beer & ale and of frozen pizza as well as for price reduction of beer.

8 Conclusion

MVL models that allow for pairwise interactions between product categories and for latent heterogeneity clearly outperform their less complex counterparts. In a similar manner, adding dynamic variables leads to better model performance. Among dynamic variables exponentially smoothed category loyalties, which previous publications have ignored, turn out to be more important than log-transformed times since the last category purchase.

Comparing two FM-MVL models, a basic model with marketing variables as independent variables and an enlarged model that in addition considers dynamic

Table 11 Cross effects of marketing variables on purchase probabilities

Independent category	Dependent category	$M1$	$M0$	$M0-M1$
<i>Features</i>				
saltsnck	carbbev	0.040	0.063	0.023
coldcer	yogurt	0.018	0.032	0.014
toitisu	carbbev	0.008	0.020	0.012
coldcer	saltsnck	0.016	0.027	0.011
yogurt	coldcer	0.025	0.036	0.011
yogurt	milk	0.017	0.027	0.010
yogurt	saltsnck	0.009	0.018	0.010
laundet	saltsnck	0.009	0.018	0.010
coffee	saltsnck	0.011	0.020	0.010
spagsauc	carbbev	0.008	0.017	0.010
coffee	carbbev	0.007	0.017	0.009
soup	saltsnck	0.013	0.022	0.009
margbutr	carbbev	0.007	0.016	0.009
saltsnck	soup	0.009	0.018	0.009
fzpizza	saltsnck	0.011	0.020	0.008
saltsnck	milk	0.003	0.011	0.008
toitisu	saltsnck	0.010	0.019	0.008
carbbev	saltsnck	0.011	0.019	0.008
coldcer	carbbev	0.007	0.015	0.008
spagsauc	saltsnck	0.013	0.020	0.008
<i>Displays</i>				
saltsnck	carbbev	0.032	0.049	0.017
paptowl	carbbev	0.018	0.034	0.017
fzpizza	saltsnck	0.016	0.032	0.016
paptowl	toitisu	0.047	0.063	0.016
toitisu	carbbev	0.011	0.026	0.014
mustketc	saltsnck	0.022	0.035	0.014
coldcer	yogurt	0.020	0.034	0.014
paptowl	saltsnck	0.012	0.024	0.012
fzpizza	carbbev	0.008	0.020	0.012
carbbev	saltsnck	0.030	0.042	0.012
laundet	saltsnck	0.008	0.020	0.012
mustketc	carbbev	0.017	0.029	0.011
coldcer	saltsnck	0.018	0.029	0.011
soup	saltsnck	0.014	0.025	0.011
beer	saltsnck	0.009	0.019	0.010
toitisu	saltsnck	0.014	0.024	0.010
laundet	carbbev	0.008	0.018	0.009
hotdog	carbbev	0.008	0.018	0.009
coffee	saltsnck	0.008	0.017	0.009
saltsnck	milk	0.002	0.011	0.009

Table 11 (continued)

Independent category	Dependent category	<i>M1</i>	<i>M0</i>	<i>M0–M1</i>
<i>Price reductions</i>				
milk	yogurt	0.007	0.014	0.007
mustketc	carbbev	0.009	0.016	0.007
mustketc	saltsnck	0.013	0.019	0.006
carbbev	saltsnck	0.020	0.025	0.005
milk	saltsnck	0.001	0.006	0.005
beer	saltsnck	0.004	0.009	0.005
peanbutr	saltsnck	0.008	0.013	0.005

For features and display 20 cross effects with highest absolute differences between models; for price reductions 7 cross effects with a minimum absolute difference of 0.005

Table 12 Positive own effect biases as percentages of relative marginal purchase frequencies

<i>Features</i>							
beer	13.16	carbbev	7.25	coldcer	8.57	diapers	30.00
factiss	7.14	fzdin	17.78	fzpizza	9.09	hhclean	23.33
hotdog	7.77	laundet	5.93	margbutr	9.49	milk	1.68
peanbutr	11.25	saltsnck	9.40	soup	3.55	toitisu	4.09
yogurt	11.39						
<i>Displays</i>							
beer	59.21	carbbev	4.00	coffee	12.50	coldcer	7.86
fzpizza	21.82	laundet	12.71	margbutr	6.96	mustketc	6.86
saltsnck	6.55	shamp	9.43	toothpa	13.21	yogurt	14.85
<i>Price reductions</i>							
beer	30.26	coffee	9.56	fzpizza	5.45	hotdog	9.71
mayo	6.42	milk	3.36	peanbutr	15.00	saltsnck	2.28

variables, shows that coefficients of marketing variables differ in a clear majority of categories. Most coefficients for features, displays and price reductions are lower for the enlarged model. A majority of pairwise interactions differs significantly between the two models being usually lower for the enlarged model. Almost all relations between categories measured by average marginal effects are different and usually lower for the enlarged model.

It also turns out that managerial implications for the two models differ. The basic model suffers from positive omitted variable biases, i.e., it overestimates the own effects of marketing variables on purchase probabilities in many product categories. The omitted variable bias provides another explanation for the well-known problem of “overpromotion” in retailing. It seems that if retailers ignore loyalty (the extent to which people would have bought a product anyway) they are inclined to promote their products too much.

We expect such biases in market situations characterized by two conditions. One condition boils down to having many product categories with high or medium loyalty values. For the data analyzed here, the omitted variable bias of a marketing variable in a category (the difference of its own effects between models $M0$ and $M1$) increases with loyalty. This increase is reflected by correlations between category loyalties (averaged across all market baskets) and biases amounting to 0.502, 0.346, 0.216 for features, displays, and price reductions, respectively.

Taking purchase frequency as proxy for loyalty, positive biases should be more common in categories with high or medium purchase frequencies. High or medium purchase frequencies are typical for categories such as food, detergents, cleaning products, hygienic products, cosmetics, pet food, some clothing products, footwear, fuel, alcohol, and digital entertainment products.

The other condition for positive biases requires positive correlations of marketing variables with loyalties across categories. Once again, we use purchase frequency as proxy of category loyalty. A study of Fader and Lodish for 331 different grocery categories shows that features, displays and price reductions are more frequent in categories with high penetration and high purchase frequency in comparison to categories with high penetration and low purchase frequency (Fader and Lodish 1990). High penetration means that a high percentage of households makes at least one purchase per year.

Let us qualify the discussion of the model comparison results. As we cannot be sure that the investigated models include the true model, we have to expect a bias corresponding to the distance between the true model and any estimated model. Clearly, a direct comparison to the unknown true model is out of the question. We had to rely on indirect comparison by using holdout data that the true model has generated. On the other hand, the basic model is too simple as its low performance shows.

Of course, several avenues for further related research on dynamic multicategory choice remain. One possibility consists in investigating the relevance of dynamic variables for non-food product categories (e.g., consumer electronics, apparel). Moreover, future research efforts could deal with models with alternative functional forms (e.g., the multivariate probit model).

Acknowledgements I thank two anonymous reviewers for several suggestions that helped to improve the paper. I also acknowledge that the editors of this special issue provided beneficial comments on the paper's original version.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Agrawal R, Srikant R (1994) Fast algorithms for mining association rules in very large databases. In: Proceedings of the 20th international conference on VLDB, Santiago, Chile
- Andrews RL, Ainslie A, Currim IS (2002) An empirical comparison of logit choice models with discrete versus continuous representations of heterogeneity. *J Mark Res* 39:479–487
- Ansari A, Mela CF (2003) E-customization. *J Mark Res* 40:24–33
- Bel K, Fok D, Paap R (2018) Parameter estimation in multivariate logit models with many binary choices. *Econ Rev* 37:534–550
- Besag J (1972) Nearest-neighbour systems and the auto-logistic model for binary data. *J R Stat Soc B* 34:75–83
- Besag J (1974) Spatial interaction and the statistical analysis of lattice systems. *J R Stat Soc B* 35:192–236
- Besag J (2004) An introduction to Markov chain Monte Carlo methods. In: Johnson ME, Khudanpur SP, Ostendorf M, Rosenfeld R (eds) *Mathematical foundations of speech and language processing*. Springer, New York, pp 247–270
- Betancourt R, Gautschi D (1990) Demand complementarities, household production, and retail assortments. *Mark Sci* 9:146–161
- Boztuğ Y, Hildebrandt L (2008) Modeling joint purchases with a multivariate MNL approach. *Schmalenbach Bus Rev* 60:400–422
- Boztuğ Y, Reutterer T (2008) A combined approach for segment-specific market basket analysis. *Eur J Oper Res* 187:294–312
- Bronnenberg BJ, Kruger MW, Mela CF (2008) Database paper: the IRI marketing data set. *Mark Sci* 27:745–748
- Chiang J (1991) A simultaneous approach to the whether, what and how much to buy questions. *Mark Sci* 10:297–315
- Chib S, Seetharaman PB, Strijnev A (2002) Analysis of multi-category purchase incidence decisions using IRI market basket data. In: Franses PH, Montgomery AL (eds) *Econometric models in marketing*. JAI, Amsterdam, pp 57–92
- Chintagunta P (1993) Investigating purchase incidence, brand choice and purchase quantity decisions of households. *Mark Sci* 12:184–208
- Cox DR (1972) The analysis of multivariate binary data. *J R Stat Soc C* 21:113–120
- Dippold K, Hruschka H (2013) A model of heterogeneous multicategory choice for market basket analysis. *Rev Mark Sci* 11:1–31
- Duvvuri SD, Ansari V, Gupta S (2007) Consumers' price sensitivities across complementary categories. *Manag Sci* 53:1933–1945
- Fader PS, Lodish LM (1990) A cross-category analysis of category structure and promotional activity for grocery products. *J Mark* 54:52–65
- Gedenk K, Neslin SA, Ailawadi KL (2010) Sales promotion. In: Krafft M, Mantrala M (eds) *Retailing in the 21st century*. Springer, Berlin, pp 303–317
- Gentzkow M (2007) Valuing new goods in a model with complementarity: online newspapers. *Am Econ Rev* 97:713–744
- Greene WH (2003) *Econometric analysis*, 5th edn. Pearson Education, Upper Saddle River
- Guadagni PM, Little JDC (1983) A logit model of brand choice calibrated on scanner data. *Mark Sci* 2:203–238
- 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
- Hruschka H (2013) Comparing small and large scale models of multicategory buying behavior. *J Forecast* 32:423–434
- Hruschka H (2014) Linking multi-category purchases to latent activities of shoppers: analysing market baskets by topic models. *Mark ZFP* 36:268–274
- Hruschka H (2017a) Analyzing the dependences of multicategory purchases on interactions of marketing variables. *J Bus Econ* 87:295–313
- Hruschka H (2017b) Multi-category purchase incidences with marketing cross effects. *Rev Manag Sci* 11:443–469
- Hruschka H (2017c) Multicategory purchase incidence models for partitions of product categories. *J Forecast* 36:230–240

- Jacobs B, Donkers B, Fok D (2016) Model-based purchase predictions for large assortments. *Mark Sci* 35:389
- Keane MP (1997) Modeling heterogeneity and state dependence in consumer choice behavior. *J Bus Econ Stat* 15:310–327
- Kwak K, Duvvuri SD, Russell GJ (2015) An analysis of assortment choice in grocery retailing. *J Retail* 91:19–33
- Manchanda P, Ansari A, Gupta S (1999) The shopping basket: a model for multi-category purchase incidence decisions. *Mark Sci* 18:95–114
- McLachlan G, Basford K (1988) Mixture models: inference and applications to clustering. Marcel Dekker, New York
- Meyer R, Erdem T, Feinberg F et al (2017) Dynamic influences on individual choice behavior. *Mark Lett* 8:349–360
- Ngatchou-Wandji J, Bulla J (2013) On choosing a mixture model for clustering. *J Data Sci* 11:157–179
- Richards TJ, Hamilton SF, Yonezkawa K (2018) Retail market power in a shopping basket model of supermarket competition. *J Retail* 94:328–342
- Ruiz F, Athey S, Blei D (2020) Shopper: a probabilistic model of consumer choice with substitutes and complements. *Ann Appl Stat* 14:1–27
- Russell GJ, Petersen A (2000) Analysis of cross category dependence in market basket selection. *J Retail* 76:69–392
- Schröder N, Hruschka H (2017) Comparing alternatives to account for unobserved heterogeneity in direct marketing models. *Decis Support Syst* 103:24–33
- Solnet D, Boztuğ Y, Dolnicar S (2016) An untapped gold mine? Exploring the potential of market basket analysis to grow hotel revenue. *Int J Hosp Manag* 56:119–125
- Wooldridge JM (2013) Introductory econometrics: a modern approach, 5th edn. Southwestern, Cengage Learning, Mason

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.