Regensburger
DISKUSSIONSBEITRÄGE
zur Wirtschaftswissenschaft

University of Regensburg Working Papers in Business,
Economics and Management Information Systems

Variable Selection for Market Basket Analysis

Katrin Dippold*

Harald Hruschka**

February 2010

Nr. 443

JEL Classification: *C13, C52, L81, M31*

Key Words: Market basket analysis, cross category effects, variable selection, multivariate logit model, pseudo likelihood estimation

* Katrin Dippold is a research assistant at the Department of Marketing, Faculty of Business, Economics and Management Information Systems at the University of Regensburg, 93040 Regensburg, Germany
Phone: +49-941-943-2276, E-mail: katrin.dippold@wiwi.uni-regensburg.de
** Prof. Dr. Harald Hruschka, Department of Marketing, Faculty of Business, Economics and Management Information Systems at the University of Regensburg, 93040 Regensburg, Germany

Variable Selection for Market Basket Analysis

Katrin Dippold Harald Hruschka

February 2010

Results on cross category effects obtained by explanatory market basket analyses may be biased as studies typically investigate only a small fraction of the retail assortment (Chib et al. 2002). We use Bayesian variable selection techniques to determine significant cross category effects in a multivariate logit model. Hence, we achieve a reduction of coefficients to be estimated which decreases computation time heavily and thus allows to consider more product categories than most previous studies. We present three different approaches to variable selection and find that an adaptation of a technique by Geweke (2005) meets the requirements of market basket analysis best, namely high numbers of observations and cross category effects. We show (1) that only a moderate fraction of possible cross category effects are significantly different from zero (one third for our data), (2) that most of these effects indicate complementarity and (3) that the number of considered product categories influences significances of cross category effects.

Keywords: Market basket analysis, cross category effects, variable selection, multivariate logit model, pseudo likelihood estimation

1 Introduction

As a rule, consumer purchase decisions involve multiple products. The most prominent example is the so called market basket, which is defined as the set of product categories purchased by one shopper in one store during a single shopping trip. The shopper is confronted with a "pick-any" decision, where he has to choose a subset of categories from a retailer's assortment. For every single category, he decides if he wants to buy it or not, leading to as many purchase or non-purchase decisions as categories are available at the store (Russell et al. 1997, 1999). In contrast to brand choice, the number of chosen alternatives, i.e., categories, is not known a priori.

The main goal of market basket analysis is to uncover the pattern of cross category relations within a retailer's assortment. Possible relations include complementarity, substitution, and independence. Usually, two categories are regarded as complements (substitutes) if their cross price elasticities are negative (positive) (e.g., Shocker et al. 2004; Bucklin et al. 1998; Russell and Petersen 2000). These concepts are modified in market basket analysis where categories are considered as complements (substitutes) if their cross effects are positive (negative), that is if categories are purchased jointly more (less) frequently than expected under stochastic independence (Betancourt and Gautschi 1990; Hruschka 1991; Hruschka et al. 1999; Mulhern and Leone 1991).

There are various causes for cross category effects. Several categories may be bought at the same time for the sake of convenience (Bell and Latin 1998; Russell et al. 1999) or to minimize transaction costs of purchase (e.g., costs of information search, purchase initiation, transport of goods or invoice settlement). This tendency for one-stop-shopping leads to an overall complementarity between categories of one assortment. On the other hand, the fact that categories compete for limited budgets of shoppers contributes to substitutability between categories (Niraj et al. 2008).

Moreover, different complementarity effects may be distinguished w.r.t. consumption and purchasing, respectively. Consumption complementarity means that the utility for the joint consumption of two categories is higher than the sum of their individual utilities (Shocker et al. 2004; Niraj et al. 2008). Cake-mix and frosting represent a well known example. Purchase complementarity is assumed in the marketing literature if marketing activities in one category influence purchase decisions not only in the promoted category but also in other categories (Erdem 1998; Manchanda et al. 1999; Shocker et al. 2004).

Complementarity and substitution are rather complex concepts which often lead to contradictory conclusions. Though these concepts may be helpful for prior determination of relevant cross category effects in small sized problems (Manchanda et al. 1999; Niraj et al. 2008), such an approach appears to be futile for larger assortments. Results of empirical studies on relations of categories in retail assortments are not consistent. The probit model of Chib et al. (2002) for 12 categories reveals positive interaction effects indicating a general assortment-wide complementarity. Also, Hruschka et al. (1999) find mainly complementary effects between various categories. In their study, only tobacco product are subject to substitutive effects. Russell and Petersen (2000) uncover only substitutive relations among paper goods categories. Boztuğ and Hildebrandt (2008) replicate the substitutive relations for the paper goods categories. They also find substi-

tutive relations among various breakfast beverages and among different detergents. On the other hand, these authors obtain complementary relations among normal beverages. Because of the difficulties to determine relationships a priori and contradictory empirical results, we conclude that the use of an appropriate statistical method is necessary to decide on strength and type of relations between categories.

Over the last decades, different techniques to analyze market basket data and study cross category effects have been developed in the fields of statistics, data mining, and marketing research. This progress has been promoted by the growing availability of market basket data acquired by conventional and electronic retailers, loyalty card programs and data providers (e.g., Boztuğ and Silberhorn 2006). We follow the established classification of market basket analysis methods into exploratory and explanatory models (Mild and Reutterer 2003; Boztuğ and Silberhorn 2006; Boztuğ and Hildebrandt 2008). Exploratory models typically aim at the discovery of purchase patterns or basket clusters from POS scanner data. For the most part, exploratory models do not include additional covariates, such as marketing mix variables or consumer demographics. Methods like association rules (e.g., Buchta 2007), vector quantization (e.g., Boztuğ and Reutterer, 2008), collaborative filtering (e.g., Mild and Reutterer 2003), and association measures (e.g., Hruschka 1985) condense a large amount of input data to a few statements, rules, prototypes or similarity measures. Of course, such methods involve loss of information (Hildebrandt and Boztuğ 2007). Besides, exploratory models are not well suited for forecasting (Boztuğ and Hildebrandt 2008). To summarize, exploratory model types can be used to uncover cross category relations, but not to explicate their causes. Still, they are useful for a first step to discover unknown relationships.

Explanatory models, on the other hand, aim at explaining effects and therefore include additional covariates. Data sets for explanatory models not only consist of market baskets, they also comprise customer attributes and marketing mix variables. Usually, models have logit or probit functional forms. Seminal work on the application of a probit model for market basket analysis was done by Manchanda et al. (1999). A multivariate probit model derived from random utility theory represents interdependent and simultaneous choices of categories. Characteristic of the probit model, cross-category effects can be asymmetric across pairs of categories. These effects are incorporated in error correlations which makes interpretation more difficult. Russell and Petersen (2000) apply the multivariate logit (MVL) model to market basket analysis.

Typically, the number of cross category effects studied by explanatory models is limited in scope. Both Manchanda et al. (1999) and Russel and Petersen (2000) investigate four categories only. We find that only a few studies with multivariate logit and probit models have investigated more than six categories at a time. An overview of publications that focus on multicategory purchase incidence decisions with logit and probit models is given in table 1.

Only two publications study a comparatively higher number of categories. Hruschka et al. (1999) implement the MVL model for 73 categories. They estimate this model after discovering significant cross category effects of univariate logit models by a stepwise forward-backward procedure. Boztuğ and Reutterer (2008) proceed in two steps. In the first step, they start from basket data on 65 categories and determine prototypes

Table 1: Maximum number of product categories investigated

Logit		**Probit**	
Publication	Categories	Publication	Categories
Hruschka et al. (1999)	73	Manchanda et al. (1999)	4
Russell & Petersen (2000)	4	Chib et al. (2002)	12
Boztuğ & Hildebrandt (2008)	5	Duvvuri et al. (2007)	6
Boztuğ & Reutterer (2008)	65		

of market baskets by vector quantization. In the second step, they estimate one MVL model for each prototype with about 5 categories.

We stick to the MVL model in this paper, but eliminate insignificant cross category effects by Bayesian variable selection methods. Therefore, we are in a position to consider a much higher number of categories than most previous studies. Moreover, we are able to investigate whether cross category effects are biased if a considerable number of categories, which market baskets of shoppers may contain, are ignored.

The MVL model is explained in section 2. Next, we state why variable selection is the appropriate concept for our goals and present three different selection methods (section 3). We apply these methods to a data set acquired at a Bavarian supermarket and discuss the results in section 4. The paper ends with conclusions and remarks on future research possibilities (section 5).

2 Model and Estimation

2.1 Multivariate Logit Model

The MVL model is based upon seminal work of Cox (1972) and Besag (1974). Data input consists of $i = 1, \cdots, I$ market baskets. A market basket i is a binary vector $Y_i = [Y_{i1}, ..., Y_{iJ}]$ of a certain combination of categories $j = 1, \cdots, J$. A binary variable Y_{ij} equal to one indicates that category j is present in market basket i. Deterministic utility $V(Y_i)$ of market basket i is specified as:

$$V(Y_i) = \sum_j \alpha_j Y_{ij} + \sum_{j<k} \theta_{jk} Y_{ij} Y_{ik} \tag{1}$$

This specification implies $\theta_{jj} = 0$. α_j denotes the constant term of category j. θ_{jk} symbolizes a first order interaction or cross category effect between categories j and k. It is important to notice that $\theta_{jk} = \theta_{kj}$. Otherwise, the model would not be identified, i.e., there would be no unique coefficient vector maximizing the likelihood (see Russell and Petersen (2000) for an intuitive proof). The model is restricted to first-order interaction effects in order to limit the number of coefficients and to keep the analysis tractable and frugal. Interactions between more than two categories are neglected. We assume that

absolute values of higher order interaction coefficients are small compared to first order interaction coefficients.

Purchase probability of market basket Y_i (which equals the joint probability of category purchases) is given by the MVL model[1] with Y^* denoting the set of all $|Y^*| = 2^J$ potential baskets:

$$P(Y_i) = \exp(V(Y_i))/\sum_{Y^*} \exp(V(Y^*)) \tag{2}$$

Because of the complex form of the joint probability distribution, we work with full conditional category probabilities which are much easier to compute. Besag (1974) and Cressie (1993) prove that the joint probability $P(Y_i)$ can be uniquely derived from a consistent set of full conditional distributions $P(Y_{ij} = 1|Y_{ik})$ (for details on the derivation, see Russell and Petersen (2000) and the appendix of Boztuğ and Hildebrandt (2008).).

The conditional purchase probability of category j given purchases of other categories $k \neq j$ can be deduced as

$$P(Y_{ij} = 1|Y_{ik}) = \exp(V_{i,j|k})/(1 + \exp(V_{i,j|k})) \tag{3}$$

$V_{ij|k} = \alpha_j + \sum_{k \neq j} \theta_{jk} Y_{ik}$ gives the conditional utility of a purchase from category j in basket i given purchases of other categories.

2.2 Estimation

Because of the complexity of the denominator of the joint probability (expression (2)), maximum likelihood (ML) estimation of the MVL model becomes intractable for a larger number of categories. That is why we use pseudo likelihood (PL) estimation which results in coefficients that are consistent but not efficient (Moon and Russell 2004).

Besag (1975) suggested PL estimation of the MVL as approximation to ML. PL estimation was developed further by Cressie (1993). Researchers in the field of Bayesian learning and pattern recognition proposed or applied PL approximation (e.g., Murray and Ghahramani 2004; Wang et al. 2000; Yu and Cheng 2003). The idea was also employed in marketing applications of the MVL model (e.g., Moon and Russell 2004) as well as in other fields (see, e.g., Ward and Gleditsch (2002) for an application in political science or Sherman et al. (2006) for an application to medical data).

The PL of the MVL model given coefficients $\beta = (\alpha, \theta)$ is defined as (Cressie 1993):

$$PL(\beta) = \prod_i \prod_j P(Y_{ij}|Y_{ik}, \beta) \tag{4}$$

One element $P(Y_{ij}|Y_{ik}, \beta)$ of the pseudo likelihood is expressed as

$$P(Y_{ij}|Y_{ik}, \beta) = \exp(\alpha_j Y_{ij} + \sum_{k \neq j} \theta_{jk} Y_{ij} Y_{ik})/(1 + \exp(\alpha_j + \sum_{k \neq j} \theta_{jk} Y_{ik})) \tag{5}$$

[1] The MVL model is also known as autologistic model and is frequently used to analyze autocorrelation in space or time (Magnussen and Reeves 2007).

Taking logs we obtain the pseudo loglikelihood (PLL):

$$PLL(\beta) = \sum_i \sum_j \log P(Y_{ij}|Y_{ik}, \beta) \tag{6}$$

3 Selection of Cross Category Effects

The model introduced in section 2 consists of $J + J(J-1)/2$ coefficients. Even for assortments of moderate size, one has to deal with the involved complexity of estimating and interpreting a large number of coefficients. Of course, adding price and promotion variables would further increase complexity.

That is why we intend to reduce the possible $J(J-1)/2$ cross category effects. A lower number of cross category coefficients not only eases interpretation, it also speeds up estimation. To calculate the conditional probability $P(Y_{ij} = 1|Y_{ik})$, we do not have to sum over all $J-1$ other categories, but only over $p^\delta - 1$ interacting categories with $p^\delta - 1$ as number of $\theta_{jk} \neq 0$. The third and maybe most important advantage of excluding irrelevant coefficients is model robustness, meaning that the PLL value does not change much if the model is applied to validation data which have not been used for estimation. Estimating all possible coefficients, on the other hand, could result in overfitting the model with many coefficients reproducing noise in the estimation data.

A priori, we do not know which pairs of categories interact ($\theta_{jk} \neq 0$) and which pairs of categories are independent ($\theta_{jk} = 0$). Therefore, we use variable selection techniques to eliminate insignificant cross category coefficients. To our knowledge, variable selection or similar techniques for variable reduction have only been applied once before in the context of market basket analysis (Hruschka 1991)[2]. In all other publications, the problem of parameter abundance has been tackled with a priori selection of a small number of categories, which could lead to biased estimates of cross category effects (Chib et al. 2002).

Given the high number of subsets of cross category effects equal to $2^{J(J-1)/2}$, it is obvious that an examination of every possible model is tedious and may even be infeasible. George and McCulloch (1993) propose stochastic search variable selection (SSVS) for such a situation, which avoids the calculation of the posterior probability of all models. Instead, SVSS suggests only more "promising" variable subsets with higher posterior probability.

We compare three different Bayesian approaches to variable selection appropriate for binary logit models. We use these variable selection approaches because the conditional purchase probabilities of each category j given purchases of other categories $k \neq j$ have a binary logit form for the MVL model (see expression 3). All three algorithms provide a vector with posterior coefficient estimates and a vector with probabilities that a coefficient is different from zero. Two of these algorithms have been applied successfully for binary logit models before, but the number of predictors was much lower than in our market

[2] Hruschka (1991) applied a model selection method based on the Marquardt algorithm that deletes interaction effects if they are determined as insignificant by likelihood ratio tests.

basket analysis study. The third algorithm is a modification of a variable selection method for linear regression.

3.1 Algorithm of Groenewald and Mokgatlhe (A1)

We choose the algorithm of Groenewald and Mokgatlhe (2005) because of its simple sampling scheme for coefficients and its forecast robustness and accuracy in tests on smaller data sets. This algorithm works with Bayes factors. The current model is named M_t with coefficient vector $\beta_j^\delta = (\alpha_j, \theta_{jk}^\delta)$ with category constant and $p^\delta - 1$ included cross category effects. Accordingly, each model M_t has a binary indicator vector δ_t of length $J(J-1)/2$ for coefficient inclusion. The marginal likelihood of a model M_t for all purchases in category j, i.e., Y_j, can be written as

$$m(Y_j|M_t) = L(\beta_j^\delta|Y_j, M_t)\pi(\beta_j^\delta, \sigma_j)\pi(\sigma_j)/\pi(\beta_j^\delta, \sigma_j|Y_j, M_t) \tag{7}$$

with scale parameter σ_j, the prior on parameters $\pi(\beta_j^\delta, \sigma_j)$ and the likelihood function $L(\beta_j^\delta|Y_j, M_t)$.

The intractable posterior likelihood, i.e., the denominator of the marginal likelihood, is calculated by introducing latent variables (Tanner and Wong 1987) and applying Gibbs sampling steps as proposed by Chib (1995) to the conditional probability components of the posterior density

$$\pi(\beta_j^\delta, \sigma_j|Y_j, M_t) = \pi(\alpha_j|Y_j, M_t)\,\pi(\theta_{j1}^\delta|\alpha_j, Y_j, M_t)\,...\,\pi(\sigma_j|\beta_j^\delta, Y_j, M_t) \tag{8}$$

Posterior coefficient values for category constant and interaction effects are computed by drawing from uniform distributions within a second Gibbs cycle. A single coefficient value $\beta_j = (\alpha_j, \theta_{jk})$ is sampled as follows:

$$\begin{aligned}
&\beta_{jk} = -\sigma_j ln((1-\upsilon_{jk})/\upsilon_{jk}) \\
&\text{with} \\
&\upsilon_{jk}|a_{jk}, b_{jk}, \sigma_j \\
&\sim U(\exp(a_{jk}/\sigma_j)/(1+\exp(a_{jk}/\sigma_j)), \exp(b_{jk}/\sigma_j)/(1+\exp(b_{jk}/\sigma_j)) \\
&a_{jk} = max_{i\in A_{jk}}[\overline{Y}_{ik}^{-1}\log(U(0,1)/(1-U(0,1))) - \sum_{k'\neq k}\beta\overline{Y}_{ik'})] \\
&b_{jk} = min_{i\in B_{jk}}[\overline{Y}_{ik}^{-1}\log(U(0,1)/(1-U(0,1))) - \sum_{k'\neq k}\beta\overline{Y}_{ik'})] \\
&A_{jk} = i : ((Y_{ij}=1)\cap(\overline{Y}_{ik}>0)) \cup ((Y_{ij}=0)\cap(\overline{Y}_{ik}<0)), \\
&B_{jk} = i : ((Y_{ij}=0)\cap(\overline{Y}_{ik}>0)) \cup ((Y_{ij}=1)\cap(\overline{Y}_{ik}<0))
\end{aligned} \tag{9, 10}$$

$U(u_1, u_2)$ denotes a random number uniformly distributed over the interval $[u_1, u_2]$.

Scale parameters σ_j are drawn from the following distribution:

$$\pi(\sigma_j|\beta_j^\delta) \propto \sigma_j^{-p^\delta-2}\,\exp(\sum\beta_{jk}/\sigma_j)/\prod(1+\exp(\beta_{jk}/\sigma_j))^2 \tag{11}$$

Averaged over the Gibbs sampling steps, estimates are used to calculate the numerator of $m(Y_j|M_t)$. Marginal likelihoods are calculated for models including and excluding each single cross category coefficient θ_{jk}. The evidence of the respective Bayes factor for the simpler null model (exclusion of θ_{jk}) is evaluated according to the guidelines of Jeffreys (1961) which favor simpler models, as suggested by Gill (2002). This result is put into the respective position of indicator vector δ_t.

3.2 Algorithm of Tüchler and Scott (A2)

We also test the algorithm of Tüchler (2008) developed as variable selection technique for logit models. It is based upon the concept of SSVS promising higher efficiency compared to algorithm A1 and only samples from standard distributions. The fundamental idea of SSVS is to derive a binary indicator vector δ with $J(J-1)/2-p^\delta$ zeros and p^δ ones. If an element of δ is 1, the respective coefficient is left in the model, otherwise it is eliminated.

By means of data augmentation (Tanner and Wong 1987), stochastic utility values $\tilde{Y}_{ij}$ for purchase or non-purchase of category j are introduced as latent variables in analogy to the utility maximization concept of McFadden (1974). Drawing two uniform random numbers $U_1 = U(0,1)$ and $U_2 = U(0,1)$, latent stochastic utilities are sampled as follows:

$$\tilde{Y}_{ij} = -\log(-\log U_1/(1+\exp(V_{i,j|k})) - \log U_2/\exp(V_{i,j|k})\,(1-Y_{ij})) \tag{12}$$

$V_{i,j|k} = \alpha_j + \sum_k \theta_{jk} Y_{ik}$ and k runs over the $p^\delta - 1$ interacting coefficients different from zero only.

The logit problem with a binary dependent variable Y_{ij} is transformed into a linear regression with Gumbel distributed error terms ϵ_i being approximated by a mixture of normal distributions (cf. Frühwirth-Schnatter and Frühwirth 2007). For the mixture approximation, every market basket is assigned to one of $r = 1, ..., 10$ normal distributions with specific mean m_r and variance s_r^2.

Indicators are sampled by a subalgorithm of Smith and Kohn (2002) using conditional priors for the indicators and marginal likelihoods $p(\tilde{Y}|\delta, R)$ with respect to the reduced coefficient vector β^δ and with utilities vector $\tilde{Y}$, indicators δ, and index of the assigned mixture component R with mean vector $m = (m'_{ri})$ and covariance matrix $\Sigma = diag(s^2_{ri})$. As estimation uses the reduced form of the coefficient vector β^δ, the market basket matrix is adapted accordingly, which is symbolized by Y^δ.

The p^δ coefficients different from zero are sampled from the normal distribution

$$\begin{aligned} &p(\beta^\delta|\tilde{Y}, R) \sim N(c, C) \\ &\text{with} \quad c = CY^\delta\beta^\delta\Sigma^{-1}(\tilde{Y}-m) \quad \text{and} \quad C^{-1} = (Y^\delta)'\Sigma^{-1}Y^\delta \end{aligned} \tag{13}$$

in one step. New coefficient values are sampled by a Metropolis-Hastings step (Scott 2006).

3.3 Algorithm of Geweke (A3)

We adapt an algorithm of Geweke (2005) developed for linear regression to logit models by introducing and sampling latent utilities the same way as in algorithm A2. The linear

regression version of this algorithm proved to be stable and efficient in applications. It also exactly discriminated relevant against irrelevant predictors. Another advantage of this algorithm is the possibility to truncate values of coefficients. Prior values indicated by an underline are set for β, error precision h, null-probability of coefficient j ρ_j and degrees of freedom ν. The starting point for estimation is a model M_t with a specific subset of coefficients $k = 1, ..., p^{\delta}$. Assuming a priori independence of coefficients, the probability $\rho_j = p(\beta_j = 0|\beta_k(k \neq j), Y, M_t, h)$ conditional on the other coefficients currently in model M_t is calculated. Derived from the conditional posterior distribution $p(\beta_j|\beta_k(k \neq j), Y, M_t, h)$, ρ_j is proportional to $\underline{p}_j \exp(-h\sum_{i=1}^{I} z_i^2/2)$ with $z_i = \tilde{Y}_{ij} - \sum_{j \neq k} \beta_j Y_{jk}$. If this probability $p(\beta_j = 0)$ is smaller than a random uniform number $U(0,1)$, the truncated value of β_j and the error precision h are sampled as follows:

$$\beta_j \sim N(\overline{\beta}_j, \overline{h}_j^{-1}) \tag{14}$$

with

$$\overline{h}_j = \underline{h}_j + h\sum_{i=1}^{I} Y_{ij}^2, \quad \overline{\beta}_j = \overline{h}_j^{-1}(\underline{h}_j\underline{\beta}_j + h\sum_{i=1}^{I} Y_{ij} z_i)$$

$$h \sim \chi^2(I + \underline{\nu})/(sse + \underline{s}^2) \tag{15}$$

β and h are sampled within a Gibbs cycle in which coefficient β_j is conditioned on the other coefficients β_k and error precision h depends on the sum of squared residuals sse given the sampled constant and interaction effects.

4 Empirical Study

4.1 Data

20,000 market baskets collected at a supermarket in Bavaria are randomly split into two data sets of equal size. One set (estimation data) is required for estimation, the second set (validation data) is used to determine the predictive accuracy of MVL models. From all 209 categories in the original data, we only use the 30 categories purchased most frequently.[3] Basket size, which is the number of categories contained in one basket, ranges between 1 and 19. Average basket size is 3.99 for the estimation data, and 4.01 for the validation data. Column 3 and 4 of table 2 show the categories considered together with their purchase frequencies.

4.2 Comparison of algorithms

Our goal is to study the suitability of the three variable selection algorithms described in section 3 for market basket analysis, primarily w.r.t. the ability to uncover significant cross category effects but also w.r.t. predictive accuracy and computation times for estimation. We measure predictive accuracy by cross-validated pseudo loglikelihood values

[3] We decide to analyze a smaller number of categories to ensure a clear presentation of results.

Table 2: Data Description and Estimated Category Constants

Number	Abbreviation	Category Name	Purchase Frequency	α_j (A1)	α_j (A2)	α_j (A3)
1	FRU	Fruit	3141 (3099)	-1.067	-1.535	-2.079
2	BRE	Bread	3098 (3078)	-0.974	-1.452	-1.719
3	VEG	Vegetables	2547 (2599)	-1.349	-1.445	-2.445
4	MAG	Magazines	2151 (2092)	-1.537	-1.296	-1.732
5	YOG	Yoghurt & Curd	2134 (2194)	-1.554	-1.779	-2.650
6	MIL	Milk	1907 (1971)	-1.721	-1.786	-2.781
7	CHO	Chocolate	1497 (1545)	-1.903	-1.716	-2.401
8	SOF	Soft Drinks	1469 (1492)	-1.860	-1.613	-2.049
9	BEE	Beer	1423 (1389)	-1.938	-1.581	-2.027
10	CIG	Cigarettes	1395 (1439)	-1.935	-1.750	-2.126
11	CHE	Cheese	1286 (1225)	-2.168	-1.907	-3.273
12	JUI	Juice	1280 (1342)	-1.407	-2.045	-2.672
13	BUT	Butter	1250 (1258)	-2.270	-1.989	-3.548
14	UHT	UHT Milk	1087 (1112)	-2.324	-2.127	-3.268
15	FAT	Fat & Oil	1055 (1121)	-2.437	-1.995	-3.447
16	SOU	Soups & Sauces	1048 (1015)	-2.444	-2.448	-3.373
17	TIN	Tinned Sour Food	1041 (1056)	-2.411	-2.074	-3.535
18	WAT	Water	1024 (1010)	-2.322	-1.623	-2.209
19	SPI	Spices & Mustard	965 (896)	-2.435	-2.106	-3.112
20	CUT	Cut Cheese	955 (1077)	-2.551	-2.049	-3.801
21	SWE	Sweets	940 (898)	-2.350	-2.439	-2.938
22	SEA	Seasonal Items	937 (923)	-2.418	-1.999	-2.954
23	BAK	Baking Ingredients	905 (992)	-2.619	-2.221	-3.335
24	ROL	Rolls	809 (778)	-2.517	-2.363	-3.144
25	SNA	Snacks & Crisps	801 (786)	-2.570	-2.581	-3.235
26	FOI	Foil & Plastic Bags	798 (720)	-2.579	-2.305	-3.037
27	COF	Coffee	775 (781)	-2.659	-2.798	-3.231
28	PAS	Pasta	724 (723)	-2.863	-2.475	-3.720
29	TRU	Truffles	713 (738)	-2.664	-2.542	-3.089
30	HYG	Hygiene Articles	699 (707)	-2.679	-2.410	-3.390

(CV-PLL), i.e., PLL values of models applied to the validation data after estimation. The PLL value for the model consisting of constants only is -112,519.76 (estimated constants of this model equal the respective log odds, i.e., logarithms of ratios of the relative purchase frequencies and relative non-purchase frequencies, for the estimation data), its CV-PLL value amounts to -112,891.57.

Table 3: Performance and Efficiency Measures

	Algorithm 1 Groenewald	Algorithm 2 Tüchler	Algorithm 3 Geweke
Duration	384.32h	54.9h	2.4h
PLL	-103,086.35	-107,419.83	-100,162.02
CV-PLL	-103,916.04	-107,921.22	-101,329.06
Included Interactions	74	148	151

All three variable selection algorithms converge quickly. The number of burn-in and saved iterations as well as the appropriate amount of chain thinning is determined individually for every algorithm to ensure a comparably good adaptation to the data. Our requirements for inclusion of coefficients are rather strict (average exclusion probability $\overline{\rho} < 0.1$, indicator average over iterations $\overline{\delta} > 0.9$, absolute value of coefficient $|\theta_{jk}| > 0.1$). All estimated models turn out to be robust as CV-PLL values demonstrate. Computation times vary between two extremes (see table 3). Computing times for A1 are very high and increase strongly with the number of categories considered.

A3 achieves the largest improvement of PLL, followed by A1, whereas improvement attained by A2 is rather modest. A1 includes approximately half the number of cross category effects of A2 or A3. Therefore, comparing A1 to its competitors may be considered unfair. Relaxing the inclusion probability from .9 to .5 and the absolute value of $|\theta_{jk}| > 0.1$ to $|\theta_{jk}| > 0.045$ in A1 results in a a model with 150 interaction effects. This enlarged model leads to PLL and CV-PLL values of -100,788.59 and -101,741.87, respectively, which are close to the values obtained by A3.

There is some variation of the relative sizes of constants due to their dependency on the number and the magnitude of included interaction effects (see table 2 columns 5 to 7). With regard to the five largest cross category effects, there is a remarkable overlap between algorithms (see table 4 for category pairs in descending order of interaction coefficients).

Using absolute values of cross category coefficients as proximities, we provide MDS graphics (see figure 1, created with SPSS Proxscal). These graphics reveal similar clusters of categories for the three selection algorithms. Categories of daily nutrition, such as milk, bread, fruit, vegetables, yogurt, etc., have large cross-category effects and interact with many other categories. Within this broad cluster, more subclusters can be identified: fresh produce (milk, butter, vegetables, cheese) as well as bread, rolls, and cut cheese or soups/sauces, fat/oil and pasta interact heavily. Beverage categories (i.e., water, beer, soft drinks) interact highly, but show weak interactions with the remaining

Table 4: Five Largest Cross Category Effects

A1	A2	A3
Cut Cheese and Bread	Pasta and Soups & Sauces	Cut Cheese and Bread
Beer and Water	Cut Cheese and Bread	Pasta and Soups & Sauces
Milk and Yogurt & Curd	Fruit and Vegetables	Beer and Water
Beer and Soft Drinks	Chocolate and Truffles	Baking Ingred. and Fat & Oil
Pasta and Soups & Sauces	Coffee and Foil & Plastic Bags	Fruit and Vegetables

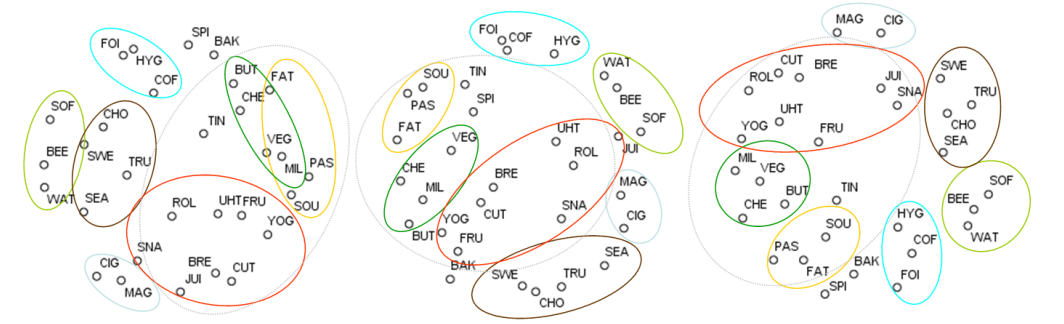

Figure 1: MDS graphics based on A1, A2 and A3

assortment. Magazines are independent from the remaining assortment with the exception of cigarettes. There exists a strong connection between categories in the candy category which could be caused by proximity of shelves. Interestingly, no algorithm finds any category that is completely independent of the other categories.

Equations with estimated coefficients show to what extent selection algorithms provide similar or different results on interactions. As examples, we choose the categories fruit, chocolate, beer, and pasta (see table 5). All algorithms reveal strong positive interactions equally well, less pronounced interactions are missed by A1 and in a few cases by A2. Differences between the algorithms are most striking for substitutive interactions. In the beer category, for example, A1 does not detect any negative interaction. A2 and A3, on the other hand, find substitutive effects but attribute it to different categories.

Forecasting accuracy of A1 is high, but A1 does not perform well in terms of computation times. This drawback of A1 will intensify, if covariates (e.g., price, promotions) are added. High computation times also rule out using A1 as component of an extended model with latent heterogeneity. Another weakness of this algorithm is its tendency to underestimate interaction effects which is to a large degree due to the high number of inclusion probabilities in the range between 20% and 90% (see figure 2)[4].

Computation times of A2 are acceptable, but A2 is clearly inferior to A1 and A3 in

[4]Mean delta and mean rho respectively are computed as average over all sampled values. Graphics include indicators for constants.

Table 5: Coefficients for fruit, chocolate, beer, and pasta

	Fruit			Chocolate			Beer			Pasta		
	A1	**A2**	**A3**	**A1**	**A2**	**A3**	**A1**	**A2**	**A3**	**A1**	**A2**	**A3**
FRU	-1.067	-1.535	-2.079	.147	.465	.381						
BRE	.139	.238	.346		.113						.126	
VEG		.990	.987							.284	.487	.648
MAG								-.110				
YOG	.291	1.034	.670		.120	.184						
MIL	.202	.341	.487		.103					.150	.407	.567
CHO	.147	.465	.381	-1.903	-1.716	-2.401			-.354			
SOF							.323	.545	.920			
BEE						-.354	-1.903	-1.581	-2.027			
CIG						.260						
CHE	.138	.242	.364								.220	.432
JUI	.105		.331					.148	.337			
BUT	.140	.365	.377			.306						.334
UHT	.148		.442									
FAT	.109										.339	.567
SOU	.164	.521	.385			.277				.322	1.094	1.235
TIN	.195		.509									.407
WAT			.224				.396	.852	1.191			
SPI								-.112				.404
CUT	.183	.135	.480									
SWE		.560	.368	.258	.881	.865						
SEA	.160	.183	.526	.240	.129	.664			-.316			
BAK		.484	.375	.157	.414	.531			-.545			
ROL			.323									
SNA					.364	.493						.456
FOI												.390
COF												
PAS										-2.863	-2.475	-3.720
TRU		.231			.976	.850						
HYG			.542									

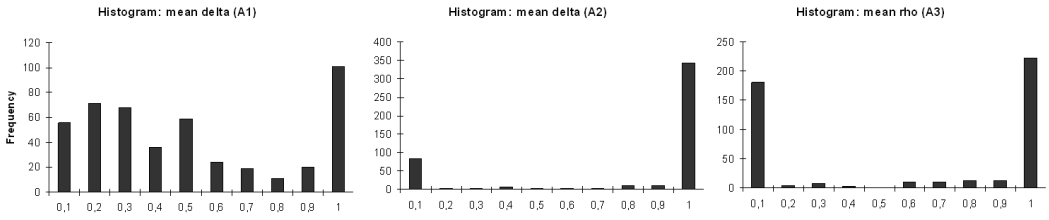

Figure 2: Histograms of inclusion/ exclusion probabilities

terms of PLL values. Figure 2 shows that A2 fails to exclude insignificant effects[5] and consequently results in many very small interaction effects ($|\theta_{jk}| < 0.1$).

A3 accomplishes the best overall performance, both in terms of computation time and PLL values. Parameter exclusion probabilities ρ have high discriminative power (see figure 2). W.r.t. coefficients, estimation is very accurate, and truncation prevents the increase of coefficients. Conditioning each coefficient on the other coefficients does not slow down estimation, as suspected by Geweke (2005). Taking all these factors into account, we propose to use A3 for market basket analysis. Accordingly, the rest of our paper discusses results obtained by A3.

4.3 Results of Algorithm A3

Contrary to Chib et al. (2002) or Russell and Petersen (2000), who analyze 12 and 4 categories, respectively, we do not find all possible cross category effects to be significantly different from zero. Our result that 34.5% of these effects are significant agrees to some extent with the only comparable publication (Hruschka et al. 1999). Hruschka et al. report only 4.9% significant interactions for 73 categories many of which have very low relative purchase frequencies. Please note that such low-frequency categories are not considered in our study.

The large increase of PLL values of our model over the model which only contains constants demonstrates that cross category coefficients are important for the explanation of purchase probabilities. Interaction effects obtained are smaller compared to several studies whose MVL models consider a small number of categories (e.g., Boztuğ and Hildebrandt 2008; Boztuğ and Reutterer 2008; Russell and Petersen 2000) and more in line with Chib et al. (2002).

Our results agree with Hruschka et al. (1999) and Chib et al. (2002). Positivity of most significant interaction effects corroborates the hypothesis of general complementarity among all categories in the assortment, e.g., due to one-stop-shopping. Still, some negative correlations are revealed, e.g., baking ingredients and cigarettes, baking ingre-

[5] In this case, A2 includes around 70% of all interactions. Recall that we additionally exclude $|\theta| < 0.1$ for our analysis reducing the number of effects by half. This reduction is justified, as the contribution of smaller effects to the PL value is negligible.

dients and water, water and truffles, soups & sauces and beer, beer and seasonal items, water and hygiene products or chocolate and beer.

Chib et al. (2002) argue that considering only a subset of categories induces underestimation of values of interaction effects, even signs might change from positive to negative. Though we already model far more categories than Chib et al., we investigate their hypothesis by expanding our data set to the 45 most often purchased categories[6] and estimate coefficients by A3 to explore possible increases or decreases of the interaction effects caused by the number of included categories. We also examine whether we obtain negative interaction coefficients if we limit our data set to the 15 most often purchased categories[7]. Results for the estimation data are reported in table 6.

Table 6: Variation of Number of Categories Included in the Model

Categories	PL	Basket Size	Complementary	Independent	Substitutive
15	-61,213.82	2.67	52 (49.5%)	51 (48.6%)	2 (1.9%)
30	-100,162.02	3.99	141 (32.4%)	284 (65.3%)	10 (2.3%)
45	-131,555.53	4.84	188 (19.0%)	794 (80.2%)	8 (0.8%)

The 51 interaction coefficients determined as insignificant considering 15 categories are also insignificant in the 30 categories case. Contrary to the underestimation hypothesis of Chib et al., the two substitutive effects do not become positive, but stay negative in the 30 categories case. The majority of constants and all significant positive cross category coefficients are larger for 15 categories compared to the 30 categories model - except for the constant of the cigarettes category- what might be caused by the lower number of cross category effects. Complementarity is found between seven category pairs that are independent relations in the 30 category case, e.g., UHT milk and juice. These results clearly contradict the underestimation hypothesis.

Similar conclusions are drawn from the comparison of the estimation with 30 categories to the estimation with 45 categories. Independent pairs for the 30 categories estimation are replicated for the 45 categories case. As a weak support of the underestimation hypothesis, only six of the ten negative interactions from the 30 categories case are identified as substitutive in the 45 categories case. However, 39 of the 141 positive interactions discovered in the 30 categories set are estimated as independent in the 45 categories set, i.e., they are overestimated in the reduced set. Surprisingly, positive interaction estimates which are significant in both data sets are smaller for the 30 categories data set.

To summarize, reducing the number of analyzed categories leads to biased estimates. However, no extreme switches from negative to positive or vice versa could be observed. Generally, the percentage of independent category pairs increases with the number of

[6] The additional categories are sugar, delicatessen, tinned vegetables, tinned fish, eggs, condensed milk, wholewheat bread, zwieback, sparkling wine, toilet paper, personal hygiene items, oral hygiene items, hair care products, cat food, gifts & candles. Purchase frequencies range from 460 (sparkling wine) to 3141 (fruit).

[7] These are fat, milk, yogurt, cheese, butter, UHT milk, bread, chocolate, cigarettes, beer, soft drinks, juice, fruit, vegetables, and magazines. For purchase frequencies, see table 2.

categories in the model due to less overestimated coefficients and more categories with low purchase frequencies.

5 Conclusions and Future Research

We use variable selection techniques to explore the cross category effects of a supermarket assortment within the framework of a MVL model. We test three variable selection techniques of which only an adaptation of an algorithm of Geweke (2005) meets the requirements of market basket analysis. We find that explanatory approaches that consider only few categories result in biased cross category effects. We conclude that the incorporation of the most important categories witin an assortment into a model is essential to obtain less biased parameters. One advantage of our model, especially in contrast to traditional exploratory methods, is the obvious way in which segmentation or covariates, such as marketing-mix data or customer demographics, may be integrated.

For reasons of simplicity and clarity we did not implement price and promotion covariates so far. However, their inclusion is straight forward: category constants and interaction effects are split into a promotion, a price and a category component. This enables the differentiation between purchase and consumption complementarity explaining consumer purchase behavior in a more detailed way (see, e.g., Hruschka et al. 1999 or Russell and Petersen 2000).

It is not clear how the assumed customer homogeneity influences the magnitude of the interaction effects. It might lead to a decrease as category interactions might have different values and even opposed signs in the various segments. Chib et al. (2002) quite contrary find that a disregard of unobserved heterogeneity leads to overestimated cross category effects. To answer this question, a finite mixture extension of the MVL model could turn out to be useful.

References

[1] Bell DR, Lattin JM (1998) Shopping Behavior and Consumer Preference for Store Price Format: Why "Large Basket" Shoppers Prefer EDLP. Marketing Sci 17:66–88

[2] Besag J (1974) Spatial Interaction and the Statistical Analysis of Lattice Systems. J R Stat Soc Ser B 36:192–236

[3] Besag J (1975) Statistical Analysis of Non-Lattice Data. J R Stat Soc Ser D (Statistician) 24:179–195

[4] Betancourt R, Gautschi D (1990) Demand Complementarities, Household Production, and Retail Assortments. Marketing Sci 9:146–161

[5] Boztuğ Y, Hildebrandt L (2007). Ansätze zur Warenkorbanalyse im Handel. In: Schuckel M, Toporowski W (eds) Theoretische Fundierung und praktische Relevanz der Handelsforschung. DUV Gabler, Wiesbaden, 218–233

[6] Boztuğ Y, Hildebrandt L (2008) Modeling Joint Purchases with a Multivariate MNL Approach. Schmalenbach Bus Rev 60:400–422

[7] Boztuğ Y, Reutterer T (2008) A Combined Approach for Segment-Specific Market Basket Analysis. Eur J Oper Res 187:294–312

[8] Boztuğ Y, Silberhorn N (2006) Modellierungsansätze in der Warenkorbanalyse im Überblick. J Betriebswirtschaft 56:105–128

[9] Buchta C (2007) Improving the Probabilistic Modeling of Market Basket Data. In: Decker R, Lenz HJ (eds) Advances in Data Analysis. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, 417–424

[10] Bucklin RE, Gupta S, Siddarth S (1998) Determining Segmentation in Sales Response across Consumer Purchase Behaviors. J Marketing Res 35:189–197

[11] Chib S (1995) Marginal Likelihood from the Gibbs Output. J Am Stat Assoc 90:1313–1321

[12] 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) Advances in Econometrics 16. Econometric Models in Marketing. JAI, Amsterdam, 57–92

[13] Cox DR (1972) The Analysis of Multivariate Binary Data. J R Stat Soc Ser C (Appl Stat) 21:113–120

[14] Cressie NAC (1993) Statistics for Spatial Data. Revised Edition. John Wiley & Sons Inc, New York

[15] Duvvuri SD, Ansari A, Gupta S (2007) Consumers' Price Sensitivities Across Complementary Categories. Manag Sci 53:1933–1945

[16] Erdem T (1998) An Empirical Analysis of Umbrella Branding. J Marketing Res 35:339–351

[17] Frühwirth–Schnatter S, Frühwirth R (2007) Auxiliary Mixture Sampling with Applications to Logistic Models. Comp Stat Data Anal 51:3509–3528

[18] George EI, McCulloch R (1993) Variable Selection via Gibbs Sampling. J Am Stat Assoc 88:881–889

[19] Geweke J (2005) Contemporary Bayesian Econometrics and Statistics. John Wiley & Sons Inc, Hoboken (NJ)

[20] Gill J (2002) Bayesian Methods. A Social and Behavioral Sciences Approach. Chapman & Hall/CRC, Boca Raton (FL)

[21] Groenewald PCN, Mokgatlhe L, Bayesian (2005) Computation for Logistic Regression. Comp Stat Data Anal 48:857–868

[22] Hruschka H (1985) Der Zusammenhang zwischen paarweisen Verbundbeziehungen und Kaufakt- bzw. Käuferstrukturmerkmalen. zfbf Z betriebswirtschaftliche Forsch 37:218–231

[23] Hruschka H (1991) Bestimmung der Kaufverbundenheit mit Hilfe eines probabilistischen Meßmodells. zfbf Z betriebswirtschaftliche Forsch 43:418–434

[24] Hruschka H, Lukanowicz M, Buchta C (1999) Cross-Category Sales Promotion Effects. J Retail Consum Serv 6:99–105

[25] Jeffreys H (1961) Theory of Probability, 3rd edition. Oxford University Press, Oxford

[26] Magnussen S, Reeves R (2007) Sample-based Maximum Likelihood Estimation of the Autologistic Model. J Appl Stat 34:547–561

[27] Manchanda P, Ansari A, Gupta S (1999) The "Shopping Basket": A Model for Multi-Category Purchase Incidence Decisions. Marketing Sci 18:95–114

[28] McFadden D (1974) Conditional Logit Analysis of Qualitative Choice Behavior. In: Zarembka P (ed), Frontiers in Econometrics. Academic Press, Inc., New York, 105–142

[29] Mild A, Reutterer T (2003) An Improved Collaborative Filtering Approach for Predicting Cross-Category Purchases Based on Binary Market Basket Data, J Retail Consum Serv 10:123–133

[30] Moon S, Russell GJ (2004) Spatial Choice Models for Product Recommendations, Working Paper, University of Iowa

[31] Murray I, Ghahramani Z (2004) Bayesian Learning in Undirected Graphical Models: Approximate MCMC Algorithms. ACM International Conference Proceeding Series 70, Proceedings of the 20th conference on Uncertainty in artificial intelligence. AUAI Press, Banff, Canada, 392–399

[32] Mulhern FJ, Leone RP (1991) Implicit Price Bundling of Retail Products: A Multiproduct Approach to Maximizing Store Profitability. J Marketing 55:63–76

[33] Niraj R, Padmanabhan V, Seetharaman PB (2008) Research Note: A Cross-Category Model of Households' Incidence and Quantity Decisions. Marketing Sci 27:225–235

[34] Russell GJ, Bell D, Bodapati A, Brown CL, Chiang J, Gaeth G, Gupta S, Manchanda P (1997) Perspectives on Multiple Category Choice. Marketing Lett 8:297–305

[35] Russell GJ, Petersen A (2000) Analysis of Cross Category Dependence in Market Basket Selection. J Retail 76:369–392

[36] Russell GJ, Ratneshwar S, Shocker AD, Bell D, Bodapati A, Degeratu A, Hildebrandt L, Kim N, Ramaswami S, Shankar VH (1999) Multiple Category Decision-Making: Review and Synthesis. Marketing Lett 10:319–332

[37] Scott SL (2006) Data Augmentation, Frequentist Estimation, and the Bayesian Analysis of Multinomial Logit Models. Working Paper, University of Southern California

[38] Sherman M, Apanasovich TV, Carroll RJ (2006) On Estimation in Binary Autologistic Spatial Models. J Stat Comput Simul 76:167–179

[39] Shocker AD, Bayus BL, Kim N (2004) Product Complements and Substitutes in the Real World. The Relevance of "Other Products". J Marketing 68:28–40

[40] Smith M, Kohn R (2002) Parsimonious Covariance Matrix Estimation for Longitudinal Data. J Am Stat Assoc 97:1141–1153

[41] Tanner MA, Wong WH (1987) The Calculation of Posterior Distributions by Data Augmentation. J Am Stat Assoc 82:528–540

[42] Tüchler R (2008) Bayesian Variable Selection for Logistic Models Using Auxiliary Mixture Sampling. J Comput Graph Stat 17:76–94

[43] Wang J, Liu J, Li SZ (2000) MRF parameter estimation by MCMC method. Pattern Recognit 33:1919–1925

[44] Ward MD, Gleditsch KS (2002) Location, Location, Location: An MCMC Approach to Modeling the Spatial Context of War and Peace. Political Anal 10:244–260

[45] Yu Y, Cheng Q (2003) MRF Parameter Estimation by an Accelerated Method. Pattern Recognit Lett 24:1251–1259