

SFB 649 Discussion Paper 2005-028

A Market Basket Analysis Conducted with a Multivariate Logit Model

Yasemin Boztuğ*
Lutz Hildebrandt*



*Institute of Marketing, Humboldt-Universität zu Berlin, Germany

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

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

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



SFB 649 ECONOMIC RISK BERLIN

A Market Basket Analysis Conducted with a Multivariate Logit Model*

Yasemin Boztuğ and Lutz Hildebrandt

Institute of Marketing, Humboldt–University Berlin,
Spandauer Str. 1, D-10178 Berlin, Germany

Abstract. The following research is guided by the hypothesis that products chosen on a shopping trip in a supermarket can indicate the preference interdependencies between different products or brands. The bundle chosen on the trip can be regarded as the result of a global utility function. More specifically: the existence of such a function implies a cross-category dependence of brand choice behavior. It is hypothesized that the global utility function related to a product bundle results from the marketing-mix of the underlying brands. Several approaches exist to describe the choice of specific categories from a set of many alternatives. The models are discussed in brief; the multivariate logit approach is used to estimate a model with a German data set.

1 Introduction

One of the major tasks of retailers is managing their product categories to maximize the overall profit of the store or chain. Using marketing mix strategies to stimulate purchases for a specific product usually has an effect both on the advertised category and related categories. Additionally, a retailer decides not only to advertise for one category, but in many simultaneously. Thus the retailer must consider cross effects between linked or related categories in their marketing measures. Ignoring dependency structures could lead to wrong decisions or at least to suboptimal marketing–mix activities. Analyzing multi–item purchases is not only of interest for the researcher, but also from a managerial point of view.

In the following, we will focus on the analysis of bundle purchases. It belongs to a ”pick–any”–choice problem (Levine (1979)), because the consumer can choose no item, one item or any possible number of items for his shopping bag. Common brand choice models, like the well known multinomial logit approach (MNL) (Guadagni and Little (1983)), consider only single category purchases and ignore cross–category relationships and influences. This could lead to wrong parameter estimates and therefore to wrong decisions for using marketing–mix activities.

* Financial support by the German Research Foundation (DFG) through the research project #BO 1952/1, and through the Sonderforschungsbereich 649 is gratefully acknowledged.

The model used in this article is based on an approach by Russell and Petersen (2000). It predicts category incidence and examines how a purchase in one category is affected by other category purchases. We assume a global utility function, which argues that the cross-category choice dependence is present within each choice process of each consumer. The modeling should include purchases conditional on purchases in other categories during the same shopping trip. Assuming such a dependence structure means that common estimation techniques cannot be used anymore because they are not able to cope for dependent observations. Instead, techniques from spatial statistics are needed to estimate the market basket model in a proper way.

The article is structured as follows. In the next section, we will describe market basket models in general, and explain our model more in detail. Afterwards, a data set is presented along with subsequent estimation results. The article concludes with a summary and an outlook.

2 Market basket models

Market baskets arise due to shopping behavior of customers. During a shopping trip, the customers are in a "pick-any"–situation because they have the possibility to choose no item, one or any other number of items in each category. Standard brand choice models, as the MNL, focus on purchases taken in one specific category, ignore cross-effects to other categories, and produce possibly biased parameter estimates.

A number of research articles started to incorporate cross-category relationships in their purchase models (see e.g., Russell et al. (1997, 1999), Seetharaman et al. (2004)). Two main research approaches can be distinguished. One is more data-driven using data-mining. It is dominated by techniques like pairwise association (e.g., Hruschka (1985)), association rules (e.g., Agrawal and Srikant (1994)), vector quantisation (e.g., Schnedlitz et al. (2001)), neural networks (e.g., Decker and Monien (2003)) and collaborative filtering (e.g., Mild and Reutterer (2001, 2003)).

Pairwise associations use simple association measures to indicate coincidence or affinity of items in market baskets to identify product category relationships. Often techniques of multidimensional scaling or cluster analysis are applied to first reduce the large number of categories. The association rules are then used to group subsets of product categories together. Vector quantisation is a more sophisticated method, which enriches the data with an additional basket vector. This vector contains information about the membership of a specific category to a subbasket class. Using neural networks for market basket analysis is related to vector quantisation. First, an affiliation to a subgroup is identified. Collaborative filtering then uses databases to identify those customers, who behave similar to the target customer and to make predictions using these similarities.

The second research approach is more explanatory driven. It tries to identify and quantify cross-category choice effects of marketing-mix variables. Here, two general methods can be identified. The multivariate probit approach (e.g., Ainslie and Rossi (1998), Manchanda et al. (1999), Seetharaman et al. (1999), Chib et al. (2002), Deepak et al. (2004)) is an extension of the standard probit approach (e.g., Hausmann and Wise (1978), Daganzo (1979), Train (2003)) for one category. It is based on Random Utility Theory and is built on a disaggregate level. The error distribution is assumed to be normal. Alternatively, the multivariate logit approach (e.g., Hruschka et al. (1999), Russell and Petersen (2000), Hansen et al. (2003), Singh et al. (2004)) can be used, which is an extension of the multinomial logit model (e.g., Guadagni and Little (1983)). It is also based on Random Utility Theory. The error term of the multivariate logit approach is assumed to be Gumbel distributed.

In our approach, adapted from Russell and Petersen (2000), we use a multivariate logit model to analyse multi-item purchases. The approach models purchase incidence and is related to the well established MNL-models. It is much easier to estimate than the multivariate probit approach. The estimation routine can be programmed with standard software modules, and the approach allows the inclusion of several marketing-mix variables. Complementary, independence and substitution of product categories can be modeled.

In our model, we assume that consumers make their category choices in some fixed order, which is not observed by the researcher. Due to this lack of information, the choice in each category is modeled conditional upon known choices in all other categories. It is assumed that the choices are made in a certain order, but it is not necessary to know this order for model construction. To estimate such a model in an unbiased way, we need to apply techniques from spatial statistics to account for relationship of dependence between the categories. With these methods, we are able to describe the conditioned observations without having any information about the concrete purchase sequence. The complete set of full conditional distributions uniquely determines the joint distribution (Besag (1974), Cressie (1993)).

Our market basket model accounts for purchases at the category level. The whole bundle description consists of zeros and ones for the existence or absence of category items in the basket. The joint distribution describing the whole basket is inferred from the full conditional distribution of the single category models which have the following form

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

The utility in Equation (1) is specified as follows

$$\begin{aligned} U(i, k, t) &= \beta_i + \text{HH}_{ikt} + \text{MIX}_{ikt} + \sum_{i \neq j} \theta_{ijk} C(j, k, t) + \epsilon_{ikt} \\ &= V(i, k, t) + \epsilon_{ikt} \end{aligned} \quad (2)$$

with $C(i, k, t) = 1$ if consumer k purchases category i at time t . The household specific variable HH is specified as

$$\text{HH}_{ikt} = \delta_{1i} \ln [\text{TIME}_{ikt} + 1] + \delta_{2i} \text{LOYAL}_{ik}, \quad (3)$$

where TIME is the time in weeks since the last purchase of consumer i in category k occurs and LOYAL the consumers long-run property to buy in one category. The marketing-mix variable MIX is defined as

$$\text{MIX}_{ikt} = \gamma_i \ln [\text{PRICE}_{ikt}] + \varphi_i \text{DISPLAY}_{ikt} \quad (4)$$

with PRICE the weighted price index across all purchased items in category i and DISPLAY a display index across all items in a category. The cross-category parameter θ_{ijk} implies a positive association between the product categories i and j for values greater zero, and a negative relationship for a values smaller than zero. The cross-category parameter consists of two parts with

$$\theta_{ijk} = \kappa_{ij} + \phi \text{SIZE}_k. \quad (5)$$

as SIZE the mean number of categories chosen by consumer k during the initial period.

Based on the full conditional model from Equation (1) with its utility specification in Equation (2), it follows, using the Theorem of Besag (Besag(1974)), the joint distribution as the final market basket model (Russell and Petersen (2000)) with

$$\Pr(B(k, t) = b) = \frac{\exp(\mu(b, k, t))}{\sum_{b^*} \exp(\mu(b^*, k, t))} \quad (6)$$

and the utility specification as

$$\begin{aligned} \mu(b, k, t) = & \sum_i \beta_i X(i, b) + \sum_i \text{HH}_{ikt} X(i, b) \\ & + \sum_i \text{MIX}_{ikt} X(i, b) + \sum_{i < j} \theta_{ijk} X(i, b) X(j, b) \end{aligned} \quad (7)$$

with $B(k, t) = \{C(1, k, t), \dots, C(N, k, t)\}$ a vector of zeros and ones, $X(i, b) = 1$ if category i is in basket b and zero elsewhere, and b^* all possible baskets excluded the Null basket. Overall, 2^{N-1} baskets are possible for N categories. The interpretation of the model in Equation (6) is that the approach is a logit choice model defined over a set of alternatives with a particular utility specification $\mu(b, k, t)$ as in Equation (7).

3 Data analysis and results

The data set used in our analysis is a one-year period consumer choices and is made available from the "Zentrum für Umfragen und Methoden, Mannheim

(ZUMA)”¹. It contains data on breakfast beverages (e.g., coffee, instant coffee, tea, canned milk and filter paper) and covers 4177 consumers purchasing 40682 baskets during 26 weeks. As explanatory variables in our model, loyalty, time, price and display are used. Examining the five categories, 31 different baskets can be bought by a consumer.

The total value of the **cross-effect** parameter is negative for substitutional and positive for complementary relationships between the categories. Regarding the parameter estimates, we pose the following hypotheses:

- The parameter for **loyalty** should be positive, as a higher loyalty to a category increases the purchase probability in that category.
- The **time** parameter is assumed to be positive because the longer it takes that a consumer did not purchase in a category the higher will be the probability that he will buy a product of that category.
- The **price** coefficient should be negative because higher prices are assumed to lower the possibility of purchasing in a category.
- The **display** parameter should be positive because the existence of display should increase the possibility to buy in a specific category.
- The **size** effect should be positive, larger basket size should lead to higher purchase incidence probability.

We estimate the fit of several stepwise extended models. The first specified model was the simplest without any cross-effects, denoted as M1. Second, we included only the "SIZE"-effect to capture a simple cross-category relationship, the model is called M2. M3 is a model which contains the full cross-category effects. M4 is the most comprehensive one, where in addition to the model in Russell and Petersen (2000), "DISPLAY" is included in the model equation. The fit values for all model types are given in Table 1. It is obvious that ignoring cross-category effects as in model M1, leads to a worse model fit. As more the cross-category relationships are included in the model (M2 to M3), the fit is improved substantially. Also adding an additional explanatory marketing-mix variable (display), results in an even better model fit, not only in the loglikelihood value, but also in the AIC value.

The parameter estimates of model M4 (best fit) are given in Table 2. First, we examine the parameter values for the direct effects. The parameter estimates for loyalty and price are as expected. Each loyalty parameter is positive and significant at the 5%-level. For price, three out of five parameters

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

Model	Number of parameters	LL	AIC
M1	20	-81087.1	162214.2
M2	30	-70507.2	141074.4
M3	31	-69147.2	138356.4
M4	36	-69138.4	138348.7

Table 1. Model fits for breakfast beverages

Variable	Coffee	Instant coffee	Tea	Canned milk	Filter paper
Direct effects					
Intercept	0.93**	2.17**	1.97**	0.55**	-2.81**
Loyalty	0.69**	0.95**	1.08**	0.72**	0.54**
Time	-0.10*	-0.24**	-0.30**	-0.22**	0.66**
Price	-0.95**	-1.35**	-1.02**	-0.07	0.44
Display	3.52**	0.09	-0.54	0.46	0.15
Cross-category effects					
Size	1.13**	1.13**	1.13**	1.13**	1.13**
Coffee	-	-2.24**	-2.29**	-2.20**	-1.41**
Instant coffee	-2.24**	-	-2.45**	-1.94**	-1.60
Tea	-2.29**	-2.45**	-	-2.21**	-1.48**
Canned milk	-2.20**	-1.94**	-2.21**	-	-1.28**
Filter paper	-1.41**	-1.60	-1.48**	-1.28**	-

Table 2. Estimation results for breakfast beverages

are negative and significant, the remaining two (for canned milk and filter paper) are not significant. The estimated parameter values for time are all significant and negative with the exception of filter paper, which is opposite of what we expected. This finding might indicate an irregular shopping behavior in the inspected categories. The display parameter is positive and significant only for coffee, a highly promoted category; the result seems to be quite plausible. In all other categories, display is not used very often; so its influence on the purchase incidence is not important.

Now we will inspect the results for the cross-category parameters. The size effect is positive and significant, as we expected. Its magnitude multiplied with the average basket size (1.39) is 1.75, a value which is larger than several cross-effects (e.g., for filter paper and coffee or for filter paper and canned milk). All cross-category effects are negative and significant. If combined with the average size times size effect, some relationships stay still negative, as e.g., the one for coffee and instant coffee or coffee and tea. This negative value leads to considering these two categories as substitutes, while others

(e.g., filter paper and coffee) have a complementary relationship. Regarding the categories, the results seem to be quite reasonable.

4 Summary and outlook

Managers make decisions for many categories simultaneously. Since ignoring relationships and interdependencies could lead to biased parameter estimates, models should include cross-category effects. We presented a model based on a multivariate logit approach, which is conducted with approaches adopted from spatial statistics. We find significant cross-category parameters, where some imply a relationship of substitutes between the inspected categories of several breakfast beverages. Others imply a complementary relationship. Ignoring these effects results not only in worse model fit, but also in biased parameters of the direct effects.

In an extended analysis, other methods to model and estimate the utility function, e.g. a generalized additive model approach (Hastie and Tibshirani (1990)) might result in a more detailed outcome. Also, considering consumer heterogeneity remains an open issue.

References

- AGRAWAL, R. and SRIKANT, R. (1994): *Fast algorithms for mining association rules*. Working Paper. IBM Almaden Research Center.
- AINSLIE, A. and ROSSI, P.E. (1998): Similarities in Choice Behavior Across Product Categories. *Marketing Science*, 17 (2), 91–106.
- BESAG, J. (1974): Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society, Series B*, 36 (2), 192–236.
- CHIB, S. and SEETHARAMAN, P.B. (2002): Analysis of multi-category purchase incidence decisions using IRI market basket data. In: P.H. Franses and A.L. Montgomery (Eds.): *Econometric Models in Marketing*. Elsevier Science, 57–92.
- CRESSIE, N.A.C. (1991): *Statistics for spatial data*. John Wiley & Sons.
- DAGANZO, C. (1979): *Multinomial Probit*. Academic Press, New York.
- DECKER, R. and MONIEN, K. (2003): Market basket analysis with neural gas networks and self-organising maps. *Journal of Targeting, Measurement and Analysis for Marketing*, 11 (4), 373–386.
- DEEPAK, S.D., ANSARI, A. and GUPTA, S. (2004): *Investigating consumer price sensitivities across categories*. Working Paper. University of Iowa.
- GUADAGNI, P.M. and LITTLE, D.C. (1983): A Logit Model of Brand Choice Calibrated on Scanner Data. *Marketing Science*, 2 (3), 203–238.
- HANSEN, K., SINGH, V.P. and CHINTAGUNTA, P. (2003): *Understanding store brand purchase behavior across categories*. Working Paper. Kellogg School of Management, Northwestern University.
- HASTIE, T.J. and TIBSHIRANI, R.J. (1990): *Generalized Additive Models*. Chapman & Hall, London.

- HAUSMAN, J.A. and WISE, D.A. (1978): A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences. *Econometrica*, 46 (2), 403–426.
- HRUSCHKA, H. (1985): Der Zusammenhang zwischen Verbundbeziehungen und Kaufakt- bzw. Käuferstrukturmerkmalen. *Zeitschrift für betriebswirtschaftliche Forschung*, 37, 218–131.
- HRUSCHKA, H., LUKANOWICZ, M. and BUCHTA, C. (1999): Cross-category sales promotion effects. *Journal of Retailing and Consumer Services*, 6, 99–105.
- LEVINE, J.H. (1979): Joint-Space analysis of pick-any data: Analysis of choices from an unconstrained set of alternatives. *Psychometrika*, 44 (1), 85–92.
- MANCHANDA, P., ANSARI, A. and GUPTA, S. (1999): The "Shopping Basket": A Model for Multicategory Purchase Incidence Decisions. *Marketing Science*, 18 (2), 95–114.
- MILD, A. and REUTTERER, T. (2001): Collaborative Filtering Methods for Binary Market Basket Analysis. In: J. Liu, P.C. Yuen, C.H. Li, J. Ng and T. Ishada (Eds.): *Active Media Technology*. Springer, Berlin, 302–313.
- MILD, A. and REUTTERER, T. (2003): An improved collaborative filtering approach for predicting cross-category purchases based on binary market data. *Journal of Retailing and Consumer Services*, 6 (4), 123–133.
- PAPASTEFANOU, G. (2001): The ZUMA data file version of the GfK Consumer-Scan Household Panel. In: G. Papastefanou, P. Schmidt, A. Börsch-Supan, H. Lüdkte and U. Oltersdorf (Eds.): *Social and Economic Analysis of Consumer Panel Data*. Zentrum für Umfragen, Meinungen und Analysen (ZUMA), Mannheim.
- RUSSELL, G.J., BELL, D., BODAPATI, A., BROWN, C.L., CHIANG, J., GAETH, G., GUPTA, S. and MANCHANDA, P. (1997): Perspectives on Multiple Category Choice. *Marketing Letters*, 8 (3), 297–305.
- RUSSELL, G.J., RATNESHWAR, S., SHOCKER, A.D., BELL, D., BODAPATI, A., DEGERATU, A., HILDEBRANDT, L., KIM, N., RAMASWAMI, S. and SHANKAR, V.H. (1999): Multiple-Category Decision-Making: Review and Synthesis. *Marketing Letters*, 10 (3), 319–332.
- RUSSELL, G.J. and PETERSEN, A. (2000): Analysis of Cross Category Dependence in Market Basket Selection. *Journal of Retailing*, 76 (3), 367–392.
- SCHNEDLITZ, P., REUTTERER, T. and JOOS, W. (2001): Data-Mining und Sortimentsverbundanalyse im Einzelhandel. In: H. Hippner, U. Küsters, M. Meyer and K. Wilde (Eds.): *Handbuch Data Mining im Marketing*. Vieweg, Wiesbaden, 951–970.
- SEETHARAMAN, P.B., AINSLIE, A. and CHINTAGUNTA, P.K. (1999): Investigating Household State Dependence Effect Across Categories. *Journal of Marketing Research*, 36, 488–500.
- SEETHARAMAN, P.B., CHIB, S., AINLSIE, A., BOATWRIGHT, P., CHAN, T., GUPTA, S., MEHTA, N., RAO, V. and STRIJNEV, A. (2004): *Models of multi-category choice behavior*. Working Paper. Rice University.
- SINGH, V.P., HANSEN, K. and GUPTA, S. (2004): *Modeling preferences for common attributes in multi-category brand choice*. Working Paper. Carnegie Mellon University.
- TRAIN, K.E. (2003): *Discrete Choice Methods with Simulation*. Cambridge University Press.

SFB 649 Discussion Paper Series

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

- 001 "Nonparametric Risk Management with Generalized Hyperbolic Distributions" by Ying Chen, Wolfgang Härdle and Seok-Oh Jeong, January 2005.
- 002 "Selecting Comparables for the Valuation of the European Firms" by Ingolf Dittmann and Christian Weiner, February 2005.
- 003 "Competitive Risk Sharing Contracts with One-sided Commitment" by Dirk Krueger and Harald Uhlig, February 2005.
- 004 "Value-at-Risk Calculations with Time Varying Copulae" by Enzo Giacomini and Wolfgang Härdle, February 2005.
- 005 "An Optimal Stopping Problem in a Diffusion-type Model with Delay" by Pavel V. Gapeev and Markus Reiß, February 2005.
- 006 "Conditional and Dynamic Convex Risk Measures" by Kai Detlefsen and Giacomo Scandolo, February 2005.
- 007 "Implied Trinomial Trees" by Pavel Čížek and Karel Komorád, February 2005.
- 008 "Stable Distributions" by Szymon Borak, Wolfgang Härdle and Rafal Weron, February 2005.
- 009 "Predicting Bankruptcy with Support Vector Machines" by Wolfgang Härdle, Rouslan A. Moro and Dorothea Schäfer, February 2005.
- 010 "Working with the XQC" by Wolfgang Härdle and Heiko Lehmann, February 2005.
- 011 "FFT Based Option Pricing" by Szymon Borak, Kai Detlefsen and Wolfgang Härdle, February 2005.
- 012 "Common Functional Implied Volatility Analysis" by Michal Benko and Wolfgang Härdle, February 2005.
- 013 "Nonparametric Productivity Analysis" by Wolfgang Härdle and Seok-Oh Jeong, March 2005.
- 014 "Are Eastern European Countries Catching Up? Time Series Evidence for Czech Republic, Hungary, and Poland" by Ralf Brüggemann and Carsten Trenkler, March 2005.
- 015 "Robust Estimation of Dimension Reduction Space" by Pavel Čížek and Wolfgang Härdle, March 2005.
- 016 "Common Functional Component Modelling" by Alois Kneip and Michal Benko, March 2005.
- 017 "A Two State Model for Noise-induced Resonance in Bistable Systems with Delay" by Markus Fischer and Peter Imkeller, March 2005.

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

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



- 018 "Yxilon – a Modular Open-source Statistical Programming Language" by Sigbert Klinke, Uwe Ziegenhagen and Yuval Guri, March 2005.
- 019 "Arbitrage-free Smoothing of the Implied Volatility Surface" by Matthias R. Fengler, March 2005.
- 020 "A Dynamic Semiparametric Factor Model for Implied Volatility String Dynamics" by Matthias R. Fengler, Wolfgang Härdle and Enno Mammen, March 2005.
- 021 "Dynamics of State Price Densities" by Wolfgang Härdle and Zdeněk Hlávka, March 2005.
- 022 "DSFM fitting of Implied Volatility Surfaces" by Szymon Borak, Matthias R. Fengler and Wolfgang Härdle, March 2005.
- 023 "Towards a Monthly Business Cycle Chronology for the Euro Area" by Emanuel Mönch and Harald Uhlig, April 2005.
- 024 "Modeling the FIBOR/EURIBOR Swap Term Structure: An Empirical Approach" by Oliver Blaskowitz, Helmut Herwartz and Gonzalo de Cadenas Santiago, April 2005.
- 025 "Duality Theory for Optimal Investments under Model Uncertainty" by Alexander Schied and Ching-Tang Wu, April 2005.
- 026 "Projection Pursuit For Exploratory Supervised Classification" by Eun-Kyung Lee, Dianne Cook, Sigbert Klinke and Thomas Lumley, May 2005.
- 027 "Money Demand and Macroeconomic Stability Revisited" by Andreas Schabert and Christian Stoltenberg, May 2005.
- 028 "A Market Basket Analysis Conducted with a Multivariate Logit Model" by Yasemin Boztaş and Lutz Hildebrandt, May 2005.

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

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

