Market Basket Analysis Using Heterogeneous Multivariate Probit Models for Groups of Product Categories

Harald Hruschka

Faculty of Economics
University of Regensburg, Germany
Email: harald.hruschka@wiwi.uni-regensburg.de

***Abstract*—Several heterogeneous multivariate probit models are used to analyze market baskets purchased by households. Each of these models is related to one group of product categories contained in seven prior partitions formed for a total of 25 product categories. The best model in terms of cross-validated log likelihood found considers all categories as one group, i.e., it does not split the 25 categories into two or more groups. In the next step of this project, we will compare this result to multivariate probit models which are related to a partition which is not fixed beforehand, but determined by stochastic model search.**

Keywords–Market basket analysis; Multivariate probit models, MCMC; Stochastic model search

I. INTRODUCTION

Using several heterogeneous multivariate probit models we analyze market baskets, i.e., multicategory purchases of households. Each of these models is related to one group of product categories contained in partitions formed from a total of 25 product categories. In the marketing literature, purchase incidence models as a rule either have a multivariate probit (MVP) or a multivariate logit (MVL) form. Papers applying MVP models typically take latent heterogeneity of households into account. To the best of our knowledge, Manchanda *et al.* [13] provide the first publication analyzing four product categories by MVP models. In their MVP models, Chib *et al.* [6] and Duvvuri *et al.* [9] consider a maximum of twelve and six categories, respectively. Russell and Petersen [15] as well as Boztuğ and Hildebrandt [3] estimate MVL models without latent heterogeneity for a maximum of four and six categories, respectively. Dippold and Hruschka [8] analyze 31 categories by one MVL model and account for latent heterogeneity.

We choose purchase incidences as response variables motivated by he expectation of Song and Chintagunta [16] that interdependences of categories emerge rather on this level than, e.g., for purchase quantities or expenditures. Error correlations are allowed only between categories belonging to the same group. In other words, error correlations are restricted to equal zero between categories which belong to different groups.

To the best of our knowledge, only three studies specify models for different category groups. Chib *et al.* [6] compare the parameter estimates of three MVP models (each model for one group with four categories) and one overall MVP model for all 12 categories. Category groups in [6] are formed by sorting category names in alphabetic order. Boztuğ and Reutterer [4] in a first step determine basket classes by online K-means of purchase incidence data. Then these authors estimate one MVL model for each class. In each MVL model, they consider as category group about five product categories which attain the highest class specific purchase frequencies using data of those households whose purchase incidences have the highest similarity to the relevant basket class.

The paper presented here differs from previous publications in two respects. Firstly, we form seven alternative prior partitions with category groups that reflect the typical uses of assigned categories by household members (e.g., drinking, eating, personal care, cleaning etc.). Then we evaluate the statistical performance of models implied by these seven partitions. Secondly, the total number of categories investigated is much higher compared to studies specifying models for different category groups.

In Section II, we introduce the basic heterogeneous MVP model and subsequently explain the overall model. We give an overview on model estimation in Section III. In section IV, we characterize the data used and present estimation results. In the final Section V, we summarize main results and mention the next step of the project presented here.

II. MODEL SPECIFICATION

The basic heterogeneous MVP model is characterized by the fact that category constants, coefficients, and residual correlations vary across households. J_g symbolizes the number of categories belonging to a category group g. Indices of product categories are denoted as $j = 1, \cdots, J_g$, indices of households as $i = 1, \cdots, I$, indices of baskets of household i as $t = 1, \cdots, T_i$. Household i purchases category j in basket t (symbolized by a purchase indicator $y_{jit} = 1$) if the stochastic utility U_{jit} of such a purchase is positive. If U_{jit} is negative, the household does not purchase category j in basket t (symbolized by a purchase indicator $y_{jit} = 0$). Stochastic utility U_{jit} results from deterministic utility V_{jit} (a linear combination of independent variables plus category constant $\beta_{1,ji}$) to which error ϵ_{jit} is added. We obtain the following expression:

$$U_{jit} = \beta_{1,ji} + \sum_{d=1}^{D} \beta_{1+d,ji}\, x_{1pi} + \sum_{m=1}^{M} \beta_{1+D+m,ji}\, x_{2mjit} + \epsilon_{jit} \quad (1)$$

The model includes two types of independent variables in (1). The first type consists of D predictors x_{1di} which differ across households, but assume the same value for all market baskets and categories of any household i. Socio-demographic household variables are examples of this type of

 ISBN: 978-1-61208-423-7

independent variable. Coefficient $\beta_{1+d,ji}$ indicates the effect of such a household-specific variable d on the utility for category j. The second type of independent variables are M marketing variables x_{2mkit} which differ across market baskets of household i and are specific to category k. Coefficient $\beta_{1+D+m,ji}$ measures the effect of marketing variable m on the deterministic utility of its category j.

We allow errors to be correlated across different categories belonging to the same group. By assuming that errors follow a multivariate normal distribution with zero mean vector and a (J_g, J_g) error covariance matrix the MVP functional form results. To attain identifiability we restrict the error covariance matrix to a correlation matrix [6].

To account for latent heterogeneity of households we use a Dirichlet process mixture (DPM) with MVP models as components. This way we allow for infinitely many household clusters in the overall population, with an unknown number of clusters observed in the finite sample [14]. The DPM is capable to reproduce multimodal and skewed distributions and determines the number of latent clusters alongside the estimation process (see, e.g., [1]). The prior of a DPM is a Dirichlet process, which in this case consists of two independent distributions. The first one is a multivariate normal distribution of category constants and coefficients [6]. The second one is a uniform distribution on the space of correlation matrices of dimension J_g which corresponds to a prior developed by Barnard *et al.* [2] on which Liu and Daniels [12] base an appropriate Metropolis-Hastings simulation step.

The overall model can be seen as union of several heterogeneous MVP models, each of which is specific to one of G groups of a partition of the total set of categories. Error correlations between categories assigned to different category groups are zero.

III. Model Estimation and Evaluation

Models are estimated by iterative Markov chain Monte Carlo (MCMC) simulation comprising the algorithm 7 of Neal [14] to construct household clusters, and additional sampling steps to estimate stochastic utilities, a correlation matrix, category constants and coefficients for each group and cluster. We evaluate performance of each overall model by the expected log likelihood over cross-validated predictive densities, which we briefly call cross-validated log likelihood (CVLL). Cross-validation predictive densities indicate which market baskets are likely if a model is fitted to all data with the exception of the respective observation, i.e., market basket t of household i [10]. To this end parameter samples $\boldsymbol{\theta}_{s,-it}$ with $s = 1, \cdots, 500$ are drawn from the density of parameters $f(\boldsymbol{\theta}_{s,-it})$ using the resampling approach described in Gelfand [10]. $CVLL$ values of a model are defined as:

$$CVLL = \sum_{i=1}^{I}\sum_{t=1}^{T_i}\sum_{j=1}^{J} 1/500 \sum_{s=1}^{500} [y_{jit} \ln p(\boldsymbol{\theta}_{s,-it})(1 - y_{jit}) \ln(1 - p(\boldsymbol{\theta}_{s,-it}))] \quad (2)$$

We compute the probability $p(\boldsymbol{\theta}_{s,-it})$ as relative frequency that the j-th element of 500 random number vectors is greater than zero. These random vectors are generated from a multivariate normal distribution with deterministic utilities as expected values and the error correlation matrix R_i all computed from parameter sample s and for the predictors of category j, household i and basket t.

IV. Empirical Study

A. Data

The data refer to 24,047 shopping visits of a random sample of 1500 households to one specific grocery store over a one year period composed from the IRI data set [5]. Each shopping visit is characterized by a market basket, which is a binary vector whose elements indicate whether a household made a purchase in each of 25 product categories

As predictors we consider two binary marketing variables, feature and display, showing whether any brand of the respective category is advertised by local newspapers and receives special placements in the store, respectively. The original data also include information on price reduction, which we omit because of high correlation with the feature variable. The other predictors are household size (number of persons) and a binary variable high income (set to 1, if income is above the median). Table I contains relative frequencies of purchases, feature and display for each category as well as overall means and standard deviations of the number of baskets, basket size (i.e., the number of categories purchased), and household size.

TABLE I. Descriptive Statistics

Category	Abbreviation	Relative purchase frequency	Relative frequency Feature	Relative frequency Display
Milk	milk	0.476	0.129	0.009
Carbonated beverages	carbbev	0.400	0.175	0.283
Salty snacks	saltsnck	0.351	0.154	0.267
Cold cereal	coldcer	0.280	0.151	0.114
Yogurt	yogurt	0.202	0.179	0.020
Soup	soup	0.197	0.112	0.061
Spaghetti sauce	spagsauc	0.181	0.169	0.072
Toilet tissue	toitisu	0.171	0.095	0.081
Margarine/Butter	margbutr	0.158	0.130	0.026
Paper towels	paptowl	0.140	0.067	0.071
Coffee	coffee	0.136	0.124	0.080
Laundry detergent	laundet	0.118	0.106	0.081
Frozen pizza	fzpizza	0.110	0.174	0.121
Mayonnaise	mayo	0.109	0.100	0.054
Frankfurters and hotdog	hotdog	0.103	0.094	0.034
Mustard/Ketchup	mustketc	0.102	0.041	0.054
Frozen dinner	fzdin	0.090	0.187	0.071
Facial tissue	factiss	0.084	0.119	0.048
Peanut Butter	peanbutr	0.080	0.133	0.053
Beer/Ale	beer	0.076	0.061	0.080
Toothpaste	toothpa	0.059	0.089	0.045
Shampoo	shamp	0.053	0.094	0.077
Deodorant	deod	0.040	0.083	0.034
Household cleaners	hhclean	0.030	0.041	0.016
Diapers	diapers	0.020	0.171	0.010

Variable	Mean	Standard Deviation
Number of baskets	16.05	13.47
Basket size	3.85	2.65
Household size	2.36	1.29

B. Estimation Results

Table II lists all prior partitions investigated. Groups of prior partitions differ with respect to the way that assigned categories are typically used by household members, e.g., for drinking, eating, personal care, cleaning etc. These partitions are also typical for category groupings, which grocery retailers use. $A5$ is the most detailed partition with five lowest level groups. We define higher-level prior groups as unions of lower level ones, i.e., Non Food as union of Personal Care and Cleaning, Other Food as union of Other Food Main and Other Food Additional, Food as union of Beverage and Other Food, and finally $A1$ which comprises all 25 categories as union of Food and Non Food. Note that in the case of $A3$ and $A4$ we

 ISBN: 978-1-61208-423-7

actually consider two alternative category partitions ($A3a$, $A3b$ and $A4a$, $A4b$) with three or four groups.

TABLE II. PRIOR PARTITIONS AND GROUPS

Prior partitions	Category groups
$A1$	One group
$A2$	Food, Non Food
$A3a$	Beverage, Other Food, Non Food
$A3b$	Food, Pers Care, Cleaning
$A4a$	Beverage, Other Food Main, Other Food Additional, Non Food
$A4b$	Beverage, Other Food, Personal Care, Cleaning
$A5$	Beverage, Other Food Main, Other Food Additional, Personal Care, Cleaning
Lowest Level Groups	**Categories**
Beverage	beer, carbev, coffee, milk
Other Food Main	coldcer, fzdin, fzpizza, hotdog, saltsnck, soup, yoghurt
Other Food Additional	margbutr, mayo, musketc, peanbutr, spagsauc
Personal Care	deod, diapers, factiss, shamp, toitisu, toothpa
Cleaning	hhclean, laundet, paptowl

Table III contains the best partition in terms of CVLL for a number of category groups varying between 2 and 5. It also contains the results for the model for which all categories belong to one group. Among prior partitions with at least two groups the most detailed one with five groups ($A5$) performs best. But the overall best performance is attained by $A1$, which treats all categories as belonging to one group. This model is, of course, the most complex one in terms of the number of parameters, as it includes 300 error correlations for all the pairs of the 25 categories.

TABLE III. CROSS-VALIDATED LOG LIKELIHOOD VALUES (CVLL)

# of category groups	prior partitions label	CVLL
1	$A1$	-135,028
2	$A2$	- 163,644
3	$A3a$	-153,520
4	$A4b$	-151,764
5	$A5$	-150,298

Values are rounded to nearest integer.

Parameter estimates are based on every 10th of 50,000 iterations, which are immediately consecutive to a burn-in phase of 50,000 iterations. The largest four household clusters are dominant. Vectors of average percentage shares of these four clusters are (58.5, 16.4, 10.1, 5.9) and (23.7, 21.3, 19.2, 16.3) for models $A1$ and $A5$, respectively.

In the following, we present a selection of higher parameter estimates for the two partitions $A1$ and $A5$. These estimates are averaged across households. Table IV shows all significant effects of the two marketing variables which are greater than 0.15 in absolute size for at least one of the two partitions.

These coefficients indicate positive effects of features and displays on utility. Effects of features are more frequent and as a rule higher compared to effects of display. For the most part, effects for partition $A5$ are higher (e.g., for features: coldcer, margbutr, yogurt, hhclean; for display: hotdog, shamp, coldcer, hhclean), a few become insignificant (features: deod, beer; display: beer).

Table V lists significant average error correlations for the two partitions which are greater than 0.200 in absolute size for at least one of the two models. Note that these correlations are all positive. Our interpretation of error correlations follows Song and Chintagunta [16]. In the case of a positive correlation a demand shock which increases (decreases) the utility of category j, also increases (decreases) utility of category j'.

TABLE IV. SELECTED COEFFICIENTS OF FEATURES AND DISPLAYS

Category	Partition $A1$	Partition $A5$	Category	Partition $A1$	Partition $A5$
			Feature		
coffee	0.352	0.367	laundet	0.287	0.332
hotdog	0.291	0.334	shamp	0.274	0.313
spagsauc	0.250	0.279	fzpizza	0.238	0.280
factiss	0.232	0.260	toothpa	0.223	0.247
deod	0.228	-	beer	0.213	-
peanbutr	0.209	0.213	mustketc	0.197	0.177
soup	0.205	0.224	margbutr	0.206	0.475
milk	0.201	0.206	yogurt	0.200	0.248
mayo	0.197	0.192	saltsnck	0.195	0.216
coldcer	0.196	0.356	hhclean	0.184	0.360
toitisu	0.156	0.178	fzdin	0.130	0.145
diapers	0.222	0.245	paptowl	0.120	0.154
			Display		
mustketc	0.247	0.269	beer	0.230	-
mayo	0.195	0.219	fzpizza	0.191	0.200
hotdog	0.186	0.235	shamp	0.185	0.229
coffee	0.181	0.197	toothpa	0.181	0.216
fzdin	0.171	0.177	peanbutr	0.176	0.197
soup	0.170	0.174	paptowl	0.169	0.170
factiss	0.163	0.161	laundet	0.160	0.179
toitisu	0.158	0.180	coldcer	0.133	0.243
hhclean	0.123	0.233			

all significant coefficients with absolute size > 0.150 in $A1$ or $A5$;
- indicates insignificance.

We obtain the highest correlation for toitisu & paptowl (0.489). Other correlations greater than 0.300 are found for the category pairs toitisu & factiss, musketc & mayo, shamp & deod, laundet & hhclean, paptowl & laundet, toitisu & laundet, and paptowl & factiss. To give an example, a positive demand shock associated with higher utilities of the two categories toitisu and factiss might be triggered by a household's decision to jointly purchase personal care items.

$A5$ restricts about 73% of error correlations to zero because it assigns the two categories involved to different groups. In addition, about 22% of error correlations are lower (including insignificant correlations) according to partition $A5$.

V. CONCLUSION

The models presented here can be used by retail managers to decide which product categories are appropriate for features and displays. In addition, management can on the basis of these models predict sales caused by these marketing decisions. Preliminary results suggest that the most accurate model A1 is preferable if management wants to predict sales. On the other hand, if managers only want to select categories for features and displays and are not interested in sales forecasts, even the models for partition A5 do a satisfactory job.

Dividing 25 product categories between two and five groups leads to worse statistical performance compared to the most complex model which treats all 25 categories as one group. Of course, one cannot rule out the possibility that other partitions than the ones investigated here (which are typical of those used by grocery retailers) could do better. Therefore, the next step of this work consists in determining post hoc

 ISBN: 978-1-61208-423-7

partitions with different numbers of category groups by a stochastic search algorithm drawing upon work of Hoeting *et al.* [11].

TABLE V. SELECTED ERROR CORRELATIONS

Category pair		Partition $A1$	Partition $A5$	Category pair		Partition $A1$	Partition $A5$
toitisu	paptowl	0.489	0	toitisu	factiss	0.336	0.285
mustketc	mayo	0.341	0.329	shamp	deod	0.330	0.176
laundet	hhclean	0.320	0.119	paptowl	laundet	0.311	0.150
toitisu	laundet	0.314	0	paptowl	factiss	0.310	0
saltsnck	carbbev	0.298	0	toitisu	shamp	0.279	0.167
paptowl	hhclean	0.255	-	yogurt	coldcer	0.274	0.092
toothpa	shamp	0.288	0.290	toothpa	deod	0.276	-
fzpizza	fzdin	0.311	0.243	toitisu	deod	0.238	-
shamp	paptowl	0.226	0	spagsauc	coldcer	0.218	0
toothpa	laundet	0.241	0	shamp	laundet	0.233	0
toitisu	hhclean	0.209	0	hhclean	factiss	0.223	0
toitisu	coffee	0.229	0	peanbutr	coldcer	0.219	0
spagsauc	soup	0.215	0	toothpa	toitisu	0.211	0.137
toitisu	margbutr	0.202	0	paptowl	deod	0.223	0
saltsnck	fzpizza	0.205	0.163	margbutr	hhclean	0.186	0
soup	margbutr	0.209	0	toitisu	saltsnck	0.199	0
paptowl	margbutr	0.200	0	mustketc	hotdog	0.191	0
paptowl	coffee	0.200	0	shamp	hhclean	0.227	0
spagsauc	mustketc	0.203	0.207	toothpa	paptowl	0.180	0

all significant correlations with absolute size $>= 0.200$ in $A1$ or $A5$;
- indicates insignificance, 0 that the error correlation is restricted to zero.

Such an approach would simultaneously estimate model parameters, assign categories to groups and households to clusters. Forming category partitions and clustering households would all be directly related to the overall statistical performance of the models. To our knowledge, such an integrated approach has not been attempted in a previous publication.

REFERENCES

[1] A. Ansari and C. F. Mela, "E-Customization," Journal of Marketing Research, vol. 40, 2003, pp. 131–145.

[2] J. Barnard, R. McCulloch, and X. L. Meng, "Modeling Covariance Matrices in Terms of Standard Deviations and Correlations with Application to Shrinkage," Statistica Sinica, vol. 10, 2000, pp. 1281-1311.

[3] Y. Boztuğ and L. Hildebrandt, "Modeling Joint Purchases with a Multivariate MNL Approach," Schmalenbach Business Review, vol. 60, 2008, pp. 400–422.

[4] Y. Boztuğ and T. Reutterer, "A Combined Approach for Segment-Specific Market Basket Analysis," European Journal of Operational Research, vol. 187, 2008, pp. 294–312.

[5] B. J. Bronnenberg, M. W. Kruger, and C. F. Mela, "The IRI Marketing Data Set," Marketing Science, vol. 27, 2008, pp. 745–748.

[6] S. Chib and E. Greenberg, "Bayesian Analysis of Multivariate Probit Models.," Biometrika, vol. 85, 1998, pp. 347–361.

[7] S. Chib, P. B. Seetharaman, and A. Strijnev, Analysis of Multi-Category Purchase Incidence Decisions Using IRI Market Basket Data. JAI, Amsterdam, 2002, pp. 57–92, in Franses, P. H., Montgomery, A. L., Econometric Models in Marketing.

[8] K. Dippold and H. Hruschka, "A Model of Heterogeneous Multi-category Choice for Market Basket Analysis," Review of Marketing Science, vol. 11, 2013, pp. 1–31.

[9] S. D. Duvvuri, V. Ansari, and S. Gupta, "Consumers' Price Sensitivities across Complementary Categories," Management Science, vol. 53, 2007, pp. 1933–1945.

[10] A. E. Gelfand, Model Determination Using Sampling-Based Methods. Chapman & Hall,Boca Raton,1996, pp. 145-161, in Gilks, W. R., Richardson, S., Spiegelhalter, D. J., Markov Chain Monte Carlo in Practice.

[11] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, "Bayesian Model Averaging: A Tutorial," Statistical Science, vol. 14, 1999, pp. 382–417.

[12] X. Liu and M. J. Daniels, "A New Algorithm for Simulating a Correlation Matrix Based on Parameter Expansion and Reparameterization," Journal of Computational and Graphical Statistics, vol. 15, 2006, 897–914.

[13] P. Manchanda, A. Ansari, and S. Gupta, "The Shopping Basket: A Model for Multi-Category Purchase Incidence Decisions," Marketing Science, vol. 18, 1999, pp. 95–114.

[14] R. M. Neal, "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," Journal of Computational and Graphical Statistics, vol. 9, 2000, pp. 249–65.

[15] G. J. Russell and A. Petersen, "Analysis of Cross Category Dependence in Market Basket Selection," Journal of Retailing, vol. 76, 2000, pp. 369–392.

[16] I. Song and P. K. Chintagunta, "A Discrete-Continuous Model for Multicategory Purchase Behavior of Households," Journal of Marketing Research, vol. 44, 2007, pp. 595–612.

 ISBN: 978-1-61208-423-7