Bachelor Thesis [Marketing]

Erasmus University Rotterdam

Juli 4, 2021

Expanded RBM model

Author	*ID*
Jochem Alblas	511281

Supervisor	*Second assessor*
Luuk van Maasakkers	Kathrin Gruber

Abstract

In this paper the cross dependence between categories of groceries is calculated. A data set that consists of online grocery orders at Instacart is used for this purpose. To analyse the effects, a restricted Boltzmann machine as proposed by Hruschka (2014) is used. The restricted Boltzmann machine is compared to a standard multivariate logit model. Unfortunately, it seems that it does not perform better in terms of the absolute deviation. The restricted Boltzmann machine is in this research expanded with a control variable for time of day. The expanded model performs better then the original restricted Boltzmann machine.The categories fresh fruits and fresh vegetables have the most significant cross effects.

Contents

List of Tables

1 Introduction

In the supermarket world there has been a lot of research regarding analysing market baskets. The main problem here is that the choice for a certain category of products in the market basket is often correlated with the choice for others (Russell and Petersen 2000). It is interesting for the supermarkets to examine the cross effects between a large number of different categories. This gives the supermarket more information that could be used for targeting advertisements. The supermarket could target customers with an advertisement for food y, when a supermarket noticed that a customer often bought food x and they know that customers that buy food x are often also interested in food y. This could improve the profits of the supermarket. Furthermore, the supermarket can take a new look at the layout of the store. If food x and y have a high cross effect, then it might be profitable to put those products together. A useful technique to deal with dependence between categories is the multivariate logit model. This model has been used by Manchanda, Ansari, and Gupta (1999) and Chib, Seetharaman, and Strijnev (2002). However, they used a small number of different categories, respectively 4 and 12. For more categories, the multivariate logit model becomes more difficult to estimate with maximum likelihood. This is due to the exponential growth of possible market baskets when the numbers of categories rise. To estimate the model for more categories, a two step approach is required. Hruschka, Lukanowicz, and Buchta (1999) and Boztuğ and Reutterer (2008) have done this to deal with a higher number of categories. However, a two step approach does often not perform well due to that the first step limits the second step (Wedel 1998).

Therefore, another model is needed to capture a higher number of categories. Hruschka (2014) suggest a restricted Boltzmann machine to capture this. In this paper this approach is followed but on a different data set. A data set of *Instacart* (2017) is used which contains data of more than three million orders from Instacart users. The first question thus becomes:

Could a restricted Boltzmann machine help with analysing market baskets that consist of too much categories for the multivariate logit model?

To answer that question different models will be estimated on the data set (*Instacart* 2017). Several different restricted Botlzmann machines (RBM's), which differ from each other in terms of the number of hidden variables used, and the multivariate logit model (MVL) will be estimated and compared to each other. Hruschka (2014) has already investigated this topic and came to the conclusion that the RBM with four hidden variables could indeed beat the MVL. In this paper, the same models are estimated on a different data set. Therefore the robustness of the result of Hruschka (2014) could be verified.

The models are without control variables just as done by Hruschka (2014). Hruschka (2014) justifies using no control variables by looking into previous research done by Boztuğ and Hildebrandt (2008) and Russell and Petersen (2000) where the control variable price did not have a significant effect or did not perform much better than the model without the control variable price. Hruschka (2014) blames also the limitations of the data as a reason that they used no

control variables. However, a different data set is used now where some variables could be used as control variables. Those control variables are added to the best performing RBM. Hence the second question:

Could control variables improve the restricted Boltzmann machine?

Then the models without control variables and with control variables are compared. The identification of the buyer could be used, but this will result in many extra parameters as there are over 200,000 different users. Therefore, time of day is used as control variable. This is supported by the research of Skogster, Uotila, and Ojala (2008). They show that shopping behavior changes depending on the hour of the day. So the control variables could therefore potentially improve the model. However, one must taken into account that this research is performed under different circumstances. For example, they researched shopping behavior in general, while this research focuses on online grocery shopping. Not taking into account control variables could lead to omitted variables. This is of course only the case if the control variable contributes to the model. When there is an omitted variable, the model could give results that do not reflect reality. Lütkepohl (1982) have shown this phenomenon in his paper. This is an extra motivation to try to include control variables in to the RBM. The control variable is added in the form of two binary dummy vectors. The first vector contains information on whether the order has been taken place in the morning and the second vector contains information on whether the order has been taken place in the afternoon. If both are zero then the order has taken place in the evening / night. The details on how the control variables are included are described in more detail in section 3.

It appears in this paper that the RBM without control variable can not beat the MVL. This is a different result from the paper from Hruschka (2014). This could be due to robustness issues of the RBM. Then the RBM performs not necessarily good on each different kind of data set. However, it can also be the result of limitations of this research. Hruschka (2014) estimated the RBM's in 50 random restarts and he has chosen the one with the best Log likelihood. Due to time limitations, the RBM's have only been estimated several times in random restarts. This could be the reason for the difference in result. When the control variable for time of day is added to the RBM with four hidden variables, the RBM does improve. It improves in terms of the log likelihood and the Bayesian information criterion. It also beats the RBM without the control variables and the MVL without control variables in terms of the absolute deviation, which is described in more detail in section 4.

The cross effects that the different categories have on each other are also calculated in the same way as proposed by Hruschka (2014). Next to that the effects that the control variable time of day have on the model is discussed. How the effects are determined is explained in section 2, the effects themselves are discussed in section 4. It seems that the category fresh vegetables has the largest cross effects, this is as expected as it has also a high univariate frequency and a lot of the largest bivariate frequencies as described in section 2. The effects of the control

variable indicated that the cross effects between categories act differently when the order has taken place in the afternoon.

2 Data

The data set used consists of over 3 million shop orders at *Instacart* (2017). Instacart is an online grocery service in America and Canada. Customers can order their groceries in the app or website. The data set contains the products that are purchased at each online order. The data set also contains several other variables like the identification of the user and time of day of the order. The data set from *Instacart* (2017) provides double purchases from a category, so when two products from the same category are bought. The model suggested by Hruschka (2014) only considers a binary variable which equals 1 if one or more products from that category are bought. Therefore, the data set is adjusted slightly to make sure that multiple purchases from the same category are not counted double.

The products of the orders are originally divided in 130 different aisles / categories. Hruschka (2014) however used 50-60 categories for the models, therefore some selection and combining in categories is needed. Here the following guidelines are used.

- An aisle with a univariate frequency higher than 1% becomes a category.
- An aisle with a univariate frequency lower than 0.1% is not considered.
- The remaining aisles are combined according to common sense, in the appendix 6.1 of the final report will be stated which aisles are combined.
- An exception can be made for aisles that seem too interesting to not consider.

This results in the 61 categories stated in table 1. Those categories will be used when estimating the models. The categories fresh vegetables, fresh fruits and packaged vegetables fruits have the highest univariate frequencies. Table 2 provides the highest bivariate marginal frequencies. Again, the categories fresh vegetables, fresh fruits and packaged vegetables fruits are almost at each of the highest bivariate marginal frequencies.

Furthermore, a control variable is used to test whether it can improve the RBM. The control variable that is used is time of day. This variable gives the time of day in hours of when the order has taken place. However, when there are too many different values for the control variable, the normalization constant of the RBM will be too hard to estimate, as will be explained further in section 3. Therefore, instead of 24 different options for each hour of day, the variable is categorised into three different categories. This is displayed in table 3. The three categories are morning, afternoon and evening / night. Evening & night are considered together, because otherwise the categories are too small. For the evaluation of the models, the data will be split in two data sets. The first data set is to estimate the models and the second data set is just for evaluation. This is done randomly and the two data sets will be of the same size. Each data set consists of 1.667.654 orders.

Table 1: Frequency of categories in percentage

Prepared soups salads	0.020	Canned food	0.158
Cheese	0.107	Cookies cakes	0.059
Energy granola bars	0.087	Grains rice dried goods	0.041
Ready made meals	0.169	Energy sports drinks	0.024
Meat	0.172	Fit products	0.022
Bakery desserts	0.038	Fresh dips tapenades	0.098
Pasta	0.120	Soup broth bouillon	0.084
Cold flu allergy	0.007	Refrigerated pudding desserts	0.010
Vega	0.053	Spices	0.105
Fresh herbs	0.094	Soft drinks	0.087
Baking ingredients	0.077	Crackers	0.115
Fat	0.206	Fresh vegetables	0.437
Packaged cheese	0.229	Milk	0.241
Hair care	0.008	Hardware	0.028
Snack	0.160	Eggs	0.137
Fresh fruits	0.539	Salad dressing toppings	0.027
Clean	0.102	Soy lactosefree	0.169
Coffee / thee	0.107	Baby food formula	0.046
Alcohol	0.016	Lunch meat	0.104
Honeys syrups nectars	0.020	Juice nectars	0.090
Foreign food	0.063	Chips pretzels	0.168
Refrigerated	0.133	Pickles goods olives	0.032
Frozen meat seafood	0.021	Other bread	0.164
Ice cream ice	0.110	Water seltzer sparkling water	0.190
Frozen meals	0.074	Frozen products	0.123
Pets	0.019	Yogurt	0.261
Bread	0.096	Cereal	0.093
Candy chocolate	0.070	Packaged vegetables fruits	0.363
Breakfast bars pastries	0.157	Frozen appetizers sides	0.052
(dip) Spreads	0.106	Hot cereal pancake mixes	0.045
Paper goods	0.063		

3 Methodology

Several different models are used: a multivariate logit model (MVL) without control variables and multiple restricted Boltzmann machines (RBM's) where the number of hidden variables varies. Also an independence model is estimated, this is done by setting the coefficients for the effects of the hidden variables equal to zero in a restricted Boltzmann machine. This estimation is done the same way as proposed by Hruschka (2014). Therefore, the MVL is estimated using a pseudo likelihood and the RBM's are estimated using a standard maximum likelihood. Then the best performing RBM is expanded with control variables for time of day as described in section 1. In this research $y_i = (y_{i1}, ..., y_{iJ})'$ is a dummy vector of dimension $J \times 1$ where $y_{ij} = 1$ if category j is in market basket i and zero otherwise and $J = 61$. This the same notation as done by Hruschka (2014). For the MVL model the following probability for market basket i is

Table 2: Bivariate marginal frequencies

Ready made meals, fresh fruits	0.104	Fresh fruits, bread	0.109
Meat, fresh fruits	0.110	Fresh fruits, water seltzer sparkling water	0.106
Meat, fresh vegetables	0.109	Fresh fruits, frozen products	0.087
Meat, packaged vegetables fruits	0.084	Fresh fruits, yogurt	0.180
Fat, fresh fruits	0.126	Fresh fruits, packaged vegetables fruits	0.259
Fat, fresh vegetables	0.112	Canned food, fresh vegetables	0.106
Fat, packaged vegetables fruits	0.089	Canned food, packaged vegetables fruits	0.080
Packaged cheese, fresh fruits	0.150	Fresh vegetables, milk	0.122
Packaged cheese, fresh vegetables	0.133	Fresh vegetables, eggs	0.083
Packaged cheese, Yogurt	0.086	Fresh vegetables, soy lactosefree	0.093
Packaged cheese, packaged vegetables fruits	0.112	Fresh vegetables, chips pretzels	0.079
Snack, fresh fruits	0.102	Fresh vegetables, bread	0.089
Fresh fruits, refrigerated	0.083	Fresh vegetables, water seltzer sparkling water	0.080
Fresh fruits, breakfast bars pastries	0.101	Fresh vegetables, yogurt	0.141
Fresh fruits, canned food	0.104	Fresh vegetables, packaged vegetables fruits	0.228
Fresh fruits, fresh vegetables	0.302	Milk, yogurt	0.093
Fresh fruits, milk	0.157	Milk, packaged vegetables fruits	0.106
Fresh fruits, eggs	0.093	Soy lactosefree, packaged vegetables fruits	0.081
Fresh fruits, soy lactosefree	0.113	Yogurt, packaged vegetables fruits	0.126
Fresh fruits, chips pretzels	0.102		

Table 3: Control variable

	Time, hour of the day	Count	Percentage
Morning	7-12	1,340,921	40.20%
Afternoon	13-18	1,489,223	44.65%
Evening & night	19-6	505,163	15.15%
Total	1-24	3,335,307	100.00%

used:

$$p(y_i) = \frac{exp(a'y_i + y_i'Vy_i)}{Z_{MVL}} \tag{1}$$

, where

$$Z_{MVL} = \sum_{y \in (0,1)^J} exp(a'y + y'Vy) \tag{2}$$

and a is a 61×1 vector with coefficients for the different categories. V is a 61×61 matrix with cross categories coefficients where $V_{jj} = 0$ and where $V_{jl} = V_{lj}$. Z_{MVL} is a normalization constant so that it is a proper distribution.

For the RBM's models, binary hidden variables are added in the form of a $K \times 1$ vector $h_i = (h_{ik}, ..., h_{iK})$, where K equals the number of hidden variables. The joint probability becomes:

$$p(y_i, h_i|c_i) = \frac{exp(b'y_i + h_i'Wy_i + c_i'My_i)}{Z_{RBM}} \tag{3}$$

, where

$$Z_{RBM} = \sum_{y \in (0,1)^J} \sum_{h \in (0,1)^K} \sum_{c} exp(b'y + h'Wy + c'My), \tag{4}$$

b is 61×1 vector with coefficients for the categories and $K \times J$ matrix W contain coefficients that links the hidden variables to the categories. Z_{RBM} is again a normalization constant. The model is expanded with control variables compared to the model of Hruschka (2014). $c_i = (c_{i1}, c_{i2})$ is a vector containing the control variable of time of day. When the order is in the morning $c_i = (1, 0)$, in the afternoon $c_i = (0, 1)$ and when the order is in the evening or at night $c_i = (0, 0)$. M is 2×61 matrix containing coefficients for the control variable. The $\sum_c$ in equation 4 is over the mentioned three options of c_i.

The normalization constants in equations 2 and 4 are hard to compute. This is due to the summation over all possible market baskets. It results in $2^{61} = 2.3 * 10^{18}$ possible baskets. For the RBM's this problem could be solved by rewriting equation 4 into equation 5 as proposed by Hruschka (2014) who got the equation from Larochelle, Bengio, and Turian (2010). Because a control variable is added in this research, equation 5 differs slightly from the expression that Hruschka (2014) used, but it is in line with the expression from Larochelle, Bengio, and Turian (2010). Where $W_{.j}$ equals the j^{th} column of matrix W. The normalization constant then becomes:

$$Z_{RBM} = \sum_{h \in (0,1)^K} \sum_c \prod_{j=1}^{J} (1 + exp(W'_{.j}h + M'_{.j}c + b_j)). \tag{5}$$

Now the total possible baskets for a RBM with four hidden variables are $2^4 * 3 * 61 = 2928$. So when the number of hidden variables and the number control variables is limited, the normalization constant can be reasonably computed. Furthermore to estimate the model the following unnormalized probability is used:

$$p^*(y_i) = exp(b'y_i)(2 + exp(c'_i M y_i)) \prod_{k}^{K} (1 + exp(W_{k,.}y_i)). \tag{6}$$

Then the parameters are estimated using maximum likelihood. The MVL model is estimated using a pseudo maximum likelihood instead of an standard maximum likelihood as described by Hruschka (2014). The maximum likelihood for the RBM's is performed by maximizing the following log likelihood equation:

$$LL = \sum_i log(p(y_i)) = \sum_i log(p^*(y_i)) - I * log(z_{RBM}), \tag{7}$$

where $I = 1.667.654$ is the number of observations. The pseudo maximum likelihood to estimate the MVL is based on the following conditional probability:

$$p(y_{ij}|y_{i,-j}) = \frac{1}{1 + exp(-(a_j + \sum_{l \neq j} V_{jl} y_{il}))}. \tag{8}$$

Then the pseudo log likelihood becomes:

$$LPL = \sum_i \sum_j [y_{ij} log(p(y_{ij}|y_{i,-j})) + (1 - y_{ij})(log(1 - p(y_{ij}|y_{i,-j})))]. \tag{9}$$

The RBM's models are compared to each other with the log likelihood value and the Bayesian information criterion (BIC) (Heij et al. 2004). However those values are not sufficient to compare the MVL model with the RBM's. Therefore, a predictive model selection approach is used, where the absolute deviation is used as an evaluation criterion (Carlin and Louis 2000). For this approach 300 artificial data sets are generated using Gibbs sampling over the conditional distribution, where the parameters have first been estimated using equation 9 for the MVL and equation 7 for the RBM. For the MVL the conditional distribution given in equation 8. For the RBM the conditional contributions are given below:

$$p(h_{ik}|y_i) = \frac{1}{1 + exp(-\sum_j w_{kj} y_{ij})}, \tag{10}$$

$$p(y_{ij}|h_i, c_i) = \frac{1}{1 + exp(-(b_j + \sum_k w_{kj} h_{ik} + m_{1j} c_{i1} + m_{2j} c_{i2}))} \tag{11}$$

and m_{1j} and m_{2j} are both a 61×1 vector with coefficients for respectively the morning and the afternoon. Then the formula for calculating the AD is:

$$AD = \frac{1}{300} \sum_{s=1}^{300} \sum_i \sum_j |y_{ij} - \tilde{y}_{ijs}|, \tag{12}$$

where $\tilde{y}_{ijs}$ is the value for sample s for category j in basket i. For each of the artificial data sets coefficients are estimated using equation 7. Those are used to calculate the significance of the coefficients for the RBM and for the significance of the (cross) effects. The mean of the 300 different estimates are divided by the standard deviation to get the t-value, as proposed by Hruschka (2014).

The marginal cross effects between categories for the best RBM are also considered. Those are computed the same way as described in Hruschka (2014), however the equations must be adapted a bit to account for the control variables. The marginal cross effects are used to analyse the effects that the categories have on each other. To derive the cross effects Hruschka (2014) proposed to use mean field theory (Salakhutdinov and Hinton 2012). The shares here are defined as follows:

$$\langle h_k \rangle = \frac{1}{1 + exp(-\sum_j w_{kj} \langle y_j \rangle)}, \tag{13}$$

$$\langle c_f \rangle = \frac{1}{1 + exp(-\sum_j m_{fj} \langle y_j \rangle)}, \tag{14}$$

$$\langle y_j \rangle = \frac{1}{1 + exp(-(b_j + \sum_k w_{kj} \langle h_k \rangle + m_{1j} \langle c_1 \rangle + m_{2j} \langle c_2 \rangle))}, \tag{15}$$

and $f = 1$ or $f = 2$. However those shares can not be determined analytically, so the shares are being set to the average of hidden variables or categories. Also $\langle c \rangle$ is set to the average of c1 or c2. The cross effects or the effect that category l has on category i are then calculated with the

following equation:

$$\frac{\delta\langle y_j\rangle}{\delta\langle y_l\rangle} = \langle y_j\rangle(1-\langle y_j\rangle)\sum_k w_{kj}w_{kl}\langle h_k\rangle(1-\langle h_k\rangle). \tag{16}$$

These are the cross effects that belong to an RBM model without control variables. This is done so that the results can be better compared with the results from Hruschka (2014). The hidden variables are interpreted by the part they add to the cross effects:

$$\langle y_j\rangle(1-\langle y_j\rangle)w_{kj}w_{kl}\langle h_k\rangle(1-\langle h_k\rangle). \tag{17}$$

This is done for all $k = 1, .., K$ where K is the number of hidden variables. To interpret the effect of the control variable in the RBM model with control variable, the derivative of equation 15 with respect to $\langle c_1\rangle$ and $\langle c_2\rangle$ is calculated as follows:

$$\frac{\delta\langle y_j\rangle}{\delta\langle c_1\rangle} = \langle y_j\rangle(1-\langle y_j\rangle)m_{1j}, \tag{18}$$

$$\frac{\delta\langle y_j\rangle}{\delta\langle c_2\rangle} = \langle y_j\rangle(1-\langle y_j\rangle)m_{2j}. \tag{19}$$

Furthermore the cross effects for the expanded model are calculated as follows:

$$\frac{\delta\langle y_j\rangle}{\delta\langle y_l\rangle} = \langle y_j\rangle(1-\langle y_j\rangle)\sum_k w_{kj}w_{kl}\langle h_k\rangle(1-\langle h_k\rangle)\sum_{f=1}^{2} m_{fj}m_{fl}\langle c_f\rangle(1-\langle c_f\rangle). \tag{20}$$

The significance of all (cross) effects is calculated by dividing the mean with the standard deviation across the 300 artificial data sets. More details on how the RBM from Hruschka (2014) is expanded can be found in the appendix 6.2

4 Results

Several RBM's are estimated. The comparison between the RBM's is stated in table 4. The log likelihood (LL) value for the estimation data and the validation data does not differ much. This implies that the RBM's are robust. Furthermore, the LL rises and the BIC declines, as the number of hidden variables becomes larger. This indicated that the model performs better with a higher number of hidden variables. However, due to the limitation of time, a model with four hidden variables is used, as it is simpler to estimate and interpret then a model with a higher number of hidden variables. Hruschka (2014) also used the RBM with four hidden variables, therefore it is also easier to compare the results. All the models have a better LL and BIC than the independence model. This indicated that the cross relations between the categories of grocery items should indeed not be ignored. When the RBM is expanded with the control variable for time of day, it improves the LL and BIC. This could indicate that including this control variable does have a positive effect on the performance of a RBM. To compare MVL with the RBM's, the absolute deviation (AD) is calculated as proposed by Hruschka (2014) and described in section 3. The AD from the RBM with four hidden variables and without a control variable equals 101.954. The AD from the MVL model is smaller with a value of 100.537. This

is a different result than the result from Hruschka (2014) where the RBM with four variables did perform better than the MVL. However Hruschka (2014) estimated the RBM's fifty times at different random starting values. Due to time limitations, the model is only estimated five times. This could have resulted in the different outcome. Furthermore the data used is different, so there is also the possibility that this data is less suited for the RBM's compared to the data set used by Hruschka (2014). The AD from the RBM model with four hidden variables, but expanded with the control variable for time of day equals 92.362. Hence the model with a control variable outperforms the RBM without a control variable and it outperforms the MVL.

Table 4: Restricted Boltzmann machines

Number of hidden variables	Number of parameters	Log Likelihood (LL)		Bayesian Information Criterion (BIC)
		Estimation data	Validation data	
Restricted Boltzmann Machines (RBM"s)				
1	122	-31,035,282	-31,033,463	62,072,311
2	183	-30,888,899	-30,886,054	61,780,419
3	244	-30,721,822	-30,719,454	61,447,138
4	305	-30,564,663	-30,563,250	61,133,696
4, with control variable	427	-30,278,212	-30,276,765	60,562,540
Independence model	61	-31,326,753	-31,325,829	62,654,380

4.1 RBM without control variable

With the model with four hidden variables without the control variable are 2235 significant cross effects at a 1% significance level, so with t-values higher then 2.34. The highest absolute cross effects are given in table 5. A negative cross effect indicates that when for example alcohol is ordered, there will be a smaller probability that fresh vegetables are also ordered. This result seems to make sense, as alcohol and fresh vegetables are not compatible categories. The most cross effects contain the category fresh vegetables, which is expected when table 1 and 2 are considered. The category fresh vegetables has a high univariate frequency and also a lot of the largest bivariate frequencies. On top of that Hruschka (2014) also found that the category vegetables is the category with the most significant cross effects. The presence of many cross effects with canned food is surprising as the category does not have many of the largest bivariate frequencies as can be seen in table 2.
Table 6 provides significant w_{kj} values that are larger than two in absolute terms. A value larger than 2 implies that the dependent odds ratio of h_{ik} is multiplied by over 7 if category j is ordered. A value smaller than -2 implies the odds ratio being multiplied by less then 1/7. For example if a product in the category cookies cake is ordered, then it becomes more than 7 times likelier that h1 equals 1 compared to h1 being equal to 0.

The partial marginal cross effects are given in table 7. Those are the individual contribu-

Table 5: Cross effects

Meat, pasta	0.165	Fresh vegetables, fresh herbs	0.546
Meat, fresh herbs	0.173	Fresh vegetables, alcohol	-0.183
Pasta, canned food	0.161	Fresh vegetables, foreign food	0.211
Fresh herbs, fresh vegetables	0.181	Fresh vegetables, bread	0.223
Packaged cheese, pasta	0.163	Fresh vegetables, Candy chocolate	-0.166
Fresh fruits, fresh vegetables	0.164	Fresh vegetables, canned food	0.334
Fresh fruits, packaged vegetables fruits	0.171	Fresh vegetables, cookies cakes	-0.198
Breakfast bars pastries, cookies cakes	0.168	Fresh vegetables, grains rice dried goods	0.344
Canned food, pasta	0.206	Fresh vegetables, energy sports drinks	-0.198
Canned food, fresh herbs	0.211	Fresh vegetables, soup broth bouillon	0.245
Canned food, foreign food	0.164	Fresh vegetables, spices	0.221
Canned food, bread	0.170	Fresh vegetables, soft drinks	-0.220
Canned food, grains rice dried goods	0.177	Fresh vegetables, eggs	0.227
Canned food, soup broth bouillon	0.175	Fresh vegetables, salad dressing toppings	0.166
Canned food, spices	0.177	Fresh vegetables, pickles goods olives	0.299
Canned food, fresh vegetables	0.172	Fresh vegetables, frozen products	0.222
Canned food, pickles goods olives	0.177	Packaged vegetables fruits	0.254
Fresh vegetables, meat	0.277	Packaged vegetables fruits, fresh herbs	0.254
Fresh vegetables, pasta	0.331	Packaged vegetables fruits, fresh vegetables	0.237
Fresh vegetables, vega	0.234		

Table 6: Selected w_{kj} coefficients

Category	Hidden variable			
	h1	h2	h3	h4
Meat		-2.41		
Pasta		-2.73	-2.09	
Fresh herbs		-2.98		
Fresh fruits				-2.27
Alcohol				2.43
Foreign food			-2.48	
Bread		-2.07	-2.09	
Breakfast bars pastries	2.12			
Canned food		-2.43	-2.14	
Cookies cakes	2.41			
Grains rice dried goods		-2.17		
Soup broth bouillon			-2.23	
Spices			-2.55	
Fresh vegetables		-2.92		-2.21
Pickles goods olives		-2.50		
Packaged vegetables fruits				-2.01
Frozen appetizers sides			-2.46	

tion to the cross effect of one hidden variable. The table shows the significant effects that are larger than 0.11. Hidden variable 2 and 4 have the most contributions to the cross effects. Hidden variable 2 contributes mostly positive to cross effects which contain the category fresh vegetables, whereas hidden variable 4 contributes more negative to other cross effects which contain the category fresh vegetables. Hidden variable 3 contributes mostly negative to cross effects that involve fresh fruits and hidden variable 1 contributes mainly to cross effects that involve cookies cake.

Table 7: Partial marginal cross effects per hidden variable

Hidden variable 1		Hidden variable 3	
Breakfast bars pastries, cookies cake	0.115	Fresh fruits, pasta	-0.125
Fresh vegetables, cookies cake	-0.111	Fresh fruits, foreign food	-0.149
		Fresh fruits, bread	-0.126
		Fresh fruits, canned food	-0.128
		Fresh fruits, soup broth bouillon	-0.134
		Fresh fruits, spices	-0.153
		Fresh fruits, frozen appetizers sides	-0.148
Hidden variable 2		Hidden variable 4	
Meat, pasta	0.123	Fresh fruits, cold flu allergy	-0.148
Meat, fresh herbs	0.135	Fresh fruits, fresh herbs	0.191
Meat, fresh vegetables	0.132	Fresh fruits, hair care	-0.133
Meat, pickles goods olives	0.113	Fresh fruits, clean	-0.139
Pasta, fresh herbs	0.111	Fresh fruits, alcohol	-0.185
Canned food, pasta	0.115	Fresh fruits, pets	-0.142
Canned food, fresh herbs	0.126	Fresh fruits, paper goods	-0.149
Canned food, fresh vegetables	0.123	Fresh fruits, energy sports drinks	-0.117
fresh vegetables, energy granola bars	-0.114	Fresh fruits, fresh dips tapenades	0.113
Fresh vegetables, meat	0.237	Fresh fruits, soft drinks	-0.125
Fresh vegetables, pasta	0.269	Fresh fruits, fresh vegetables	0.168
Fresh vegetables, vega	0.139	Fresh fruits, hardware	-0.132
Fresh vegetables, fresh herbs	0.293	Fresh fruits, packaged vegetables fruits	0.153
Fresh vegetables, packaged cheese	0.112	Fresh vegetables, cold flu allergy	-0.140
Fresh vegetables, hair care	0.132	Fresh vegetables, fresh herbs	0.180
Fresh vegetables, clean	0.124	Fresh vegetables, hair care	-0.126
Fresh vegetables, foreign food	0.131	Fresh vegetables, fresh fruits	0.164
Fresh vegetables, frozen meat seafood	0.115	Fresh vegetables, clean	-0.131
Fresh vegetables, pets	0.111	Fresh vegetables, alcohol	-0.175
Fresh vegetables, bread	0.203	Fresh vegetables, pets	-0.134
Fresh vegetables, candy chocolate	-0.118	Fresh vegetables, paper goods	-0.140
Fresh vegetables, paper goods	0.133	Fresh vegetables, energy sports drinks	-0.110
Fresh vegetables, canned food	0.240	Fresh vegetables, soft drinks	-0.118
Fresh vegetables, grains rice dried goods	0.214	Fresh vegetables, hardware	-0.124
Fresh vegetables, soup broth bouillon	0.188	Fresh vegetables, packaged vegetables fruits	0.145
Fresh vegetables, spices	0.167	Packaged vegetables fruits, cold flu allergy	-0.119
Fresh vegetables, eggs	0.186	Packaged vegetables fruits, fresh herbs	0.153
Fresh vegetables, salad dressing toppings	0.153	Packaged vegetables fruits, fresh fruits	0.139
Fresh vegetables, lunch meat	0.128	Packaged vegetables fruits, clean	-0.111
Fresh vegetables, pickles goods olives	0.246	Packaged vegetables fruits, alcohol	-0.149
Fresh vegetables, other bread	0.119	Packaged vegetables fruits, pets	-0.114
Fresh vegetables, frozen produce	0.124	Packaged vegetables fruits, paper goods	0.119
Fresh vegetables, frozen appetizers sides	0.142	Packaged vegetables fruits, fresh vegetables	-0.135

4.2 Expanded model

The model performs better, when the control variable for time of day is added as described in section 3. Therefore some significant coefficients are expected that improved the model. Table 10 in the appendix 6.3 gives the significant coefficients of the control variable. However, the coefficient which would correspond with the morning as time of day are all insignificant, which

indicates that there is no difference between orders in the morning and in the evening/night. All the coefficients corresponding with an order taken place in the afternoon are significantly smaller then zero. The marginal effects are used to interpret what this suggest. Those effects are given in table 8. All the effects are negative and significant. Allowing the model for this change in effect does improve the model as the AD is lower and therefore better than the model without this control variable. The highest effects are seen with fresh vegetables and packaged vegetables fruits. This indicates that the odds ratio between for example fresh vegetables and prepared soups salads becomes smaller. In other words, in the afternoon is the probability of purchase of a product of the category prepared soups salads relative higher compared to the probability of purchase of fresh vegetables given the other purchased categories.

The highest significant cross effects for the expanded model are given in table 9. The most surprising change in comparison with the cross effects from the RBM without a control variable in table 5, is that the category canned food has far less of the largest cross effects. Instead the category fresh fruits is more present. This seems to be a better representative of the data as fresh fruits has the highest univariate frequency and a lot of the highest bivariate frequencies as can be seen in table 1 and 2 respectively.

5 Conclusion

To deal with the cross effects between different categories, the multivariate logit model is often used. However, for a large number of categories the model performs not sufficient. Therefore another model is needed. Hruschka (2014) has proposed a restricted Boltzmann machine to deal with a larger number of categories. In this paper, that restricted Boltzmann machine has been repeated on a different data set. However, in this paper it did not perform better then the multivariate logit model in terms of the absolute deviation. In an attempt to improve the restricted Boltzmann machine, a control variable for the time of day was added. The model improved in terms of log likelihood, Bayesian information criteria and the absolute deviation. It even outperformed the multivariate logit model. So adding a control variable can improve the restricted Boltzmann machines. In further research, it will be interesting to investigate if other control variables or more control variables could improve the model even further.

The (cross) effects have also been investigated. It appears that the category fresh vegetables has the most cross effects with other categories. Hruschka (2014) found a similar result in his research where the category vegetables appeared the most in the significant cross effects. It appears that the dummy variables for orders taken place in the afternoon are all significant. This indicated that the cross effects behave different in the afternoon from in the morning or at evening / night. It would be interesting for further research to examine those effects even more and to find out what drives this effect.

Table 8: Effects control variable afternoon

Prepared soups salads	-0.004	Canned food	-0.145
Cheese	-0.075	Cookies cakes	-0.028
Energy granola bars	-0.055	Grains rice dried goods	-0.014
Ready made meals	-0.178	Energy sports drinks	-0.006
Meat	-0.171	Fit products	-0.005
Bakery desserts	-0.012	Fresh dips tapenades	-0.065
Pasta	-0.090	Soup broth bouillon	-0.049
Cold flu allergy	-0.001	Refrigerated pudding desserts	-0.001
Vega	-0.022	Spices	-0.074
Fresh herbs	-0.061	Soft drinks	-0.063
Baking ingredients	-0.044	Crackers	-0.087
Fat	-0.243	Fresh vegetables	-0.829
Packaged cheese	-0.276	Milk	-0.313
Hair care	-0.001	Hardware	-0.008
Snack	-0.160	Eggs	-0.116
Fresh fruits	-1.211	Salad dressing toppings	-0.006
Clean	-0.077	Soy lactosefree	-0.171
Coffee / thee	-0.082	Baby food formula	-0.017
Alcohol	-0.002	Lunch meat	-0.071
Honeys syrups nectars	-0.004	Juice nectars	-0.058
Foreign food	-0.030	Chips pretzels	-0.167
Refrigerated	-0.115	Pickles goods olives	-0.009
Frozen meat seafood	-0.004	Other Bread	-0.154
Ice cream ice	-0.082	Water seltzer sparkling water	-0.232
Frozen meals	-0.041	Frozen produce	-0.097
Pets	-0.004	Yogurt	-0.348
Bread	-0.062	Cereal	-0.061
Candy chocolate	-0.039	Packaged vegetables fruits	-0.614
Breakfast bars pastries	-0.145	Frozen appetizers sides	-0.021
(dip) Spreads	-0.075	Hot cereal pancake mixes	-0.017
Paper goods	-0.034		

5.1 Discussions

The performance of the restricted Boltzmann machine without control variables does not outperform the the multivariate logit model. However, in the paper of Hruschka (2014) it does. This could be due to the different data set used, which would mean that the restricted Boltzmann machine is not robust for other data. The data used is data from online orders. This may have caused the difference in results form the paper of Hruschka (2014). Another issue may be that of the estimation of the model. Hruschka (2014) used 50 random restarts to estimate the best possible restricted Boltzmann machine. However, there were only a few random restarts in this paper. This could explain the difference in performance and it is an interesting point to further investigate.

This also rises the question whether the control variable did improve the restricted Boltzmann machine or it was only a better estimation of the restricted Boltzmann machine. In that case

Table 9: Cross effects, expanded model

Meat, fresh vegetables	0.225	Fresh vegetables, fresh fruits	0.778
Pasta, fresh vegetables	0.142	Fresh vegetables, bread	0.195
Fresh herbs, fresh vegetables	0.157	Fresh vegetables, canned food	0.477
Packaged cheese, fresh vegetables	0.190	Fresh vegetables, soup broth bouillon	0.185
Fresh fruits, fresh herbs	0.169	Fresh vegetables, spices	0.210
Fresh fruits, Clean	-0.178	Fresh vegetables, soft drinks	-0.174
Fresh fruits, breakfast bars pastries	-0.133	Fresh vegetables, eggs	0.281
Fresh fruits, Soft drinks	-0.153	Fresh vegetables, bread	0.171
Fresh fruits, fresh vegetables	0.916	Fresh vegetables, water seltzer sparkling water	-0.253
Fresh fruits, milk	0.169	Fresh vegetables, frozen products	0.248
Fresh fruits, chips pretzels	-0.184	Fresh vegetables, yogurt	0.137
Fresh fruits, frozen products	0.136	Fresh vegetables, packaged vegetables fruits	0.831
Fresh fruits, yogurt	0.279	Water seltzer sparkling water, fresh vegetables	-0.161
Fresh fruits, packaged vegetables fruits	0.790	Yogurt, fresh fruits	0.167
Canned food, fresh vegetables	0.230	Packaged vegetables fruits, fresh herbs	0.170
Fresh vegetables, cheese	0.137	Packaged vegetables fruits, fresh fruits	0.597
Fresh vegetables, meat	0.434	Packaged vegetables fruits, canned food	0.135
Fresh vegetables, pasta	0.362	Packaged vegetables fruits, fresh vegetables	0.740
Fresh vegetables, fresh herbs	0.462	Packaged vegetables fruits, Yoghurt	0.134
Fresh vegetables, packaged cheese	0.303		

it only seems that the control variable improved the model, but in reality it could be that it is only a better local optimum. To examine this in further research may give some clearer answers. It is important then that there are a lot more random restarts performed than in this research. There were namely a lot of difference in performing amongst different restarts in this research. Therefore, more restarts could improve the model. The effect of adding more control variables is also interesting to investigate. However, the number of control variables should not be too large, otherwise the normalization constant will be too hard to compute. To avoid this, one may consider adding one set of control variables and compare the results with adding another set of control variables. In this manner the best performing model could be chosen. Possible suggestion for a control variable are a variable for display advertising or price advertising.

6 Appendix

6.1 Variable selection

First the aisles with a higher univariate frequency then 1% are chosen to be a category. Second the aisles with a lower univariate frequency then 0.1% are not considered. At the start there were 134 aisles. There are 26 aisles that are immediately selected to be a category and 33 aisles that are not considered. Then 75 aisles remain to be combined. Those aisles are combined according to common sense as follows:

6.2 RBM model expanded

The following two equations gives the distribution for the RBM model described by Hruschka (2014):

$$p(y_i, h_i) = \frac{exp(b' y_i + h_i' W y_i)}{Z_{RBM}} \tag{21}$$

Aisle	Category	Aisle	Category
Prepared soups salads	Prepared soups salads	Red wines	Alcohol
Specialty cheeses	Cheese	Honeys syrups nectars	Honeys syrups nectars
Other creams cheeses	Cheese	Latino foods	Foreign food
Instant foods	Ready made meals	Asian foods	Foreign Food
Prepared meals	Ready made meals	Frozen meat seafood	Fish
Packaged produce	Ready made meals	Dog food care	Pets
Frozen pizza	Ready made meals	Cat food care	Pets
Marinades meat preparation	Meat	Buns rolls	Bread
Poultry counter	Meat	Frozen breads doughs	Bread
Packaged poultry	Meat	Tortillas flat bread	Bread
Meat counter	Meat	Candy chocolate	Candy chocolate
Hot dogs bacon sausage	Meat	Breakfast bars pastries	Breakfast
Bakery desserts	Baking	Frozen breakfast	Breakfast
Doughs gelatins bake mixes	Baking	Granola	Breakfast
Pasta sauce	Pasta	Breakfast bakery	Breakfast
Fresh pasta	Pasta	Preserved dips spreads	(dip) Spreads
Dry pasta	Pasta	Spreads	(dip) Spreads
Cold flu allergy	Medicine	Paper goods	Paper goods
Tofu meat alternatives	Vega	Canned meals beans	Canned food
Frozen vegan vegetarian	Vega	Canned jarred vegetables	Canned food
Oils vinegars	Fat	Canned meat seafood	Canned food
Cream	Fat	Canned fruit applesauce	Canned food
Butter	Fat	Cookies cakes	Cookies cakes
Popcorn jerky	Snack	Grains rice dried goods	Grains rice dried goods
Fruit vegetable snacks	Snack	Energy sports drinks	Energy sports drinks
Trail mix snack mix	Snack	Protein meal replacements	Fit products
Nuts seeds dried fruit	Snack	Vitamins supplements	Fit products
Soap	Clean	Refrigerated pudding desserts	Refrigerated pudding desserts
Dish detergents	Clean	Condiments	Spices
Laundry	Clean	Spices seasonings	Spices
Cleaning products	Clean	Plates bowls cups flatware	Hardware
Body lotions soap	Clean	Food storage	Hardware
Oral hygiene	Clean	Salad dressing toppings	Salad dressing toppings
Coffee	Coffee / tea	Pickled goods olives	Pickled goods olives
Tea	Coffee / tea	Frozen appetizers sides	Frozen appetizers sides
Beers coolers	Alcohol	Hot cereal pancake mixes	Hot cereal pancake mixes

,where

$$Z_{RBM} = \sum_{y \in (0,1)^J} \sum_{h \in (0,1)^K} exp(b'y + h'Wy). \tag{22}$$

This model is in this paper extended with the control variable time of day as described in section 3. The control variable must interact with the y vector to make sure that the effects are connected to a purchase of a specific category. The distribution then becomes

$$p(y_i, h_i | c_i) = \frac{exp(b'y_i + h_i'Wy_i + c_i'My_i)}{Z_{RBM}} \tag{23}$$

,where

$$Z_{RBM} = \sum_{y \in (0,1)^J} \sum_{h \in (0,1)^K} \sum_{c} exp(b'y + h'Wy + c'My) \tag{24}$$

and c is the vector which contains the control variable time of day. To estimate the model the following normalization constants are used:

$$Z_{RBM} = \sum_{h \in (0,1)^K} \prod_{j=1}^{J} (1 + exp(W_{.j}'h)) \tag{25}$$

$$Z_{RBM} = \sum_{h \in (0,1)^K} \sum_{c} \prod_{j=1}^{J} (1 + exp(W'_{\cdot j} h + M'_{\cdot j} c + b_j)), \tag{26}$$

where equation 25 is the equation from Hruschka (2014). To make sure the expanded model is also a proper distribution, the normalization constant in equation 26 is used. The sum over the three possible values of the vector c must be used to make the integral from minus infinity to infinity equal to 1. In the maximum likelihood the following unnormalized probability are used:

$$p^*(y_i) = exp(b' y_i) \prod_{k}^{K} (1 + exp(W_{k,\cdot} y_i)) \tag{27}$$

and

$$p^*(y_i) = exp(b' y_i)(2 + exp(c'_i M y_i)) \prod_{k}^{K} (1 + exp(W_{k,\cdot} y_i)). \tag{28}$$

Equation 27 is the unnormalized probility of the RBM form Hruschka (2014) and equation is the RBM expanded with the control variable. When the coefficients in matrix M are set to zero, the expanded model must be the same as the original RBM model. Therefore the term $2 + exp(c'_i M y_i)$ is used. When M equals 0, then $2 + exp(c'_i M y_i) = 3$. The summation over c in equation 26, when M equals 0 also results in a multiplication of three compared to the original model. When divided in the log likelihood those terms thus cancels out. The cross effects are calculated as follows:

$$\frac{\delta \langle y_j \rangle}{\delta \langle y_l \rangle} = \langle y_j \rangle (1 - \langle y_j \rangle) \sum_{k} w_{kj} w_{kl} \langle h_k \rangle (1 - \langle h_k \rangle), \tag{29}$$

$$\frac{\delta \langle y_j \rangle}{\delta \langle y_l \rangle} = \langle y_j \rangle (1 - \langle y_j \rangle) \sum_{k} w_{kj} \frac{\delta \langle h_k \rangle}{\delta \langle y_l \rangle} \sum_{f=1}^{2} m_{fj} \frac{\delta \langle c_f \rangle}{\delta \langle y_l \rangle} \tag{30a}$$

and

$$\frac{\delta \langle y_j \rangle}{\delta \langle y_l \rangle} = \langle y_j \rangle (1 - \langle y_j \rangle) \sum_{k} w_{kj} w_{kl} \langle h_k \rangle (1 - \langle h_k \rangle) \sum_{f=1}^{2} m_{fj} m_{fl} \langle c_f \rangle (1 - \langle c_f \rangle). \tag{30b}$$

where equation 29 are the cross effects as stated by Hruschka (2014). Equation 30 gives the derivation of the cross effect of the expanded RBM.

6.3 Coefficients for control variable afternoon

Table 10: Coefficients for control variable afternoon

Prepared soups salads	-0.212	Canned food	-1.685
Cheese	-1.145	Cookies cakes	-0.632
Energy granola bars	-0.924	Grains rice dried goods	-0.437
Ready made meals	-1.816	Energy sports drinks	-0.258
Meat	-1.842	Fit products	-0.238
Bakery desserts	-0.406	Fresh dips tapenades	-1.048
Pasta	-1.286	Soup broth bouillon	-0.889
Cold flu allergy	-0.071	Refrigerated pudding desserts	-0.109
Vega	-0.565	Spices	-1.120
Fresh herbs	-0.997	Soft drinks	-0.931
Baking ingredients	-0.817	Crackers	-1.231
Fat	-2.203	Fresh vegetables	-4.632
Packaged cheese	-2.450	Milk	-2.570
Hair care	-0.084	Hardware	-0.300
Snack	-1.711	Eggs	-1.462
Fresh fruits	-5.743	Salad dressing toppings	-0.282
Clean	-1.089	Soy lactosefree	-1.800
Coffee / thee	-1.135	Baby food formula	-0.494
Alcohol	-0.160	Lunch meat	-1.115
Honeys syrups nectars	-0.219	Juice nectars	-0.959
Foreign food	-0.674	Chips pretzels	-1.790
Refrigerated	-1.419	Pickles goods olives	-0.341
Frozen meat seafood	-0.224	Other Bread	-1.743
Ice cream ice	-1.172	Water seltzer sparkling water	-2.031
Frozen meals	-0.796	Frozen produce	-1.313
Pets	-0.204	Yogurt	-2.757
Bread	-1.029	Cereal	-0.996
Candy chocolate	-0.741	Packaged vegetables fruits	-3.853
Breakfast bars pastries	-1.668	Frozen appetizers sides	-0.556
(dip) Spreads	-1.135	Hot cereal pancake mixes	-0.486
Paper goods	-0.677		

References

[1] Yasemin Boztuğ and Lutz Hildebrandt. “Modeling joint purchases with a multivariate MNL approach”. In: *Schmalenbach Business Review* 60.4 (2008), pp. 400–422.

[2] Yasemin Boztuğ and Thomas Reutterer. “A combined approach for segment-specific market basket analysis”. In: *European Journal of Operational Research* 187.1 (2008), pp. 294–312.

[3] Bradley P Carlin and Thomas A Louis. *Bayes and empirical Bayes methods for data analysis*. Chapman & Hall/CRC, 2000.

[4] Siddhartha Chib, PB Seetharaman, and Andrei Strijnev. "Analysis of multi-category purchase incidence decisions using IRI market basket data". In: *Advances in Econometrics*. Emerald Group Publishing Limited, 2002.

[5] Christiaan Heij, Paul de Boer, Philip Hans Franses, Teun Kloek, and Herman van Dijk. *Econometric Methods with applications in Business and Economics*. Oxford University Press, 2004, pp. 92–93, 99, 279, 281, 315–316, 387.

[6] Harald Hruschka. "Analyzing market baskets by restricted Boltzmann machines". In: *OR spectrum* 36.1 (2014), pp. 209–228.

[7] Harald Hruschka, Martin Lukanowicz, and Christian Buchta. "Cross-category sales promotion effects". In: *Journal of Retailing and Consumer Services* 6.2 (1999), pp. 99–105.

[8] *Instacart*. `https://www.kaggle.com/c/instacart-market-basket-analysis`. Accessed: 30-04-2021. 2017.

[9] Hugo Larochelle, Yoshua Bengio, and Joseph Turian. "Tractable multivariate binary density estimation and the restricted Boltzmann forest". In: *Neural computation* 22.9 (2010), pp. 2285–2307.

[10] Helmut Lütkepohl. "Non-causality due to omitted variables". In: *Journal of Econometrics* 19.2 (1982), pp. 367–378. ISSN: 0304-4076. DOI: `https://doi.org/10.1016/0304-4076(82)90011-2`. URL: `https://www.sciencedirect.com/science/article/pii/0304407682900112`.

[11] Puneet Manchanda, Asim Ansari, and Sunil Gupta. "The "shopping basket": A model for multicategory purchase incidence decisions". In: *Marketing science* 18.2 (1999), pp. 95–114.

[12] Gary J Russell and Ann Petersen. "Analysis of cross category dependence in market basket selection". In: *Journal of Retailing* 76.3 (2000), pp. 367–392.

[13] Ruslan Salakhutdinov and Geoffrey Hinton. "An efficient learning procedure for deep Boltzmann machines". In: *Neural computation* 24.8 (2012), pp. 1967–2006.

[14] Patrik Skogster, Varpu Uotila, and Lauri Ojala. "From mornings to evenings: is there variation in shopping behaviour between different hours of the day?" In: *International Journal of Consumer Studies* 32.1 (2008), pp. 65–74.

[15] M Wedel. "Kamakura. WA (1998) Market Segmentation: Conceptual and Methodological Foundations". In: *Kluwer, Boston, Dordrecht, London. Received* 22 (1998), p. 98.