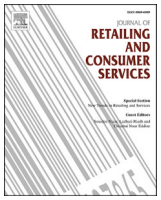
Journal of Retailing and Consumer Services 85 (2025) 104310

Contents lists available at ScienceDirect

Journal of Retailing and Consumer Services

journal homepage: www.elsevier.com/locate/jretconser



Multicategory choice modeling by recurrent neural nets

Harald Hruschka

Faculty of Business, University of Regensburg, Universitätsstraße 31, Regensburg, 93953, Germany

ARTICLE INFO

Keywords:
Multicategory choice
Market basket analysis
Neural nets
Optimization

ABSTRACT

In multicategory choice, a customer may purchase multiple products or product categories at the same time. Hidden variables of recurrent nets depend on current inputs and hidden variables of the previous period. We investigate the three main variants of recurrent neural nets, which we compare to multilayer perceptrons and multivariate logit models. Model evaluation is based on binary cross-entropies for a holdout sample. We restrict further analyses to the best non-recurrent model, a multilayer perceptron, and the best performing recurrent neural net, which both include category-specific advertising (features) as inputs. We interpret these two models looking at category dependences and feature effects. Category dependences measure the strength of either complementary or substitutive relations. We show what the stronger dependences inferred from the recurrent net imply for cross-selling decisions. We also compare what these two models imply for sales promotion by optimizing features. For the multilayer perceptron we obtain features for each category, which are constant across weeks, equaling either zero or the maximum value. For the recurrent net, features assume many intermediate values and vary considerably across weeks. To illustrate managerial implications of the recurrent net, we determine weekly features for six selected categories that differ as much as possible from each other. Finally, we discuss limitations of our approach and opportunities for future research.

1. Introduction

In multicategory choice, a customer may purchase multiple products (product categories) at the same time. The set of products (categories) a customer acquires at the same time is also known as market basket. In single category choice, on the other hand, a customer selects one product from a set of alternatives.

For a long time, models for multicategory choice did not include dynamic effects. To the best of our knowledge, only Gabel and Timoshenko (2022) and Hruschka (2022) considered dynamic effects by a fixed time lag of recent purchases and by exponentially smoothing past purchases, respectively. We focus on recurrent neural nets as they allow reproducing dynamic effects in a more flexible way.

Our work differs from previous publications using recurrent nets or transformers (Yu et al., 2016; Bai et al., 2018; Le et al., 2019; van Maasakkers et al., 2023; Gabel and Ringel, 2024) in the following respects:

- We apply not only one, but all three main variants of recurrent neural nets.
- We compare these recurrent nets to one dominant econometric model, the multivariate logit (MVL) and to a classic static neural net, the multilayer perceptron (MLP).
- We add category-specific marketing variables (i.e., sales promotion variables) as input to our neural nets. These marketing variables affect hidden variables. Therefore, for the investigated recurrent nets current marketing variables affect future choices.
- We derive dependences between product categories based on the best performing static and recurrent models. These dependences may be either substitutive or complementary. We illustrate cross-selling marketing implications for stronger dependences.
- We determine average effects of category-specific sales promotions for these two models.
- We compare the implications for sales promotion for these two models. To this end we optimize category-specific sales promotion variables.

We give a literature review of publications that use neural nets and econometric models to analyze multicategory choice data in Section 2. In Section 3 we specify the investigated models. We explain how we

E-mail address: harald.hruschka@ur.de.

https://doi.org/10.1016/j.jretconser.2025.104310
Received 28 November 2024; Received in revised form 14 April 2025; Accepted 15 April 2025
Available online 22 April 2025

estimate models and how we evaluate their statistical performances in Section 4. In Section 5 we show how we interpret models using dependences between categories and average effects of category-specific sales promotion variables. Section 6 deals with determining optimal sales promotion variables across time by an evolutionary algorithm. We characterize the data by means of descriptive statistics in section 7.1 and evaluate model performances in section 7.2.

We interpret estimation results in section 7.3. We present optimization results based on the two selected models in section 7.4. In the last section 8 we summarize results and also discuss limitations of our approach as well as opportunities for future research.

2. Literature review

Our review deals with two literature streams. We start with neural nets, emphasizing their recurrent variants. We continue by reviewing the dominant econometric models that have been used to analyze multicategory choice in marketing.

2.1. Neural nets

Neural nets have been applied in marketing research since the 1990s (for an overview of early work, see Hruschka (2008)). None of these earlier publications considered multicategory choice.

Gabel and Timoshenko (2022) develop a neural net for multicategory choice that accounts for dynamics. To determine purchase history statistics, these authors apply several linear time series filters with equal weights across products to recent product purchases, i.e., purchases up to a fixed time lag. Purchase history statistics result from transforming the outputs of these filters by a neural activation function. Three bottleneck layers compress purchase histories, frequencies of older purchases and current coupons for each product, respectively. These bottleneck layers serve to reproduce dependences and coupon cross-effects between products. Finally, a binary logistic function computes the purchase probability of each product for each period. Inputs of this binary logistic function are both product-specific variables (product constant, purchase history, old purchase frequency, current discount) and outputs of bottleneck layers. Except for product constants, coefficients of each binary logistic function do not differ between products.

Gabel and Timoshenko (2022) compare their neural net to less complex models. One of these models is an independent binary logit model that completely ignores product dependences. Another model results from a two-step approach in which these authors add similarities between latent product and customer attributes as input variables to account for product dependences. In the first step, these latent product attributes are determined by another neural net that uses co-occurrences of products in market baskets. In a binary optimization exercise based on simulated data Gabel and Timoshenko (2022) select for each investigated model the one product for each customer in each period which leads to the greatest revenue uplift for couponing.

Recurrent neural nets are capable to reproduce dynamic effects in a more flexible way compared to recent purchases with a fixed time lag (Gabel and Timoshenko, 2022) or exponentially smoothed past purchases (Hruschka, 2022). Recurrent nets encompass hidden variables that depend on current input variables and on hidden variables of the previous period. As main variants of recurrent nets, simple recurrent nets (SRNNs), long-short time memory (LSTM) nets and gated recurrent units (GRU) nets can be distinguished. We explain these variants in more detail in section 3.1.

Marketing applications of recurrent neural nets as a rule improve on more conventional models. Dhillon and Aral (2021) infer users' interests as a time-varying convex combination of hidden variables from clickstream data on online news offered by a newspaper. The model corresponds to a SRNN, which processes the words in news' headlines that are read by users. Hong and Hoban (2022) develop a tool to revise essays for the application of donations based on two sets of GRU nets that deal with word sequences and sentence sequences, respectively. Essays revised by this tool turn out to be more successful. Mena et al. (2023) predict churn, i.e., a company's loss of customers, using as input variables recency, frequency, and monetary values (RFM) together with socio-demographics. Cross-validation shows that GRU nets often outperform LSTM nets.

A few publications apply recurrent nets to direct marketing problems. Salehinejad and Rahnamayan (2016) predict RFM variables by recurrent nets, with LSTM nets outperforming SRNNs. Quite contrary, the following two applications do not rely on feature engineering and consequently do not use RFM variables as inputs. Sarkar and De Bruyn (2021) apply LSTM nets for two output variables, likelihood of a positive response and the donation amount of a charity.

Valendin et al. (2022) model purchases in several categories as well as charity donations by means of LSTM nets. Toth et al. (2017) analyze clickstream data for one website to model three output variables (purchase, abandoned shopping-cart, browsing-only) with page type and viewing time as inputs. A LSTM net outperforms a SRNN. Similarly, Sheil et al. (2018) investigate a binary classification (purchase vs. non-purchase sessions in ecommerce) with categorical input variables such as product category, price, and time stamp. These authors do not obtain great performance differences between SRNNs, LSTM, and GRU nets, which they explain by the high number of short sessions in their data. Please note that the models of Toth et al. (2017), Sheil et al. (2018), and Valendin et al. (2022) do not take dependences between products into account.

We are aware of the following four publications applying recurrent nets to multicategory choice data. Yu et al. (2016) represent categories in a low dimensional space for each user and basket. They obtain aggregate values as maximum or average across dimensions. Hidden variables are formed by a SRNN that depends on aggregate values and previous values of hidden variables. Their model focuses on binomial logit functions with the difference of scores for items included and items not included in a basket as arguments. These scores are obtained by multiplying aggregate values and hidden variables. In Le et al. (2019) the hidden variables of a LSTM net also depend on a correlation matrix computed from co-occurrence frequencies of category pairs. Le et al. (2019) use binomial logit functions similar to Yu et al. (2016). Bai et al. (2018) consider two latent representations with the same dimensionality, one for categories, the other one for groups of categories. Hidden variables are formed by a LSTM net. The model focuses on softmax (multinomial logit) probabilities for the items belonging to a basket. van Maasakkers et al. (2023) deal with next basket prediction using a GRU network to compute hidden variables. These authors also consider covariates by adding a single network layer to their model. Effects of covariates are static, as they are not related to hidden variables. The models of van Maasakkers et al. (2023) focus on binomial logit functions for categories that (do not) belong to a basket.

Full-fledged transformer models are non-recurrent. Nonetheless, these neural nets are capable to reproduce dynamics effects as they obtain sequential information by positional encoding and weighting different parts of input by self-attention mechanisms (Vaswani et al., 2017). Recently, Gabel and Ringel (2024) analyze market baskets by a stripped-down transformer without positional encoding. Therefore, their transformer represents a static neural net.

The publications mentioned in the previous two paragraphs do not measure the effect of marketing variables on purchases. These publications also do not give results on complementary or substitutive relations between pairs of products, except for Gabel and Ringel (2024). In Gabel and Ringel (2024) attention weights measure the strength of complementary relations for product pairs. Please note that these attention weights do not allow for substitutive relations.

Table 1
Main Characteristics of Econometric Models and Neural Nets.

	Dependence of purchase probabilities on	Dynamic effects
Econometric Models		
Multivariate probit model	a constant error covariance	no
Multivariate logit model	two-way interactions with other purchased categories	no
Multivariate logit model with category loyalties	two-way interactions with other purchased categories	by exponentially smoothed past purchases
Neural Nets		
Transformer with no positional encoding (Gabel and Ringel, 2024), Multilayer Perceptron	several hidden variables	no
Neural net of Gabel and Timoshenko (2022)	several hidden variables	by transformed linear filters of recent purchases with a fixed time lag
Recursive neural nets	several hidden variables	by hidden variables that are also affected by hidden variables of the previous period
Model flexibility: several hidden variables > two-way interactions > constant error covariance; recurrent hidden variables > transformed filters with a fixed time lag > exponentially smoothed purchases		

2.2. Econometric models

We now review publications using the two econometric models predominant in multicategory choice analysis, the multivariate logit (MVL) and the multivariate probit (MVP) model.

The MVL model is frequently applied to multicategory choice data. It allows for two-way interactions between purchases of different product categories (Hruschka et al., 1999; Russell and Petersen, 2000; Boztuğ and Hildebrandt, 2008; Boztuğ and Reutterer, 2008; Dippold and Hruschka, 2013b; Aurier and Mejia, 2014; Richards et al., 2018; Solnet et al., 2016; Hruschka, 2022). A positive two-way interaction exists if the purchase of category j_1 increases the purchase probability of another category j_2. For example, the purchase of hot dogs could increase the purchase probability of mustard. In a negative two-way interaction, on the other hand, the purchase of category j_1 decreases the purchase probability of another category j_2 (e.g., the purchase of beer could decrease the purchase probability of coffee).

Hruschka et al. (1999) start from the MVL model with all two-way interactions, followed by greedy stepwise backward elimination that stops if only significant interactions remain. The approach of Boztuğ and Reutterer (2008) consists of two steps. In the first step, these authors cluster market baskets by an online K-means algorithm. In the second step, they estimate one MVL model for the categories assigned to a cluster. This approach only allows two-way interactions between the categories of a cluster and sets interactions with categories belonging to other clusters to zero. Dippold and Hruschka (2013b) determine significant interactions by Bayesian variable selection techniques.

The multivariate probit (MVP) model represents an alternative multicategory choice model, which is quite often applied to market basket data (Chib et al., 2002; Duvvuri et al., 2007; Manchanda et al., 1999; Hruschka, 2017; Aurier and Mejia, 2014). The MVP model reproduces dependences between categories by correlations of error terms. Error terms are assumed to follow a multivariate normal distribution whose parameters are constant across time.

The majority of multicategory choice models do not include dynamics. One exception is the multivariate logit (MVL) model of Hruschka (2022) that includes category loyalties, i.e., exponentially smoothed past purchases, as additional inputs and clearly outperforms its basic static variant.

2.3. Main characteristics of econometric models and neural nets

We summarize the main characteristics of both econometric models and neural nets in Table 1. Please note that this table also contains the multilayer perceptron, a static neural net, which we will be presenting in Section 3.1.

Neural nets allow more flexibility in representing the interdependence of product categories by latent variables compared to the error covariance or the two-way interactions of econometric models. In addition, certain neural nets attain more flexibility in reproducing dynamic effects. This latter advantage especially applies to recursive neural nets.

3. Investigated models

J column vector y_{mt} consists of binary purchase indicators (J symbolizes the number of product categories). If household m purchases category j in period t, the respective element y_{jmt} equals one. In the following we explain how we specify the investigated neural nets and MVL models. The binary logistic function defined as $1/(1+\exp(-V))$ with output values in $[0,1]$ is denoted as $\sigma(V)$. $\tanh(V)$ symbolizes the hyperbolic tangent function $(exp(V)-exp(-V))/(exp(V)+exp(-V))$ with output values in $[-1,1]$.

3.1. Neural nets

The purchase probability of category j of household m in period t conditional on the $(K,1)$-dimensional vector of hidden variables h_{mt} is computed as follows:

$$\tilde{P}_{jmt} \equiv P(y_{jmt}=1|h_{mt}) = \sigma(b_j + W_j h_{mt}) \quad (1)$$

b_j is the constant for category j, W_j the $(1,K)$-dimensional coefficient vector relating hidden variables to purchases of category j. The hidden variables (neurons) h_{mt} on their turn are linked to a $(2J,1)$ dimensional input vector x_{mt} that consists of J observed purchase indicators y_{jmt} and J category-specific sales promotion variables f_{jt} for period t.

Note that purchase incidences y_{jmt} of each category are used to compute hidden variables h_{mt} that in their turn determine conditional purchase probabilities (also see the equations for hidden variables and output variables in Yu et al. (2016) and Le et al. (2019)). These computations may seem to be circular but do not pose a problem if the number of hidden variables is much lower than the number of categories.

Neural nets that ignore category-specific sales promotion variables by being linked to the observed purchase indicators only, can be seen as autoencoders (also known as auto-associative nets). Autoencoders provide compressed representations of purchases (Bishop, 1995; Goodfellow et al., 2016). Autoencoders are similar to factor analytic methods, restricted Boltzmann machines and topic models (for more details on these models and their application to market basket analysis see Hruschka (2021) and Hruschka (2014)).

Hidden variables of recurrent nets depend both on current input variables and on hidden variables of the previous period. Therefore, recurrent neural nets are able to consider purchase event feedback effects

(i.e., effects of past purchases on current purchases). Recurrent nets also reproduce indirect effects that arise if marketing variables directly influence hidden variables that in their turn affect future hidden variables. Then these hidden variables have an effect on future purchases. Such indirect effects can be reproduced if marketing variables are added to the inputs of a recurrent net. van Maasakkers et al. (2023) mention that such an extension constitutes an interesting research effort. An extensive literature demonstrates the importance of feedback effects in marketing (Heckman, 1981; Guadagni and Little, 1983; Erdem and Keane, 1996). With respect to indirect effects, it has been shown that, e.g., sales promotion actions as well as frequent or large direct mailings tend to reduce purchase event feedback (Gedenk and Neslin, 1999; van Diepen et al., 2009; Schröder and Hruschka, 2016).

We look at three variants of recurrent neural nets that differ with respect to the computation of hidden variables, namely simple recurrent neural nets (SRNNs), gated recurrent units (GRU) nets and long-short term memory (LSTM) nets. In these recurrent nets constant vectors, input coefficient matrices, and recurrent coefficient matrices have dimensions $(K, 1)$, (K, I) and (K, K), respectively. The SRNN, also known as Elman's net (Elman, 1990), computes the vector of hidden variables for period t as:

$$h_{mt} = \tanh(b_h + W_h x_{mt} + R_h h_{mt-1}) \tag{2}$$

b_h, W_h, and R_h denote the constant vector, the input coefficient matrix and the recurrent coefficient matrix, respectively.

SRNNs have difficulties with long term dependences (Bengio et al., 1994), as gradients tend to either explode or vanish. This weakness of SRNNs becomes problematic if purchase event feedback is at work. Hochreiter and Schmidhuber (1997) introduced LSTM nets to tackle long term dependences. LSTM nets include input gate units i_{mt}, forget gate units f_{mt}, candidate cell states $\tilde{C}_{mt}$, and output units o_{mt}. These units depend on current inputs x_{mt} and previous hidden variables h_{mt-1} in the following way:

$$\begin{aligned} i_{mt} &= \sigma(b_i + W_i x_{mt} + R_i h_{mt-1}) \\ f_{mt} &= \sigma(b_f + W_f x_{mt} + R_f h_{mt-1}) \\ \tilde{C}_{mt} &= \tanh(b_C + W_c x_{mt} + R_c h_{mt-1}) \\ o_{mt} &= \sigma(b_o + W_o x_{mt} + R_o h_{mt-1}) \end{aligned} \tag{3}$$

b_i, b_f, b_c, b_o denote constant vectors of these units, W_i, W_f, W_c, W_o their input coefficient matrices and R_i, R_f, R_c, R_o their recurrent coefficient matrices. $\odot$ stands for elementwise multiplication.

The current cell states C_{mt} are determined as:

$$C_{mt} = f_{mt} \odot C_{mt-1} + i_{mt} \odot \tilde{C}_{mt} \tag{4}$$

For $f_{mt} = 1$ ($f_{mt} = 0$) the old cell state is kept (eliminated), for $i_{mt} = 1$ ($i_{mt} = 0$) the candidate cell state is completely considered (ignored). Finally, the current hidden variables depend on current cell states C_{mt} filtered by output gate units o_{mt}:

$$h_{mt} = o_{mt} \odot \tanh(C_{mt}) \tag{5}$$

The GRU net (Cho et al., 2014) simplifies the LSTM net by combining the forget and input gates into a single update gate and merging cell states and hidden variables. This net has constant vectors b_r, b_h, b_z, input coefficient matrices W_r, W_h, W_z, and recurrent coefficient matrices R_r, R_h, R_z. Reset gate units $\tilde{r}_{mt}$ and candidate variables $\tilde{h}_{mt}$ are computed as:

$$\tilde{r}_{mt} = \sigma(b_r + W_r x_{mt} + R_r h_{mt-1}) \tag{6}$$

$$\tilde{h}_{mt} = \tanh(b_h + W_h x_{mt} + R_h[r_{mt} \odot h_{mt-1}] \tag{7}$$

For r_{mt} close to 0 the candidate hidden variables are determined primarily by the new input.

Update gate units z_{mt} and hidden variables h_{mt} are computed as follows:

$$z_{mt} = \sigma(b_z + W_z x_{mt} + R_z h_{mt-1}) \tag{8}$$

$$h_{mt} = z_{mt} \odot h_{mt-1} + (1 - z_{mt}) \odot \tilde{h}_{mt} \tag{9}$$

For z_{mt} close to 1 the hidden variables are determined primarily by their previous values.

In the empirical study we compare these recurrent nets to several, less complex benchmark models, e.g., multilayer perceptrons (MLPs). MLPs have since the late 1980s been applied to marketing problems (Hruschka, 2008). As a rule, MLPs ignore purchase event feedback as their hidden variables depend on current inputs only:

$$h_{mt} = \tanh(b_h + W_h x_{mt}) \tag{10}$$

We also investigate whether recurrent nets outperform a MLP with category loyalties as additional input, as the latter offer a parsimonious way to consider purchase event feedback. We compute the loyalty of household m for category j in market basket t in analogy to exponentially smoothed brand loyalties (Guadagni and Little, 1983):

$$loy_{jmt} = \alpha\, y_{jmt-1} + (1 - \alpha)\, loy_{jmt-1} \tag{11}$$

$0 \le \alpha \le 1$ denotes the smoothing constant. The binary purchase incidence y_{jmt-1} equals one, if household m purchases category j in the previous period $t-1$.

The current category loyalty depends on the previous purchase incidence y_{jmt-1} and the previous loyalty loy_{jmt-1}. The lower smoothing constant α is, the less the loyalty variable reflects fluctuating purchases. In a manner similar to the brand loyalty of Guadagni and Little (1983) we set initial values loy_{jm0} equal to the relative purchase frequency of the respective category j across all households and periods.

We also estimate MLPs, SRNNs, GRU nets, and LSTM nets without category-specific sales promotions as inputs.

3.2. Multivariate logit models

The multivariate logit (MVL) model encompasses pairwise interactions between category purchases. In contrast to the neural nets presented in Section 3.1 MVL models do not include hidden variables. Inputs of our multivariate logit models may consist of category-specific sales promotions and category loyalties, which we explained in Section 3.1. Taking category loyalties into account has led to improved performance in Hruschka (2022).

Maximum likelihood estimation of MVL models is not feasible even for a moderate number of categories. That is why we resort to pseudo-probabilities, which are equivalent to conditional probabilities of purchases or non-purchases for the individual product categories (Bel et al., 2018).

We investigate two variants of MVL model, the homogeneous MVL model and its finite mixture extension FM-MVL. The homogeneous MVL model has been applied to market basket data by Russell and Petersen (2000) building upon earlier publications in statistics (Cox, 1972; Besag, 1974). Dippold and Hruschka (2013a) introduced the FM-MVL model to market basket analysis.

For the homogeneous MVL model, we write the purchase probability of category j of household m in period t conditional on purchases of the other categories collected in vector y_{-jmt} and other regressors x_{mt} such as category sales promotions pr_{jt} and loyalties loy_{jmt}:

$$\begin{aligned} &\tilde{P}_{jmt} \equiv P(y_{jmt} = 1 | y_{-jmt}, x_{mt}) = \sigma(Z_{jmt}) \\ &\text{with} \quad Z_{jmt} = \beta_j^0 + \beta_j^1 pr_{jt} + \beta_j^2 loy_{jmt} + \sum_{l \neq j} \beta_{j,l}^3 y_{lmt} \end{aligned} \tag{12}$$

Z_{jmt} can be interpreted as latent variable referring to category j, household m and period t. β_j^0 is a category constant, β_j^1 and β_j^2 are the

coefficients for sales promotions and loyalty of category j. Interaction coefficients are symmetric, i.e., $\beta^3_{j1,j2} = \beta^3_{j2,j1}$.

Coefficients of the FM-MVL model differ between household segments. The purchase probability of category j of household m in period t conditional on purchases of the other categories and regressors x_{mt} is:

$$\tilde{P}_{jmt} \equiv P(y_{jmt} = 1 | y_{-jmt}, x_{mt}) = \sum_{s=1}^{S} u_{sm} P_s(y_{jmt} = 1 | y_{-jmt}) = \sum_{s=1}^{S} u_{sm} \sigma(Z_{sjmt})$$
$$\text{with} \quad Z_{sjmt} = \beta^0_{sj} + \beta^1_{sj} pr_{jt} + \beta^2_{sj} loy_{jnmt} + \sum_{l \neq j} \beta^3_{sj,l} y_{lmt} \tag{13}$$
$$\text{for} \quad s = 1, \cdots, S$$

S denotes the number of segments. u_{sm} is a binary membership indicator set to one if household m is assigned to segment s. P_s is the segment-specific conditional probability function. The latent variable Z_{sjmt} refers to segment s, category j period t and household m. Due to the binary membership indicators, the same segment-specific conditional probability function is used for all market baskets of any household m.

We also estimate the MVL model and the FM-MVL model with pairwise interactions only excluding sales promotions and loyalties. Similar to autoencoders, the MVL model with interactions only is also known as auto-logistic model (Besag, 1972).

4. Estimation and evaluation of models

We estimate the investigated models by minimizing the binary cross-entropy (Bishop, 1995):

$$-1/(JMT) \sum_{j=1}^{J} \sum_{m=1}^{M} \sum_{t=1}^{T} [y_{jmt} \log \tilde{P}_{jmt} + (1 - y_{jmt}) \log(1 - \tilde{P}_{jmt}] \tag{14}$$

Expression (14) shows that binary cross-entropy corresponds to the negative log-likelihood averaged across households, baskets, and categories. The binary cross-entropy can be interpreted as distance of the respective model to the true distribution. A value of zero indicates that the model and the true distribution agree completely. To estimate neural nets, we use the open source library Keras (Chollet et al., 2015) and its implementation of Adam, a stochastic gradient descent method that is based on adaptive estimation of first-order and second-order moments (Kingma and Ba, 2015).

Estimation of homogeneous MVL models turns out to be straightforward because its binary cross-entropy function has only one local minimum. On the contrary, the FM-MVL model may be prone to multiple local maxima. That is why we start estimation of the FM-MVL models ten times by randomly assigning each household to one of S segments. Our estimation approach for the FM-MVL model is akin to maximizing the classification likelihood (McLachlan and Basford, 1988; Ngatchou-Wandji and Bulla, 2013) replacing the intractable likelihood by segment-specific pseudo-probabilities. We estimate homogeneous MVL models as well as segment-specific MVL Models by means of the BFGS algorithm contained in the Optimize module of the Python package SciPy (Virtanen et al., 2020).

The binary cross-entropy for holdout data serves to evaluate models. This way we consider the complexity of models. A model, whose complexity is too high, leads to a worse (higher) cross-entropy for the holdout data. In contrast to information criteria such as AIC or BIC, holdout validation has the advantage to do without assumptions about the true underlying model. Our holdout data encompass all baskets of a randomly selected one third of the households. This approach is more stringent compared to that of previous publications applying recurrent nets to multicategory choice data. These publications form holdout data by considering for each household only either the last basket (Yu et al., 2016; Bai et al., 2018; van Maasakkers et al., 2023) or all baskets falling in the last half-month (Le et al., 2019).

Following suggestions of one anonymous reviewer, we also investigate the predictive performance of models both separately for each category and with respect to market basket composition. Firstly, we determine which category-specific purchase prediction benefit the most (the least) from incorporating dynamic cross-category effects and dynamic promotional effects. To this end we compute log-likelihood differences lld_{jmt} between a recurrent net and a static model for each category j, household m and basket t using the holdout data:

$$lld_{jmt} = [y_{jmt} \log \tilde{P}^1_{jmt} + (1 - y_{jmt}) \log(1 - \tilde{P}^1_{jmt}] - [y_{jmt} \log \tilde{P}^0_{jmt} + (1 - y_{jmt}) \log(1 - \tilde{P}^0_{jmt}] \tag{15}$$

$\tilde{P}^1_{jmt}$ and $\tilde{P}^0_{jmt}$ denote the purchase probabilities for the recurrent net and the static model, respectively.

We determine the mean difference as $1/(MT) \sum_m \sum_t lld_{jmt}$ for each category j. A positive (negative) mean difference measures to what extent the recurrent net provides better (worse) predictions than the static model for the respective category across households and baskets.

Secondly, we demonstrate a model's performance in predicting the composition of baskets. To this end we compute the average predicted rank of the purchased categories similar to van Maasakkers et al. (2023). In contrast to these authors, who consider only one test basket for each household, we take all baskets of the holdout data set into account as shown in the following expression:

$$1/(MT) \sum_{m=1}^{M} \sum_{t=1}^{T} 1/size_{mt} \sum_{j=1}^{J} y_{jmt} \, rank(\tilde{P}_{jmt} | \tilde{P}_{1mt}, \tilde{P}_{2mt}, \cdots, \tilde{P}_{Jmt})$$
$$\text{with} \quad size_{mt} = \sum_{j=1}^{J} y_{jmt} \tag{16}$$

$rank(\tilde{P}_{jmt} | \tilde{P}_{1mt}, \tilde{P}_{2mt}, \cdots, \tilde{P}_{Jmt})$ denotes the rank of the probability of category j in basket t of household m with respect to the probabilities of all categories for the same basket and household where a rank of 1 is assigned to the highest purchase probability. The average rank of the probabilities of the categories contained in each basket of each household is determined as $1/size_{mt} \sum_{j=1}^{J} y_{jmt} \, rank(\tilde{P}_{jmt} | \tilde{P}_{1mt}, \tilde{P}_{2mt}, \cdots, \tilde{P}_{Jmt})$ with $size_{mt}$ as basket size (number of categories in the basket). Averaging once more across households and baskets, we finally obtain the average predicted rank of purchased categories. The lower the average rank, the better a model performs in finding out which categories are relevant for households.

5. Model interpretation

We interpret models by looking at average dependences of ordered category pairs and average effects of category-specific sales promotions. Let us remind you that the specifications of the investigated models refer to the conditional probability of each category (see Section 3). If purchases of product categories were independent, we could simply process each conditional probability expression separately in the further analyses (independence of categories applies, e.g., to the independent logit model that corresponds to the MVL model without interaction terms). On the other hand, the conditional probabilities of the investigated models depend on (functions of) product categories. Therefore, we have to rely on the joint distribution (Bee et al., 2015). We determine joint probabilities of product categories by mean field approximation, a method that often serves to reduce computation times for estimation and inference of neural nets (Peterson and Anderson, 1987; Haykin, 1994). Mean field approximation consists in replacing binary purchase indicators by purchase probabilities. We solve the resulting nonlinear equation system by fixed point iterations as described by Algorithm 1.

$P_0(jmt)$ and $P_1(jmt)$ denote the purchase probability of category j for household m in period t for the previous and the current iterations, respectively. Fixed point iterations stop, if the maximum absolute difference xad of purchase probabilities between two successive iterations

set initial probabilities equal to relative marginal frequencies
for all j, m, t **do**
  $P_0(jmt) \leftarrow \bar{y}(j)$
end for
$xad \leftarrow 1.0$
while $xad \geq 0.0001$ **do**
  $xad \leftarrow 0.0$
  for all j, m, t **do**
    $P_1(jmt) \leftarrow g_j(P_0(jmt), pr_t)$
    $ad \leftarrow abs(P_1(jmt) - P_0(jmt))$
    if $ad > xad$ **then**
      $xad \leftarrow ad$
    end if
    $P_0(jmt) \leftarrow P_1(jmt)$
  end for
end while

Algorithm 1: Determining Joint Purchase Probabilities.

across $J \times M \times T$ constellations is less than 0.0001. $g_j()$ denotes the conditional probability of category j according to a (recurrent) neural net, a MVL model or a FM-MVL model with purchase probabilities of the previous iteration $P_0(jmt)$ and current sales promotions $pr_t = (pr_{1t}, \cdots, pr_{Jt})$ as inputs.

We measure pairwise dependence by the average probability change of category j due to a marginal increase of the probability of another category $j' \neq j$. We define two categories to be purchase complements if the probability change is positive and to be purchase substitutes if the probability change is negative. This definition is equivalent to the one put forward by Betancourt and Gautschi (1990), who consider two products as purchase complements (purchase substitutes) if they are purchased jointly more (less) frequently than expected under stochastic independence. We investigate ordered category pairs, as dependences are as a rule asymmetric, i.e., the probability change of category $j' \neq j$ due to a marginal increase of the probability of category j differs from the probability change of category j due to a marginal increase of the probability of category j'.

We determine purchase probabilities $P(y_{jmt})$ by Algorithm 1. Then we increase the purchase probability $P(j'mt)$ of each category j' by a small value δ. Running Algorithm 1 again now holding the purchase probability of category j' fixed at $P(j'mt) + \delta$, we obtain modified purchase probabilities $P(y_{jmt}, P(j'mt) + \delta)$ for each of the other categories $j \neq j'$. The average dependence of any other category j follows as:

$$1/(MT) \sum_{m=1}^{M} \sum_{t=1}^{T} P(y_{jmt}, P(j'mt) + \delta) - P(y_{jmt}) \tag{17}$$

Our second interpretation approach uses average marginal effects of sales promotions on purchase probabilities. We increase the observed sales promotion variable of a category j' by a small value δ, keeping all the other input variables at their observed values. The average marginal effect of features of category j' on purchases of category j can be written as:

$$1/(MT) \sum_{m=1}^{M} \sum_{t=1}^{T} P(y_{jmt}, pr_t^{+j'}) - P(y_{jmt}, pr_t) \tag{18}$$

$pr_t^{+j'}$ equals pr_t except for category j' where it is $pr_{j't} + \delta$. $P(y_{jmt}, pr_t^{+j'})$ and $P(y_{jmt}, pr_t)$ denote the joint purchase probabilities of household m for category j in period t if the sales promotion variable of category j' is increased and kept at the observed value, respectively. To determine purchase probabilities for higher sales promotion variables we replace $g_j(P_0(jmt), pr_t)$ by $g_j(P_0(jmt), pr_t^{+j'})$ in Algorithm 1. We distinguish own effect and cross effects. The former are effects on the same category ($j = j'$), the latter are effects on another category ($j \neq j'$).

Table 2
Product Categories and Abbreviations.

Beer & ale	beer	Blades	blades
Carbonated beverages	carbbev	Cigarettes	cigets
Coffee	coffee	Cold Cereal	coldcer
Deodorant	deod	Diapers	diapers
Facial tissue	factiss	Frozen dinners	fzdin
Frozen pizza	fzpizza	Household cleaners	hhclean
Frankfurters & hotdog	hotdog	Laundry detergent	laundet
Margarine & butter	margbutr	Mayonnaise	mayo
Milk	milk	Mustard & ketchup	mustketc
Paper towels	paptowl	Peanut butter	peanbutr
Photographic supplies	photo	Razors	razors
Salty snacks	saltsnck	Shampoo	shamp
Soup	soup	Spaghetti sauce	spagsauc
Sugar substitutes	sugarsub	Toilet tissue	toitisu
Tooth brush	toothbr	Toothpaste	toothpa
Yogurt	yogurt		

6. Optimization

We want to set promotions in each week maximizing the average revenue per basket, an objective that is proportional to total revenue. Frequently, revenue maximization rather than profit maximization is consistent with actual behavior of oligopolistic firms (Baumol, 1967). As we explain below, the computation of average revenue per basket is based on estimated joint purchase probabilities. For dynamic multicategory choice models such as the MVL with category loyalties or recurrent neural nets, we expect that optimal sales promotion variables are not constant across periods.

We opt for an evolutionary algorithm that stores diverse sets of solutions in each iteration. Storing solution sets provides a means to escape from local optima and to cope with large and discontinuous search spaces (Eiben and Smith, 2015). Local optima may occur both for (recurrent) neural nets and MVL models.

As the promotion variables of our empirical study are real valued shares (see Section 7.1), we choose a so-called evolutionary strategy. Specifically, we determine optimal feature values by means of the covariance matrix adaptation evolution strategy (CMA-ES), which is one of the leading algorithms for difficult real valued functions (Eiben and Smith, 2015). CMA-ES samples the new candidate solutions according to a multivariate normal distribution. Dependences between decision variables are kept in a covariance matrix that increases (decreases) in the case of low (high) confidence (see Hansen and Ostermeier (2001) and Hansen (2023) for more details). For computations we use the implementation of CMA-ES in the Python module cma (Hansen, 2018).

We input sales promotion variables contained in a J-dimensional vector $pr(t)$ for each week $t = 1, \cdots, T$ to the fixed-point iterations of Algorithm 1. This algorithm returns the joint purchase probabilities that are needed in each evaluation step of the CMA-ES. Finally, we obtain the average revenue per basket as $\sum_{j=1}^{J} rev_j \, 1/(MT) \sum_{m=1}^{M} \sum_{t=1}^{T} P_1(j, m, t)$ with rev_j denoting the observed average revenue of a purchase of category j.

7. Empirical study

7.1. Data

Our data refer to the purchases made by a random sample of 1500 households in one specific grocery store. We compose weekly market baskets over a one-year period from the IRI data set (Bronnenberg et al., 2008). We represent a market basket by a binary vector whose elements indicate whether a household purchases any of 31 product categories in a week (see Table 2). Note that we also include zero baskets that result if households do not purchase any of the 31 categories in a week. The average basket size (i.e., the number of purchased categories) is 1.189, its standard deviation is 2.311.

Table 3
Relative Marginal Frequencies.

milk	0.147	carbbev	0.123	saltsnck	0.108	coldcer	0.086	yogurt	0.062
soup	0.061	spagsauc	0.057	toitisu	0.053	margbutr	0.049	paptowl	0.043
coffee	0.042	laundet	0.036	fzpizza	0.034	mayo	0.034	hotdog	0.032
mustketc	0.031	fzdin	0.028	factiss	0.026	peanbutr	0.025	beer	0.023
toothpa	0.018	shamp	0.016	deod	0.012	cigets	0.010	hhclean	0.009
diapers	0.006	blades	0.006	toothbr	0.004	sugarsub	0.003	photo	0.002
razors	0.001								

Table 4
Average Features (Advertising in Local Newspapers and Flyers).

fzdin	0.196	spagsauc	0.179	diapers	0.165	coffee	0.164	saltsnck	0.148
peanbutr	0.141	mayo	0.140	yogurt	0.137	fzpizza	0.129	factiss	0.128
carbbev	0.114	beer	0.100	paptowl	0.097	margbutr	0.095	laundet	0.094
deod	0.094	razors	0.091	soup	0.081	shamp	0.079	coldcer	0.075
toitisu	0.074	toothpa	0.073	milk	0.059	hotdog	0.058	hhclean	0.042
photo	0.040	blades	0.036	mustketc	0.020	toothbr	0.016	sugarsub	0.009
cigets	0.000								

Table 3 shows relative marginal purchase frequencies for the 31 categories. Milk is the category most frequently purchased.

The category-specific sales promotion variables that we consider in our models are features, i.e., advertising in local newspapers or flyers. We measure features as weekly market share-weighted averages of UPC level variables in the respective category. Consequently, features take values between zero and one. Table 4 shows average values of features for each category. We obtain the highest (lowest) feature value for frozen dinners (cigarettes).

7.2. Model evaluation results

We base the evaluation of models on binary cross-entropies for the holdout data. Tables 5 and 6 show the results for models excluding and including features, respectively. For each model type we estimate models with $1, 2, \ldots \cdots, K, K+1$ hidden variables (segments). $K+1$ denotes the lowest number of hidden variables (segments) for which the holdout binary cross-entropy decreases by less than 0.005 compared to its value for K hidden variables (segments). Finally, we select the model with K hidden variables (segments) for each type.

Of course, we should not consider any model that does not outperform the simplest model consisting of category constants only. This simplest model is outperformed by all the other models, no matter whether they include features or not. Our evaluations show that homogeneous MVL models (= FM-MVL models with one segment) are sufficient. A smoothing constant $\alpha = 0.1$, which puts most weight on the loyalty of the previous week leads to the best performing MVL model with category loyalties according to a grid search over $[0.1, 0.2, 0.3, \cdots, 0.9]$. Given such a value, past purchases are strongly smoothed.

Several results apply no matter whether features are excluded (Table 5) or included (Table 6):

- Binary cross-entropies improve (deteriorate) both in the estimation and the holdout data if the number of hidden variables (segments) increases (decreases).
- Addition of category loyalties as predictors in MVL models improves binary cross-entropies similar to Hruschka (2022). Nonetheless, MVL models are outperformed by all the neural nets investigated, even the MLP without both features and loyalties.
- GRU and LSTM nets outperform all MVL models and MLPs. This result indicates inter alia that category loyalties reproduce purchase event feedback only in a limited way.
- Both GRU and LSTM nets outperform SRNN nets, demonstrating that long term dependences are important.
- LSTM nets outperform GRU nets.

If features are excluded, the MLP with loyalties and three hidden variables turns out as the best non-recurrent model, the LSTM net with six hidden variables as best recurrent net (see Table 5). Adding category-specific features as covariates influencing hidden variables leads to clearly better model performances for both MLPs and LSTM nets. The six hidden variable LSTM net with features outperforms all the other models.

In the following analyses we only consider two models, the best non-recurrent model, the six hidden variables MLP with features, and the best recurrent net, the six hidden variables LSTM net with features (see Table 6). From now on, we simply call these two models MLP and LSTM net, respectively.

We now investigate which product categories benefit the most (and which categories benefit the least) from incorporating dynamic cross-category effects and dynamic promotional effects. Table 7 shows for each category the average log likelihood difference and the value of the t-test against the null hypotheses that it equals zero. Significant positive differences confirm that the LSTM net leads to better predictions for 26 categories. Categories with t-values greater than 20 (beer & ale, cold cereal, frozen dinner, facial tissue, frankfurters & hot dogs, laundry detergent, margarine & butter, mayonnaise, mustard & ketchup, salty snacks, soup, spaghetti sauce, toilet tissue, and yogurt) benefit the most from applying the LSTM net.

Predictions of the LSTM are significantly worse compared to the MLP for two categories only (carbonated beverages and milk). An anonymous reviewer asked for an explanation of the performance of the LSTM net for these two most frequently purchased categories (see Table 3). We think that taking dynamic effects into account may not lead to better predictions, if purchases of a category are regular. We measure purchase regularity by the coefficient of variation of interpurchase times (the number of weeks between successive purchases) across households (Dunn et al., 1983; Gupta, 1988). A low coefficient of variation indicates regular purchase timing. As we expected, purchases of carbonated beverages and milk turn out to be regular. Coefficients of variation of their interpurchase times are lower than the values for the remaining categories (see Table 8).

We demonstrate the performance of the MLP and the LSTM net in predicting basket composition on the holdout data by the average predicted rank of purchased categories (see expression (16)). We obtain average ranks amounting to 6.59 and 3.11 for the MLP and the LSTM net, respectively. The LSTM net clearly does a better job in finding out which categories are relevant for households. Let us remind you that the MLP only considers current data in contrast to the LSTM. Therefore, these results show that taking past purchase occasions and promotional activities into account leads to a clearly better performance in predicting basket composition.

Table 5
Binary Cross-Entropy Values of Models Excluding Features.

Model Type	Number of Hidden Variables or Segments	Estimation Cross-Entropy	Holdout Cross-Entropy	Holdout Cross-Entropy Decrease
Category Constants Only		0.1624	0.1633	
FM-MVL	**1**	**0.1093**	**0.1138**	
	2	0.1087	0.1141	-0.0003
FM-MVL plus Loyalties	**1**	**0.1029**	**0.1061**	
	2	0.1023	0.1060	0.0001
MLP	**1**	**0.0973**	**0.0997**	
	2	0.0929	0.0951	0.0046
MLP plus Loyalties	1	0.0982	0.1006	
	2	0.0917	0.0941	0.0064
	3	**0.0834**	**0.0858**	**0.0083**
	4	0.0818	0.0841	0.0017
SRNN	1	0.0947	0.0970	
	2	0.0889	0.0913	0.0057
	3	0.0812	0.0834	0.0079
	4	0.0760	0.0781	0.0053
	5	**0.0676**	**0.0699**	**0.0082**
	6	0.0645	0.0665	0.0034
GRU	1	0.0931	0.0955	
	2	0.0830	0.0859	0.0096
	3	0.0743	0.0773	0.0086
	4	0.0678	0.0710	0.0063
	5	0.0611	0.0644	0.0066
	6	0.0545	0.0577	0.0067
	7	**0.0488**	**0.0521**	**0.0056**
	8	0.0450	0.0485	0.0037
LSTM	1	0.0931	0.0956	
	2	0.0820	0.0853	0.0103
	3	0.0675	0.0708	0.0145
	4	0.0601	0.0637	0.0071
	5	0.0507	0.0547	0.0090
	6	**0.0435**	**0.0475**	**0.0072**
	7	0.0420	0.0463	0.0013

The selected number of hidden variables (segments) for each model type is shown in boldface. It is set to K if the holdout binary cross-entropy for K+1 hidden variables (segments) decreases by less than 0.005.

7.3. Model interpretation results

We interpret models by looking at category dependences and effects of category-specific sales promotions.

Table 9 contains the average dependences of ordered category pairs of at least 0.01 in absolute size for both the MLP and the LSTM net. A positive (negative) value shows that a purchase probability increase of the category is associated with a purchase probability increase (decrease) of the other category. In other words, a positive (negative) value indicates that two categories are purchase complements (purchase substitutes).

In the following, we discuss the dependences shown in Table 9. We obtain more dependences (also more negative dependences) for the LSTM net. The MLP implies 17 dependences, of which three are negative. The LSTM net implies 39 dependences, most of which (i.e., 20) are negative. Managers run the risk to overestimate the relative number of complementary dependences if they rely on the MLP in spite of its worse statistical performance. The two models agree on one dependence only (yogurt on coffee). All the other dependences are different.

Let us illustrate what the stronger dependences of the best performing LSTM net imply for cross-selling. Management may enhance cross-selling by appropriate positioning of categories in aisles and shelves of a store. Because of positive values the categories deodorant, frozen dinners, frozen pizza, margarine & butter, mayonnaise, mustard & ketchup, paper towels, salty snacks, soup, spaghetti sauce, sugar substitutes, and yogurt (salty snacks, soup, spaghetti sauce, sugar substitutes, and yogurt) should be positioned near to coffee (milk). On the other hand, the category beer & ale should be positioned far from coffee due to the negative dependence. Paper towels, salty snacks, soup, spaghetti sauce, sugar substitutes, and yogurt should be positioned far from frankfurters & hotdogs. On the other hand, mustard & ketchup should be positioned near to frankfurters & hotdogs.

Recommending other categories by printouts at checkout or by mobile phone messages could also foster cross-selling. Let us illustrate this approach for two examples given in the previous paragraph. Coffee (milk) may be recommended if a buyer has purchased deodorant, frozen dinners, frozen pizza, margarine & butter, mayonnaise, mustard & ketchup, paper towels, salty snacks, soup, spaghetti sauce, sugar substitutes, and yogurt (salty snacks, soups, spaghetti sauce, sugar substitutes, and yogurt).

Table 10 displays own effects for features of at least 0.005 in absolute size for both models. The MLP implies ten, the LSTM net seven such own effects. These own effects are all positive, i.e., more features increase the purchase probability of the same category.

Tables 11 and 12 contain cross effects for the MLP, Table 13 for the LSTM net. These tables only contain cross effects, which are at least 0.005 in absolute size.

We obtain more cross effects for the MLP than for the LSTM net. All higher cross effects for the MLP are positive. In contrast, the LSTM implies many negative cross effects. Features of all the other categories exert positive effects on coffee purchases. Features of household cleaners, mustard & ketchup, and sugar substitutes exert positive effects on photographic supplies. On the other hand, purchases of categories such as soup, salty snacks, and yogurt are negatively affected by features of other categories.

To illustrate the dynamics of hidden variables of the LSTM net, we determine three households as centers of a three-cluster medoid cluster analysis using the weekly values of the six hidden variables. Computa-

Table 6
Binary Cross-Entropy Values of Models Including Features.

Model Type	Number of Hidden Variables or Segments	Estimation Cross-Entropy	Holdout Cross-Entropy	Holdout Cross-Entropy Decrease
FM-MVL	**1**	**0.1092**	**0.1137**	
	2	0.1086	0.1135	0.0002
FM-MVL plus Loyalties	**1**	**0.1027**	**0.1059**	
	2	0.1022	0.1058	0.0001
MLP	1	0.0983	0.1006	
	2	0.0918	0.0941	0.0065
	3	0.0829	0.0854	0.0088
	4	0.0769	0.0792	0.0061
	5	0.0696	0.0723	0.0069
	6	**0.0606**	**0.0634**	**0.0089**
	7	0.0566	0.0595	0.0039
MLP plus Loyalties	**1**	**0.0952**	**0.0975**	
	2	0.0928	0.0950	0.0025
SRNN	1	0.0968	0.0989	
	2	0.0879	0.0899	0.0090
	3	0.0816	0.0839	0.0060
	4	0.0754	0.0776	0.0062
	5	**0.0673**	**0.0694**	**0.0082**
	6	0.0650	0.0670	0.0025
GRU	1	0.0930	0.0957	
	2	0.0837	0.0864	0.0093
	3	0.0741	0.0772	0.0092
	4	0.0655	0.0693	0.0079
	5	**0.0551**	**0.0589**	**0.0104**
	6	0.0510	0.0550	0.0039
LSTM	1	0.0927	0.0956	
	2	0.0775	0.0817	0.0139
	3	0.0627	0.0672	0.0145
	4	0.0552	0.0594	0.0079
	5	0.0450	0.0499	0.0095
	6	**0.0369**	**0.0414**	**0.0085**
	7	0.0345	0.0394	0.0020

The selected number of hidden variables (segments) for each model type is shown in boldface. It is set to K if the holdout binary cross-entropy for K+1 hidden variables (segments) decreases by less than 0.005.

Table 7
Log Likelihood Differences LSTM net - MLP: means and t-values.

beer	0.0738	30.79	blades	-0.0007	1.21	carbbev	-0.0056	6.44
cigets	0.0197	12.07	coffee	0.0321	16.71	coldcer	0.0941	44.74
deod	-0.0010	1.52	diapers	0.0023	2.35	factiss	0.0420	24.51
fzdin	0.0562	25.06	fzpizza	0.0106	9.09	hhclean	-0.0018	1.79
hotdog	0.0526	29.17	laundet	0.0396	20.44	margbutr	0.1089	53.89
mayo	0.0920	43.44	milk	-0.0499	31.61	mustketc	0.0562	29.69
paptowl	0.0350	17.51	peanbutr	0.0044	4.72	photo	0.0039	5.06
razors	0.0011	2.72	saltsnck	0.0624	29.45	shamp	0.0063	5.27
soup	0.1157	56.57	spagsauc	0.0698	32.75	sugarsub	0.0019	3.80
toitisu	0.0373	21.14	toothbr	0.0028	4.30	toothpa	0.0054	6.00
yogurt	0.1220	49.98						

significant at $\alpha = 0.05$ for t-values >= 1.96

Table 8
Interpurchase Times: coefficients of variation.

beer	1.169	blades	1.361	carbbev	0.755	cigets	1.371
coffee	0.997	coldcer	0.824	deod	1.235	diapers	1.401
factiss	1.095	fzdin	1.144	fzpizza	1.066	hhclean	1.295
hotdog	1.059	laundet	1.008	margbutr	0.949	mayo	1.037
milk	0.707	mustketc	1.049	paptowl	0.995	peanbutr	1.102
photo	1.456	razors	1.495	saltsnck	0.780	shamp	1.190
soup	0.934	spagsauc	0.917	sugarsub	1.427	toitisu	0.950
toothbr	1.400	toothpa	1.157	yogurt	0.939		

tion of the vector of K hidden variables h_{mt} can be seen from expressions (5), (3) and (4). m and t denote the household and week, respectively.

These three households show great differences with respect to the hidden variables. Plotting the weekly hidden variables for these households in Fig. 1 demonstrates considerable variation of dynamic effects. We note, for example, high or very low effects around certain weeks, regular purchase patterns every three and four weeks, purchase in each of several weeks, and so on. Hidden variables reflect that interpurchase

Table 9
Average Category Dependences.

Purchase probability increase for	Associated purchase probability changes							
	for the MLP:							
toitisu	coffee	0.0141	factiss	0.0100	laundet	0.0149	margbutr	0.0117
	paptowl	0.0145	yogurt	-0.0243				
toothbr	coffee	0.0121	laundet	0.0134	paptowl	0.0137	yogurt	-0.0233
toothpa	coffee	0.0120	laundet	0.0130	paptowl	0.0131	yogurt	-0.0230
yogurt	coffee	0.0104	laundet	0.0109	paptowl	0.0108		
	for the LSTM net:							
beer	coffee	-0.0243						
deod	blades	-0.0106	coffee	0.0302	photo	-0.0128	spagsauc	0.0136
	yogurt	-0.0122						
fzdin	coffee	0.0337						
fzpizza	coffee	0.0346						
margbutr	coffee	0.0362						
mayo	coffee	0.0336	soup	-0.0100				
mustketc	coffee	0.0346	hotdog	0.0145	soup	-0.0110	yogurt	-0.0124
paptowl	coffee	0.0327	hotdog	-0.0149				
saltsnck	carbbev	-0.0120	coffee	0.0462	fzdin	-0.0102	hotdog	-0.0393
	milk	0.0320						
soup	carbbev	-0.0123	coffee	0.0381	hotdog	-0.0419	milk	0.0318
spagsauc	carbbev	-0.0125	coffee	0.0313	hotdog	-0.0367	milk	0.0317
sugarsub	carbbev	-0.0126	coffee	0.0376	hotdog	-0.0357	milk	0.0316
yogurt	carbbev	-0.0133	coffee	0.0416	hotdog	-0.0244	milk	0.0315
	saltsnck	-0.0100						

shows dependences of at least 0.01 in absolute size

Table 10
Average Own Effects of Features.

for the MLP							
soup	0.0264	spagsauc	0.0239	fzdin	0.0206	margbutr	0.0204
hotdog	0.0145	peanbutr	0.0125	fzpizza	0.0098	mustketc	0.0093
mayo	0.0086	yogurt	0.0054				
for the LSTM net							
coffee	0.0469	soup	0.0291	beer	0.0131	hhclean	0.013
toothpa	0.010	laundet	0.085	deod	0.053		

shows average own effects of at least 0.005 in absolute size

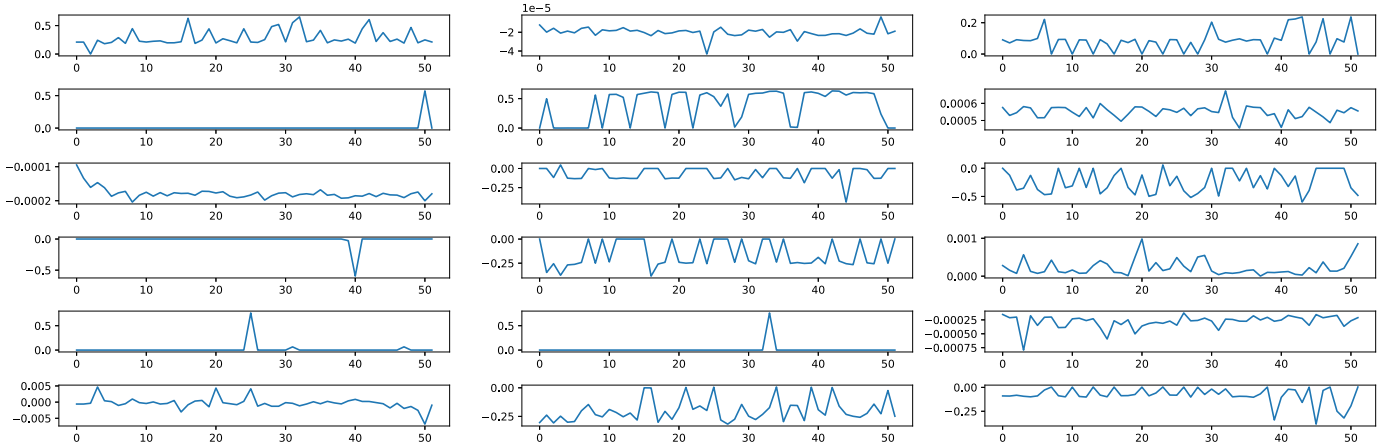


Fig. 1. Six Weekly Hidden Variables of the LSTM net for Three Selected Households.

times differ to a great extent between product categories. To determine purchase probabilities of any category j the hidden variables are weighted differently by a vector W_j as shown in expression (1).

7.4. Optimization results

We optimize 1612 features, i.e., 31 category-specific features for 52 weeks, to obtain insights into managerial relevant differences between

Table 11
Average Cross Effects of Features for the MLP (1).

cigets	fzdin	0.0181	fzpizza	0.0073	hotdog	0.0120	margbutr	0.0167
	mayo	0.0055	mustketc	0.0061	peanbutr	0.0110	soup	0.0214
	spagsauc	0.0214	yogurt	0.0052				
coffee	fzdin	0.0197	fzpizza	0.0083	hotdog	0.0140	margbutr	0.0201
	mayo	0.0080	mustketc	0.0077	peanbutr	0.0120	soup	0.0251
	spagsauc	0.0229						
coldcer	fzdin	0.0218	fzpizza	0.0100	hotdog	0.0156	margbutr	0.0219
	mayo	0.0093	mustketc	0.0090	peanbutr	0.0131	soup	0.0279
	spagsauc	0.0264						
deod	fzdin	0.0203	fzpizza	0.0088	hotdog	0.0146	margbutr	0.0211
	mayo	0.0086	mustketc	0.0082	peanbutr	0.0124	soup	0.0263
	spagsauc	0.0241						
diapers	fzdin	0.0206	fzpizza	0.0091	hotdog	0.0148	margbutr	0.0212
	mayo	0.0085	mustketc	0.0084	peanbutr	0.0127	soup	0.0266
	spagsauc	0.0250						
factiss	fzdin	0.0205	fzpizza	0.0090	hotdog	0.0146	margbutr	0.0207
	mayo	0.0085	mustketc	0.0082	peanbutr	0.0124	soup	0.0262
	spagsauc	0.0243						
fzdin	fzpizza	0.0090	hotdog	0.0146	margbutr	0.0207	mayo	0.0084
	mustketc	0.0082	peanbutr	0.0124	soup	0.0261	spagsauc	0.0244
fzpizza	fzdin	0.0216	hotdog	0.0151	margbutr	0.0208	mayo	0.0091
	mustketc	0.0086	peanbutr	0.0127	soup	0.0267	spagsauc	0.0253
hhclean	fzdin	0.0190	fzpizza	0.0078	hotdog	0.0139	margbutr	0.0207
	mayo	0.0078	mustketc	0.0077	peanbutr	0.0120	soup	0.0253
	spagsauc	0.0227						
hotdog	fzdin	0.0201	fzpizza	0.0087	margbutr	0.0209	mayo	0.0084
	mustketc	0.0081	peanbutr	0.0123	soup	0.0260	spagsauc	0.0238
laundet	fzdin	0.0198	fzpizza	0.0085	hotdog	0.0143	margbutr	0.0207
	mayo	0.0081	mustketc	0.0080	peanbutr	0.0123	soup	0.0258
	spagsauc	0.0238						
margbutr	fzdin	0.0199	fzpizza	0.0085	hotdog	0.0142	mayo	0.0081
	mustketc	0.0079	peanbutr	0.0122	soup	0.0255	spagsauc	0.0235
mayo	fzdin	0.0205	fzpizza	0.0090	hotdog	0.0148	margbutr	0.0213
	mustketc	0.0084	peanbutr	0.0126	soup	0.0266	spagsauc	0.0246
milk	fzdin	0.0200	fzpizza	0.0086	hotdog	0.0142	margbutr	0.0203
	mayo	0.0080	mustketc	0.0079	peanbutr	0.0122	soup	0.0255
	spagsauc	0.0238						
mustketc	fzdin	0.0215	fzpizza	0.0098	hotdog	0.0159	margbutr	0.0232
	mayo	0.0091	peanbutr	0.0137	soup	0.0291	spagsauc	0.0278
	yogurt	0.0056						

shows average marginal cross effects of at least 0.005 in absolute size

the MLP and the LSTM net. Less than 100 fixed-point iterations are sufficient to obtain joint purchase probabilities. In accordance with our observations, we restrict features to the interval $[0, 0.2]$, starting with a standard deviation of 0.05 and let the optimization algorithm run for 300 iterations.

Both for the MLP and the LSTM net we determine average revenue per basket for no features, all features set to 0.1, all features set to 0.2, and optimal features. Two aspects suggest preferring the results obtained for the LSTM net. One aspect relates to its better statistical performance. The other aspect concerns the fact that average revenues per basket computed based on the LSTM net given in Table 14 are more in line with the observed average basket revenue.

Optimization based on the LSTM net increases expected average revenue per basket by about 35% compared to its observed value. Optimal features differ between the MLP and the LSTM net in 30 categories. In more detail, optimal features are lower in 16 categories and higher in 14 categories for the LSTM net according to t-tests, each at a significance level of 0.05. For the MLP, optimal solutions are limited to the bounds of the decision variables. The optimization recommends no features in 14 categories and features at the upper bound of 0.2 in 17 categories. Optimal feature values of each category (which equal either 0.0 or 0.2) are constant across weeks for the MLP. Optimal solutions are more heterogeneous for the LSTM net with many intermediate values. We obtain 0.01, 0.07 and 0.16 as quartiles of optimal features based on the LSTM net. This higher heterogeneity can be explained by the fact that in contrast to the MLP the LSTM net also implies negative cross effects as shown in Section 7.3.

We illustrate managerial implications of the LSTM net for six selected categories. We determine these categories as centers of a six-cluster medoid cluster analysis using optimal weekly features to make sure that categories differ as much as possible from each other. Because in contrast to the MLP the LSTM net includes dynamic effects, optimal features for these categories vary considerably across weeks (see Fig. 2). These results mean that managers should set promotion values that vary across time.

Frequently, managers should set a low value after a high value and vice versa. On the other hand, managers should decide for low feature values in certain successive weeks (blades: May – June, household cleaners, beer and peanut butter: September – mid-October, paper towels: mid-January – February, mid-May – mid-July, mid-August – September, diapers: mid-April – June, September – November). For five of the six categories, management should choose high feature values just before the start of both the Easter and Christmas holidays. Because optimization based on the LSTM net is more in line with average observed revenues and implies features values that vary across time, we think that feature decisions should be based on this model rather than on the MLP.

8. Conclusion

We compare recurrent neural nets for market basket analysis to that of several, less complex models. In the following, we deal with the reasons for performance differences. Among the static models, the MVL model is clearly outperformed by the MLP, because the hidden variables of the latter reproduce category interdependences and sales promotion effects in a more flexible way. The MVL model on the other hand allows

Table 12
Average Cross Effects of Features for the MLP (2).

paptowl	fzdin	0.0203	fzpizza	0.0088	hotdog	0.0146	margbutr	0.0211
	mayo	0.0085	mustketc	0.0083	peanbutr	0.0125	soup	0.0264
	spagsauc	0.0243	yogurt	0.0052				
peanbutr	fzdin	0.0203	fzpizza	0.0088	hotdog	0.0147	margbutr	0.0212
	mayo	0.0086	mustketc	0.0083	soup	0.0264	spagsauc	0.0242
	yogurt	0.0051						
photo	fzdin	0.0194	fzpizza	0.0082	hotdog	0.0145	margbutr	0.0217
	mayo	0.0082	mustketc	0.0082	peanbutr	0.0125	soup	0.0265
	spagsauc	0.0238						
razors	fzdin	0.0206	fzpizza	0.0091	hotdog	0.0147	margbutr	0.0208
	mayo	0.0085	mustketc	0.0083	peanbutr	0.0125	soup	0.0262
	spagsauc	0.0246	yogurt	0.0053				
saltsnck	fzdin	0.0203	fzpizza	0.0089	hotdog	0.0146	margbutr	0.0209
	mayo	0.0084	mustketc	0.0082	peanbutr	0.0125	soup	0.0262
	spagsauc	0.0243	yogurt	0.0052				
shamp	fzdin	0.0207	fzpizza	0.0091	hotdog	0.0148	margbutr	0.0210
	mayo	0.0087	mustketc	0.0083	peanbutr	0.0125	soup	0.0264
	spagsauc	0.0243	yogurt	0.0052				
soup	fzdin	0.0208	fzpizza	0.0092	hotdog	0.0147	margbutr	0.0208
	mayo	0.0085	mustketc	0.0083	peanbutr	0.0126	spagsauc	0.0249
	yogurt	0.0053						
spagsauc	fzdin	0.0200	fzpizza	0.0086	hotdog	0.0145	margbutr	0.0211
	mayo	0.0084	mustketc	0.0082	peanbutr	0.0124	soup	0.0262
	yogurt	0.0051						
sugarsub	fzdin	0.0223	fzpizza	0.0103	hotdog	0.0156	margbutr	0.0212
	mayo	0.0097	mustketc	0.0089	peanbutr	0.0128	soup	0.0274
	spagsauc	0.0256	yogurt	0.0055				
toitisu	fzdin	0.0201	fzpizza	0.0087	hotdog	0.0143	margbutr	0.0206
	mayo	0.0082	mustketc	0.0080	peanbutr	0.0123	soup	0.0258
	spagsauc	0.0238	yogurt	0.0053				
toothbr	fzdin	0.0217	fzpizza	0.0100	hotdog	0.0152	margbutr	0.0209
	mayo	0.0090	mustketc	0.0086	peanbutr	0.0128	soup	0.0269
	spagsauc	0.0258	yogurt	0.0056				
toothpa	fzdin	0.0187	fzpizza	0.0075	hotdog	0.0140	margbutr	0.0214
	mayo	0.0082	mustketc	0.0078	peanbutr	0.0120	soup	0.0257
	spagsauc	0.0220						
yogurt	fzdin	0.0209	fzpizza	0.0093	hotdog	0.0150	margbutr	0.0214
	mayo	0.0089	mustketc	0.0085	peanbutr	0.0127	soup	0.0270
	spagsauc	0.0249						

shows average marginal cross effects of at least 0.005 in absolute size

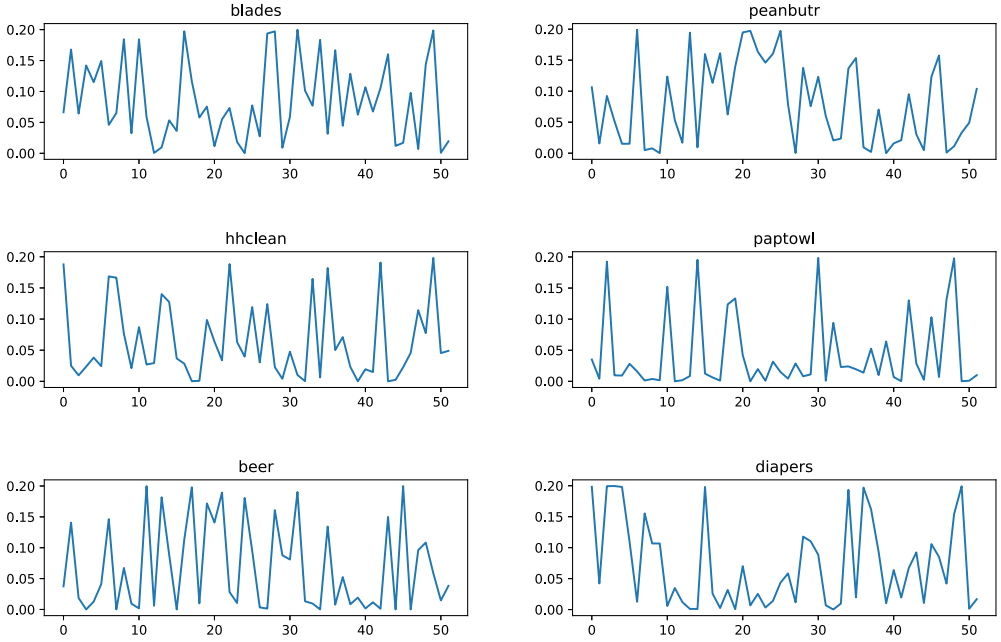


Fig. 2. Optimal Weekly Features of Six Selected Categories for the LSTM net.

only two-way interactions between categories and log-linear effects of sales promotion variables.

A simple way to add dynamics to these models consists in including category loyalties, exponentially smoothed past purchases, as predic-

Table 13
Average Cross Effects of Features for the LSTM net.

blades	coffee	0.0383						
carbbev	coffee	0.0389						
cigets	coffee	-0.0059	soup	-0.0051				
coffee	soup	-0.0055						
coldcer	coffee	0.0381	soup	-0.0052				
deod	coffee	0.0391						
diapers	coffee	0.0383	soup	-0.0066				
factiss	coffee	0.0402	soup	-0.0066				
fzdin	coffee	0.0413	soup	-0.0066				
fzpizza	coffee	0.0504	soup	0.0053				
hhclean	coffee	0.0462	fzdin	-0.0051	photo	0.0052		
hotdog	coffee	0.0442						
laundet	coffee	0.0460	saltsnck	-0.0051				
margbutr	coffee	0.0396	fzdin	-0.0064	saltsnck	-0.0052		
mayo	coffee	0.0439	saltsnck	-0.0056				
milk	coffee	0.0445	saltsnck	-0.0051				
mustketc	blades	0.0052	carbbev	-0.0050	coffee	0.0538	fzdin	-0.0050
	photo	0.0108	yogurt	-0.0052				
paptowl	coffee	0.0435	saltsnck	-0.0058	yogurt	-0.0059		
peanbutr	carbbev	-0.0050	coffee	0.0415	hotdog	-0.0057	saltsnck	-0.0057
	yogurt	-0.0065						
photo	coffee	0.0460	saltsnck	-0.0064				
razors	coffee	0.0453	saltsnck	-0.0060	yogurt	-0.0060		
saltsnck	coffee	0.0419	yogurt	-0.0055				
shamp	coffee	0.0416	saltsnck	-0.0057	yogurt	-0.0060		
soup	coffee	0.0421	saltsnck	-0.0060	yogurt	-0.0061		
spagsauc	coffee	0.0431	saltsnck	-0.0062	yogurt	-0.0064		
sugarsub	carbbev	-0.0054	coffee	0.0533	peanbutr	-0.0057	photo	0.0107
	saltsnck	-0.0088						
toitisu	carbbev	-0.0055	coffee	0.0413	peanbutr	-0.0056	saltsnck	-0.0079
	yogurt	-0.0064						
toothbr	carbbev	-0.0058	coffee	0.0444	fzdin	-0.0052	hotdog	-0.0054
	peanbutr	-0.0056	saltsnck	-0.0073	yogurt	-0.0058		
toothpa	carbbev	-0.0061	coffee	0.0445	hotdog	-0.0068	peanbutr	-0.0057
	saltsnck	-0.0088	toitisu	-0.0063	yogurt	-0.0076		
yogurt	carbbev	-0.0061	coffee	0.0417	peanbutr	-0.0057	saltsnck	-0.0087
	toitisu	-0.0056						

shows average marginal cross effects of at least 0.005 in absolute size

Table 14
Average Revenue per Basket.

Model	no features	features all set to 0.1	features all set to 0.2	optimal features
LSTM	4.9990	4.9049	4.8703	5.7864
MLP	0.8246	0.7716	0.7319	1.7111

observed average revenue: 4.2928

tors. For the MVL model category loyalties improve performance in contrast to the MLP. Obviously, the MLP with its flexibility with respect to category interdependences and sales promotion effects does not benefit from the simple dynamics generated this way.

The hidden variables of the MLP depend on current input variables (category purchase and sales promotion variables) only. Recurrent nets generalize the MLP, because hidden variables of the former also depend on hidden variables of the previous period. Hidden variables of recurrent nets allow forms of purchase event feedback (effects of past purchases on current purchases), which are more complex than exponentially smoothed past purchases. Moreover, recurrent nets may also reproduce indirect effects of sales promotion variables. Indirect effects arise by sales promotion variables directly influencing hidden variables that in their turn affect future hidden variables. Then these hidden variables affect future purchases.

We apply all three main variants of recurrent neural nets. The SRNN does not outperform the MLP, which can be explained by the vanishing gradient problem of the SRNN that interferes with estimation of long term effects. The LSTM net was developed to avoid the vanishing gradient problem. The GRU net constitutes a simplification of the LSTM net. Both the LSTM net and GRU net outperform the MLP due to their ability to reproduce purchase event feedback and indirect sales promotion effects.

The higher complexity of the LSTM net compared to the GRU net pays off. The overall best performing model is a LSTM net that includes category-specific features (advertising in local newspapers or flyers) as one of the input variables.

In the following we only discuss two of the investigated models, the best performing non-recurrent model, a MLP, and a LSTM net, both with six hidden variables. The LSTM net leads to better predictions for 26 of 31 categories, only for two categories its predictions are worse than those of the MLP. Taking dynamic effects into account for these two categories, whose purchases are more regular than those of the remaining categories, obviously does not improve predictions. The LSTM net excels in predicting basket composition. The LSTM net clearly does a better job in finding out which categories are relevant for households.

We derive dependences between product categories for these two models as average probability changes of each category due to the marginal increase of the probability of another category. These dependences may be either substitutive or complementary. We obtain more dependences (also more substitutive dependences) for the LSTM net. Managers run the risk to overestimate the relative number of complementary dependences of they rely on the MLP despite its worse statistical performance.

We therefore focus on the LSTM net and illustrate what stronger dependences imply for fostering cross-selling by positioning categories in aisles and shelves of a store or by recommending other, not purchased categories.

In addition, we also look at average effects of features on marginal purchase probabilities. These effects are own effects if they involve the same category, cross effects if they affect purchases of other categories.

Both models imply several higher own effects, which are all positive, i.e., more features increase the purchase probability of the same category.

We obtain more high cross effects for the MLP than for the LSTM net. These high cross effects are all positive. In contrast, the LSTM also implies many higher (in absolute size) negative cross effects.

We investigate what these two models imply for sales promotion. To this end we optimize features for each category in each of 52 weeks by means of the covariance matrix adaptation evolution strategy (CMA-ES). Average revenue per basket, which is proportional to total revenue, serves as objective to be maximized. Optimization based on the LSTM net increases expected average revenue per basket by about 35% compared to its observed value. For the MLP optimal solutions are limited to the bounds of the decision variables. Optimal solutions are more varied for the LSTM net with many intermediate values and vary considerably across weeks confirming our expectation that optimal weekly sales promotion variables are not constant for dynamic models.

We illustrate managerial implications of the LSTM net for six selected categories that differ as much as possible from each other. Because in contrast to MLP the LSTM net includes dynamic effects, optimal features for these categories vary considerably across weeks. This result means that managers should set promotion values that vary across time. Frequently, managers should set a low value after a high value and vice versa. On the other hand, manager should decide for low feature values in certain successive weeks. In five of these six categories, management should choose high feature values just before the start of both the Easter and Christmas holidays. Because optimization based on the LSTM net is more in line with average observed revenues and implies features values that vary across time, we think that feature decisions should be based on this model rather than on the MLP.

The approach presented here is, of course, not free from limitations, all of which provide opportunities for future research. The data we analyze originate from a food retailing context. Investigating the relevance of recurrent nets for non-food product categories (e.g., consumer electronics, apparel, footwear, books, furniture, financial services, tourism) seems to be interesting. In this paper we look at (binary) purchases of different product categories. One might take other output variables (e.g., brand choices, purchase quantities) into account.

Future research may deal with completely different application fields. As proposed by Sarkar and De Bruyn (2021), recurrent neural networks might be appropriate to investigate online customer behavior across multiple websites as they capture inter-sequence and inter-temporal interactions from multiple streams of clickstream. Recurrent nets constitute alternatives to widespread static approaches, such as topic models (Schröder et al., 2019; Falke and Hruschka, 2022). Similarly, consumption of different media across time could be analyzed by recurrent nets.

In addition, technical extensions may be possible as well. Our neural nets include only one hidden layer. Therefore, it might be interesting to investigate whether the addition of hidden layers improves statistical performance. One also could compare to alternative optimization methods such as other evolutionary algorithms (Eiben and Smith, 2015), e.g., genetic algorithms, differential evolution, particle swarm optimization, or randomized heuristics, e.g., simulated annealing (Kirkpatrick et al., 1983; Cerny, 1985).

Data statement

The dataset generated during the current study is not publicly available as it contains proprietary information that the authors acquired through a license.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

I have nothing to declare.

Data availability

The data that has been used is confidential.

References

Aurier, P., Mejia, V., 2014. Multivariate logit and probit models for simultaneous purchases: presentation, uses, appeal and limitations. Rech. Appl. Mark. 29, 79–98.
Bai, T., Nie, J.Y., Zhao, W.X., Zhu, Y., Du, P., Wen, J.R., 2018. An attribute-aware neural attentive model for next basket recommendation. In: SIGIR'18. ACM.
Baumol, W.J., 1967. Business Behavior, Value and Growth. Harcourt Brace Jovanovich, New York.
Bee, M., Espa, G., Giuliani, D., 2015. Approximate maximum likelihood estimation of the autologistic model. Comput. Stat. Data Anal. 84, 14–26.
Bel, K., Fok, D., Paap, R., 2018. Parameter estimation in multivariate logit models with many binary choices. Econom. Rev. 37, 534–550.
Bengio, Y., Simard, P., Frasconi, P., 1994. Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw. 5, 157–166.
Besag, J., 1972. Nearest-neighbour systems and the auto-logistic model for binary data. J. R. Stat. Soc., Ser. B 34, 75–83.
Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. J. R. Stat. Soc., Ser. B 35, 192–236.
Betancourt, R., Gautschi, D., 1990. Demand complementarities, household production, and retail assortments. Mark. Sci. 9, 146–161.
Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, Oxford, UK.
Boztuğ, Y., Hildebrandt, L., 2008. Modeling joint purchases with a multivariate MNL approach. Schmalenbach Bus. Rev. 60, 400–422.
Boztuğ, Y., Reutterer, T., 2008. A combined approach for segment-specific market basket analysis. Eur. J. Oper. Res. 187, 294–312.
Bronnenberg, B.J., Kruger, M.W., Mela, C.F., 2008. Database paper: the IRI marketing data set. Mark. Sci. 27, 745–748.
Cerny, V., 1985. A thermodynamical efficient simulation algorithm. J. Optim. Theory Appl. 45, 41–51.
Chib, S., Seetharaman, P.B., Strijnev, A., 2002. Analysis of multi-category purchase incidence decisions using IRI market basket data. In: Franses, P.H., Montgomery, A.L. (Eds.), Econometric Models in Marketing. JAI, Amsterdam, pp. 57–92.
Cho, K., Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Doha, Qatar, pp. 1724–1734.
Chollet, F., et al., 2015. Keras. https://github.com/fchollet/keras.
Cox, D.R., 1972. The analysis of multivariate binary data. J. R. Stat. Soc., Ser. C 21, 113–120.
Dhillon, P.S., Aral, S., 2021. Modeling dynamic user interests: a neural matrix factorization approach. Mark. Sci. 40, 1059–1080.
Dippold, K., Hruschka, H., 2013a. A model of heterogeneous multicategory choice for market basket analysis. Rev. Mark. Sci. 11, 1–31.
Dippold, K., Hruschka, H., 2013b. Variable selection for market basket analysis. Comput. Stat. 28, 519–539.
Dunn, R., Reader, S., Wrigley, N., 1983. An investigation of the assumptions of the NBD model as applied to purchasing at individual stores. J. R. Stat. Soc., Ser. C, Appl. Stat. 32, 249–259.
Duvvuri, S.D., Ansari, V., Gupta, S., 2007. Consumers' price sensitivities across complementary categories. Manag. Sci. 53, 1933–1945.
Eiben, A.E., Smith, J.E., 2015. Introduction to Evolutionary Computing, 2nd ed. Springer, Berlin.
Elman, J.L., 1990. Finding structure in time. Cogn. Sci. 14, 179–211.
Erdem, T., Keane, M.P., 1996. Decision-making under uncertainty: capturing dynamic brand choice processes in turbulent consumer goods markets. Mark. Sci. 15, 1–20.
Falke, A., Hruschka, H., 2022. Analyzing browsing across websites by machine learning methods. J. Bus. Econ. 92, 829–852.
Gabel, S., Ringel, D., 2024. The market basket transformer: a new foundation model for retail. https://doi.org/10.2139/ssrn.4335141.
Gabel, S., Timoshenko, A., 2022. Product choice with large assortments: a scalable deep-learning model. Manag. Sci. 68, 1808–1827.
Gedenk, K., Neslin, S.A., 1999. The role of retail promotion in determining future brand loyalty: its effect on future purchase event feedback. J. Retail. 75, 433–459.
Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press, Cambridge, MA.
Guadagni, P.M., Little, J.D.C., 1983. A logit model of brand choice calibrated on scanner data. Mark. Sci. 2, 203–238.
Gupta, S., 1988. Impact of sales promotions on when, what, and how much to buy. J. Mark. Res. 25, 342–355.

Hansen, N., 2018. Module cma. http://www.cmap.polytechnique.fr/~nikolaus.hansen/html-pythoncma/frames.html.
Hansen, N., 2023. The cma evolution strategy: a tutorial. https://doi.org/10.48550/arXiv.1604.00772.
Hansen, N., Ostermeier, A., 2001. Completely derandomized self-adaptation in evolution strategies. Evol. Comput. 9, 159–195.
Haykin, S., 1994. Neural Networks: A Comprehensive Foundation. Macmillan Publishing.
Heckman, J.J., 1981. Statistical models for discrete panel data. In: Manski, C.F., McFadden, D. (Eds.), Advanced Methods for Modeling Markets. MIT Press, Cambridge, MA, pp. 114–178.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780.
Hong, J., Hoban, P.R., 2022. Writing more compelling creative appeals: a deep learning-based approach. Mark. Sci. 41, 941–965.
Hruschka, H., 2008. Neural nets and genetic algorithms in marketing. In: Wierenga, B. (Ed.), Handbook of Marketing Decision Models. Springer, New York, pp. 114–178.
Hruschka, H., 2014. Linking multi-category purchases to latent activities of shoppers: analysing market baskets by topic models. Mark. ZFP 36, 268–274.
Hruschka, H., 2017. Analyzing the dependences of multicategory purchases on interactions of marketing variables. J. Bus. Econ. 87, 295–313.
Hruschka, H., 2021. Comparing unsupervised probabilistic machine learning methods for market basket analysis. Rev. Manag. Sci. https://doi.org/10.1007/s00291-022-00690-z.
Hruschka, H., 2022. Relevance of dynamic variables in multicategory choice models. OR Spektrum. https://doi.org/10.1007/s00291-022-00690-z.
Hruschka, H., Lukanowicz, M., Buchta, C., 1999. Cross-category sales promotion effects. J. Retail. Consum. Serv. 6, 99–105.
Kingma, D.P., Ba, J., 2015. Adam: A Method for Stochastic Optimization. Cornell University.
Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., 1983. Optimization by simulated annealing. Science 220, 671–680.
Le, D.T., Lauw, H.W., Fang, Y., 2019. Correlation-sensitive next-basket recommendation. In: IJCA-2019, pp. 2808–2814.
van Maasakkers, L., Fok, D., Donkers, B., 2023. Next-basket prediction in a high-dimensional setting using gated recurrent units. Expert Syst. Appl. 212.
Manchanda, P., Ansari, A., Gupta, S., 1999. The shopping basket: a model for multicategory purchase incidence decisions. Mark. Sci. 18, 95–114.
McLachlan, G., Basford, K., 1988. Mixture Models: Inference and Applications to Clustering. Marcel Dekker, New York.
Mena, G., Coussement, K., De Bock, K.W., De Caigny, A., Lessmann, S., 2023. Exploiting time-varying RFM measures for customer churn prediction with deep neural networks. Ann. Oper. Res. https://doi.org/10.1007/s10479-023-05259-9.
Ngatchou-Wandji, J., Bulla, J., 2013. On choosing a mixture model for clustering. J. Data Sci. 11, 157–179.
Peterson, C., Anderson, 1987. A mean field theory learning algorithm for neural networks. Complex Syst. 1, 995–1019.
Richards, T.J., Hamilton, S.F., Yonezkawa, K., 2018. Retail market power in a shopping basket model of supermarket competition. J. Retail. 94, 328–342.
Russell, G.J., Petersen, A., 2000. Analysis of cross category dependence in market basket selection. J. Retail. 76, 69–392.
Salehinejad, H., Rahnamayan, S., 2016. Customer shopping pattern prediction: a recurrent neural network approach. In: 2016 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1–6.
Sarkar, M., De Bruyn, A., 2021. LSTM response models for direct marketing analytics: replacing feature engineering with deep learning. J. Interact. Mark. 53, 80–95.
Schröder, N., Falke, A., Hruschka, H., Reutterer, T., 2019. Analyzing the browsing basket: a latent interests-based segmentation tool. J. Interact. Mark. 35, 181–197.
Schröder, N., Hruschka, H., 2016. Investigating the effects of mailing variables and endogeneity on mailing decisions. Eur. J. Oper. Res. 250, 579–589.
Sheil, H., Rana, O., Reilly, R., 2018. Predicting purchasing intent: automatic feature learning using recurrent neural networks. In: Proceedings of the SIGIR 2018 eCom. Ann Arbor, Michigan.
Solnet, D., Boztuğ, Y., Dolnicar, S., 2016. An untapped gold mine? Exploring the potential of market basket analysis to grow hotel revenue. Int. J. Hosp. Manag. 56, 119–125.
Toth, A., Tan, L., Di Fabbrizio, G., Datta, A., 2017. Predicting shopping behavior with mixture of RNNs. In: Proceedings of the SIGIR 2017 eCom. Tokyo.
Valendin, J., Reutterer, T., Platzer, M., Kalcher, K., 2022. Customer base analysis with recurrent neural networks. Int. J. Res. Mark. 39, 988–1018.
van Diepen, M., Donkers, B., Franses, P.H., 2009. Dynamic and competitive effects of direct mailings: a charitable giving application. J. Mark. Res. 46, 120–133.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst., 5998–6008.
Virtanen, P., et al., 2020. Scipy 1.0: fundamental algorithms for scientific computing. Nat. Methods 17, 261–272.
Yu, F., Liu, Q., Wu, S., Wang, L., Tan, T., 2016. A dynamic recurrent model for next basket recommendation. In: SIGIR'16. ACM.