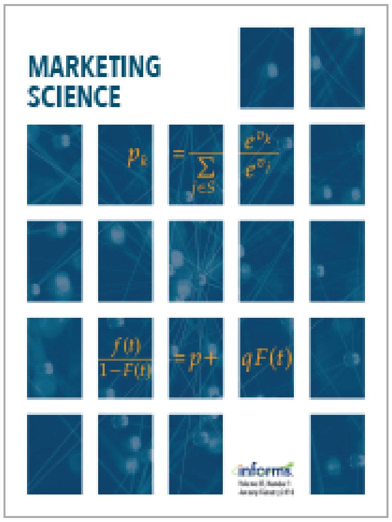


Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Using Conditional Restricted Boltzmann Machines to Model Complex Consumer Shopping Patterns

Feihong Xia, Rabikar Chatterjee, Jerrold H. May

To cite this article:
Feihong Xia, Rabikar Chatterjee, Jerrold H. May (2019) Using Conditional Restricted Boltzmann Machines to Model Complex Consumer Shopping Patterns. Marketing Science 38(4):711-727. https://doi.org/10.1287/mksc.2019.1162

CORRECTED VERSION OF RECORD; SEE LAST PAGE OF ARTICLE

http://pubsonline.informs.org/journal/mksc/

MARKETING SCIENCE
Vol. 38, No. 4, July–August 2019, pp. 711–727
ISSN 0732-2399 (print), ISSN 1526-548X (online)

Using Conditional Restricted Boltzmann Machines to Model Complex Consumer Shopping Patterns

Feihong Xia,[a] Rabikar Chatterjee,[b] Jerrold H. May[b]

[a] University of Rhode Island College of Business Administration, Kingston, Rhode Island 02881; [b] Katz Graduate School of Business, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

Contact: fex8@pitt.edu (FX); rabikar@katz.pitt.edu, http://orcid.org/0000-0003-1131-0728 (RC); jerrymay@katz.pitt.edu (JHM)

Received: August 19, 2015
Revised: February 27, 2017; June 22, 2018; February 5, 2019
Accepted: February 11, 2019
Published Online in Articles in Advance: July 12, 2019

https://doi.org/10.1287/mksc.2019.1162

Abstract. Marketers have recognized that the probability of a consumer's (or household's) purchase in a particular product category may be influenced by past purchases in the same category and also, purchases in other related categories. Past studies of crosscategory effects have focused on a limited number of product categories, and they have often ignored intertemporal effects in their analyses. Those studies have generally used multivariate logit or probit models, which are limited in their ability to analyze enormous data sets that contain consumer purchase records across a large number of categories and time periods. The availability of such enormous consumer shopping data sets and the value of analyzing the complex relationships across categories and over time (for example, for personalized promotions) suggest the need for computationally efficient modeling and estimation methods. Such models can capture associations among buying decisions across all product categories and over all time periods for which data are available, but they must also have a tractable and clear model structure that permits meaningful interpretation of the results. We explore the nature of intertemporal crossproduct patterns in an enormous consumer purchase data set using a model that adopts the structure of conditional restricted Boltzmann machines (CRBMs). Our empirical results demonstrate that our proposed approach using the efficient estimation algorithm embodied in the CRBM enables us to process very large data sets and capture the consumer decision patterns for both predictive and descriptive purposes that might not otherwise be apparent. In addition to persistent intertemporal within-category effects, we find that there are also significant intertemporal cross effects between product categories.

History: K. Sudhir served as the editor-in-chief and Puneet Manchanda served as associate editor for this article.

Supplemental Material: Data and the online appendix are available at https://doi.org/10.1287/mksc.2019.1162.

Keywords: big data • sparsity • regularization • consumer decision patterns • deep learning

1. Introduction

Big data and mobile marketing have transformed the landscape of marketing, with companies using consumer data to launch effective personalized coupons and highly targeted advertisements. In this brave new world, marketing managers must "get the timing right" and "rely on data to make offers relevant" (Miles 2013). Machine learning, especially deep learning, has fueled much of the rapid progress in consumer product recommendation and purchase prediction. Deep networks have achieved remarkable results in image and speech recognition, bringing huge benefits to businesses and consumers alike. Google, Microsoft, Apple, Facebook, Netflix, Amazon, and others have created their own deep learning engines to generate recommendations based on analysis of consumer data. Some companies have released their open source deep learning framework (e.g., Google with TensorFlow). However, most deep networks used in business applications, such as convolutional neural networks, deep reinforcement learning, and recurrent neural networks, are nonprobabilistic, with model structures that provide no basis for theoretical interpretation of the results.

For personalized marketing, an ideal model should capture and predict shopping patterns efficiently and accurately from large consumer data sets to both assist in forecasting consumers' future purchase decisions and permit inferences with marketing decision-making implications. We propose an efficient probabilistic model that captures intertemporal multicategory purchasing patterns from such data to address these needs. Our ability to model intertemporal crossproduct effects on a massive scale has been influenced by two factors. First, we now have access to massive amounts of consumer-level (or household-level) purchase data recorded at the point of sale by individual retailers and compiled by marketing research companies, such as Nielsen. Second, efficient methods have been developed that permit the analysis of such rich and massive data sets.

Traditional modeling approaches in marketing are econometrics based, emphasizing unbiased parameter estimation rather than algorithm efficiency (Mehta 2007). However, machine learning models (e.g., neural networks) can handle large data sets, but the results are typically difficult to interpret. To model large-scale, intertemporal, crosscategory effects using the massive data sets available today, we need a parsimonious model with parameters that must be estimated efficiently but a tractable and clear model structure that permits meaningful interpretation of the results.

Previous research on multicategory consumer purchasing decisions has focused primarily on a few product categories owing to model estimation limitations. Furthermore, discussion of the intertemporal aspects of multicategory buying decision is meagre (Manchanda et al. 1999, Edwards and Allenby 2003, Seetharaman et al. 2005). One major challenge for such analysis is the rapid growth in the size of the correlation matrix as the numbers of product categories and time periods increase.[1]

To make the analysis less time consuming, the typical approach is to preselect a few categories, which are expected be either complements or substitutes ex ante, and then, use multivariate probit (MVP) or multivariate logit methods to model the cross effects (Manchanda et al. 1999, Russell and Petersen 2000, Duvvuri et al. 2007, Boztug and Hildebrandt 2008).

A more recent stream of literature in marketing focuses on empirical methods for handling high-dimensional data. One intuitive approach is to first reduce the dimensionality to a manageable level and then, conduct the analysis on the lower-dimensional data set (Lawrence 2004). Latent linear models, such as factor analysis and principal component analysis, are the most popular dimension reduction methods in marketing research. However, the benefits of applying such methods in multicategory analysis are not apparent. Duvvuri and Gruca (2010) develop a Bayesian multilevel factor model for consumer price sensitivity analysis across product categories, but this model yields less accurate predictions than the regular MVP model. The algorithm of Talhouk et al. (2012) for efficient and sparse MVP modeling improves efficiency and reduces the number of parameters, but the estimation relies on Markov chain Monte Carlo (MCMC) sampling, which is inefficient when analyzing very large data sets.[2]

A model that utilizes a parsimonious (or technically, sparse) representation of the associations between product purchasing decisions and an efficient inference algorithm is needed for high-dimensional multicategory purchasing decision analysis. The restricted Boltzmann machine (RBM) possesses such advantages (Hruschka 2014).

We significantly extend the RBM to develop an efficient interpretable model for the analysis of consumer purchasing patterns on a very large scale that can capture (i) *intertemporal* in addition to contemporaneous multicategory cross effects, (ii) the impact of *additional variables* (e.g., marketing mix) that may influence purchasing patterns, and (iii) consumer *heterogeneity*. To the best of our knowledge, this is the first time that a *conditional* restricted Boltzmann machine (CRBM) has been adapted for application to a marketing problem. We describe in detail how the model parameters are estimated, including our extension of the CRBM in the machine learning literature to handle the sparseness that is typical of marketing data (as in our illustrative application), and then, compare its performance with that of currently popular models to demonstrate its strengths, especially in analyzing very large data sets.

The empirical illustration analyzes household-level supermarket shopping histories based on panel data on purchases made by each of 4,000 households across 1,055 product categories over 208 weeks, totaling nearly 6.7 million observations. We focus on describing the methodology and illustrating its application, and we only briefly touch on the substantive implications of the results. Thus, the contribution of this paper is primarily methodological, and a more comprehensive application (or multiple applications) of the model for improved marketing decision making (for example, in personalizing mobile couponing) is left to future research.

Next, Section 2 develops the model and motivates our methodological contribution. Section 3 describes the model estimation procedure, which is both efficient and effective. Section 4 provides an illustrative empirical application with an evaluation of the model's performance relative to the multivariate probit benchmark. Section 5 concludes by summarizing our contributions and noting some avenues for future research.

2. Model Development

We first describe the latent variable model in general and then, the basic RBM as it applies to modeling large crosscategory purchase data. Next, we extend the RBM using a CRBM structure to model the contemporaneous and intertemporal crosscategory effects efficiently while including other variables that might impact consumer behavior.

2.1. Latent Variable Model

In probabilistic modeling, associations between variables can be captured in several ways. The "coincidence" of, say, milk and bread purchases can be represented simply by a correlation coefficient. Alternately, one can explicitly model an unobservable common cause driving purchases of both milk and bread, which can help *explain* the coincidence.[3]

A nonlinear latent structure has several advantages over (linear) correlations in modeling associations. First, the correlation matrix assumes that all variables are Gaussian, which works well if the associations are indeed linear. For latent variable models, the association can be linear (linear factor model) or nonlinear (RBM), with the latter providing greater modeling flexibility (Burnap et al. 2014). Second, although correlations capture pairwise associations, a latent variable structure can model associations among multiple variables. Third, in consumer choice modeling, there can be various types of associations among purchasing decisions, where a single correlation matrix is insufficient to represent them all. For example, at the universal product code (UPC) level, there are associations within and between brands as well as within and between product categories.

2.2. Restricted Boltzmann Machines

RBMs, unlike nonprobabilistic feedforward neural networks, such as convolutional neural network, are essentially hierarchical Bayesian models with structures that can be adjusted and interpreted just as for "traditional" statistical models, albeit with latent structures. Researchers can construct the RBM based on theory or hypotheses and gain insights from the parameters and latent variables. In theory, the RBM is a probabilistic model that defines the joint distribution of binary visible and hidden (latent) variables (Hinton 2002). The network distribution is based on the energy function in physics:

$$E(\boldsymbol{X},\boldsymbol{H}) = -\sum_I \sum_J \beta_{ij} x_i h_j - \sum_I \beta_i x_i - \sum_J \beta_j h_j, \tag{1}$$

where $\boldsymbol{X}$ is a vector of $\boldsymbol{I}$ visible variables and $\boldsymbol{H}$ is a vector of J hidden variables. High-energy states tend to be unstable, and hence, they are rare events in statistical terms. The joint probability distribution function of the network then should assign a low probability to high-energy states and a high probability to low-energy states:

$$\begin{aligned} P(\boldsymbol{X},\boldsymbol{H}) &= \frac{\exp(-E)}{z} \\ &= \frac{\exp(\Sigma_I \Sigma_J \beta_{ij} x_i h_j + \Sigma_I \beta_i x_i + \Sigma_J \beta_j h_j)}{z}. \end{aligned} \tag{2}$$

The exponential function ensures that the probability is nonnegative; z is the normalizing constant. In the basic RBM, all variables are binary. Instead of using correlation coefficients to capture the interdependencies between visible variables directly, the RBM uses hidden variables to model the multivariate distribution in the visible layer. The structure of the RBM is analogous to that of the latent factor model—all of the visible variables are connected to the hidden variables, and there are no direct connections among visible variables. However, unlike the factor model, where only the conditional distribution of visible variables based on hidden variables is defined, the RBM has a symmetric structure between hidden and visible layers so that the conditional distributions are defined in both directions (Figure 1).

Following Bayes' rule, the joint distribution of the visible variables conditioned on all hidden variables is

$$P(\boldsymbol{X} = 1|\boldsymbol{H}) = \frac{P(\boldsymbol{X} = 1, \boldsymbol{H})}{\Sigma_X P(\boldsymbol{X},\boldsymbol{H})} = \prod_X \frac{\exp\left(\Sigma_J \beta_{\mathrm{ij}} \mathrm{h_j} + \beta_{\mathrm{i}}\right)}{\exp\left(\Sigma_J\, \beta_{ij} h_j + \beta_i\right) + 1}. \tag{3}$$

Equation (3) shows an interesting property of the RBM: that all visible variables are conditionally independent or

$$P(\boldsymbol{X} = 1|\boldsymbol{H}) = \prod_X P(x_i = 1|\boldsymbol{H}).$$

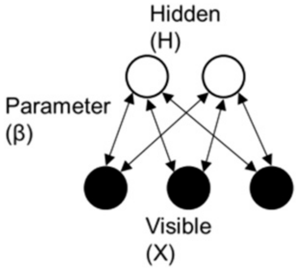
Because of the symmetric structure of the RBM, the hidden variables are also independent conditioned on all visible variables. In our model setting, the visible layer consists of binary variables representing the observed purchasing decisions for a single household s.[4] Let i index a product category and t denote a time period, typically a week. The variable x_{sit} in the visible layer takes on the value of one if household s purchased at least one unit of one item from category i in period t and zero otherwise. The conditional distribution of each visible variable takes the simple logistic form

$$P(x_{sit} = 1|\boldsymbol{H}_s) = \frac{1}{1 + \exp\left(-\beta_i - \Sigma_J \beta_{ji} h_{sj}\right)}, \tag{4}$$

where $\boldsymbol{H}_s$ is a household-specific J-dimensional hidden vector, $\boldsymbol{H}_s = (h_{s1}, h_{s2}, \cdots h_{sJ})$, and β_i is a constant term for each product category.

For multicategory analysis ignoring intertemporal effects, our task is to model the joint distribution $P(x_{s1t}, x_{s2t} \ldots x_{s3t})$. In the MVP model, the symmetric covariance matrix captures the crosscategory associations explicitly. As discussed, the MVP becomes rapidly

Figure 1. Graphical Representation of the Restricted Boltzmann Machine



inefficient (computationally) as the number of categories increases. In contrast, the RBM captures relationships among the variables in time period t with far fewer parameters than are necessary for MVP by using hidden variables to capture relationships among the variables in the visible layer. In fact, the complexity of the RBM model grows at a rate less than $O(n)$.

The inclusion of hidden variables in the RBM model is consistent with the intuition that purchasing decisions are influenced by unobservable household attributes (Shocker et al. 2004). For instance, some households may be more likely to buy frozen pizza for dinner than vegetables owing to different lifestyles, which may not be observable but can be inferred from each household's purchasing pattern. One can use unobservable household attributes to explain the relationships among purchasing decisions. Being unobservable, they are modeled as binary hidden variables, with values derived from the observed purchasing decisions. Several hidden variables may be needed to describe a household's complex purchasing behavior. The household decisions then become a mixture of patterns determined by these hidden variables.

For illustration, as part of our empirical application described later in Section 4, we analyzed household decision patterns for the soup, pasta sauce, dog food, and pasta categories, which are ordered in terms of purchasing frequency. We used two hidden units in a two-layer RBM to capture the associations. Table 1 shows the parameter values of the connections between hidden and visible variables.

Just as in factor analysis, we can infer the associations between purchasing decisions by examining the signs and magnitudes of the parameters. For each hidden variable, the product categories are positively associated if the connections to those product categories have the same sign. The magnitudes capture the strength of the associations. Viewing the connections to the first hidden variable in Table 1, the strongest associations are among soup, pasta sauce, and pasta (all positive), whereas dog food is independent of other product categories. The parameters of the second hidden units tell a slightly different story: only soup and pasta sauce are positively associated. Households in general would exhibit either purchasing pattern or a mix. The parameters are akin to those in logistic regression if we think of the hidden units as the independent variables and purchase decisions as the dependent variables.

The binary hidden variables can be viewed as discrete labels that summarize household shopping patterns, performing the task of market segmentation implicitly. For example, with two hidden variables, we have 2^2 possible combinations of hidden variables, yielding 4 market segments based on different shopping patterns.

Table 1. RBM Coefficients for Small-Scale Illustrative Application

	Hidden variable 1	Hidden variable 2
Soup	–0.752	–0.898
Pasta sauce	–0.066	–0.268
Dog food	0.000	0.000
Pasta	–0.201	0.000

2.3. Proposed Model: Conditional RBM

The basic RBM model only captures contemporaneous crosscategory effects. However, we know from the marketing literature that households' past purchasing decisions are predictive of their future purchases (Keenan 1982, Fader et al. 2005). In practice, contemporaneous and intertemporal effects have been typically modeled and analyzed separately owing to computational power constraints. We propose a model adapting an extension to the RBM, the CRBM, to incorporate both contemporaneous and intertemporal effects for a complete picture of household purchasing patterns. The base model is the same RBM described in previous section:

$$P(\boldsymbol{X}_{st}, \boldsymbol{H}_s) = \frac{\exp(-\boldsymbol{\beta}_X \boldsymbol{X}_{st} - \boldsymbol{\beta}_H \boldsymbol{H}_s - \boldsymbol{\beta}_{XH} \boldsymbol{X}_{st} \boldsymbol{H}_s)}{z}. \qquad (5)$$

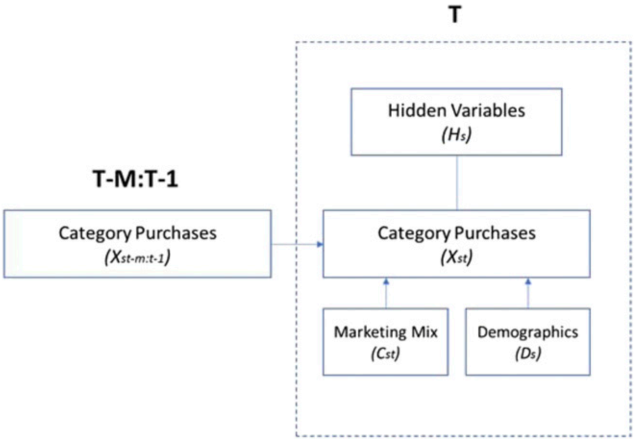
All parameters and variables are in vector form in the equation. For example, $\boldsymbol{X}_{st}$ is a binary vector representing purchasing decisions of household s over the set of product categories at time t. The literature suggests that household demographics (such as size and income) may account for some of the heterogeneity in purchasing patterns. Moreover, marketing mix variables, such as price and promotion, have causal effects on buying decisions. We incorporate these effects by modeling the parameter vector β_X as a function of marketing mix $\boldsymbol{C}_{st}$ at time t and household demographics $\boldsymbol{D}_s$. Finally, the intertemporal effects are captured similarly so that

$$\boldsymbol{\beta}_X = \boldsymbol{\beta}_C \boldsymbol{C}_{st} + \boldsymbol{\beta}_D \boldsymbol{D}_s + \sum_{t'=T-m}^{m} \boldsymbol{\beta}_{t'} \boldsymbol{X}_{st'}, \qquad (6)$$

where $X_{st'}$ is a vector of purchasing decisions in each of the previous m periods. Figure 2 provides a graphical representation of the full model constructed to capture the effects and interactions discussed above.

Although we cannot determine the direction of contemporaneous effects a priori, we know with certainty that the intertemporal effects are directional and should be modeled as such.[5] In a CRBM, we add another type of connection to the model that is directional and may imply causal effects. A graphical model with directed edges is called a "directed acyclic graph" (DAG) or Bayesian Network, and it has been widely used for causal inference (Pearl 2014). Thus, intertemporal

Figure 2. (Color online) Graphical Representation of the Proposed CRBM Model



effects should be modeled using a DAG as should the effects of demographics and marketing mix variables. The joint distribution of hidden and visible variables (represented by the RBM) is *conditioned* on these causal variables—hence, it is a CRBM. In essence, we have a hierarchical Bayesian model with an implicit prior over the hidden units. A random effects specification captures unobserved heterogeneity, with the hidden units sampled for each household. The joint distribution of visible and hidden variables is expressed in standard Bayesian form as

$$f(\boldsymbol{X}_{st}, \boldsymbol{H}_s) = f(\boldsymbol{X}_{st}|\boldsymbol{H}_s)f(\boldsymbol{H}_s), \tag{7}$$

where f denotes the probability density function. The alternative representation indicates that we can treat household s's purchasing decision at time t as a sample from a conditional distribution determined by the hidden variables at the household level. The hidden variables, which can be interpreted as parameters representing each household, are then sampled from an implicit common prior defined by integrating out $\boldsymbol{X}$ in the CRBM. We can also make the implicit common prior explicit by adding another layer of hidden units. Other variables, such as price and promotion, do not change the hierarchical structure, because they are constructed as independent of the hidden variables.

Unlike the intertemporal effects, the contemporaneous effects are captured implicitly by the hidden variables. We can infer the cross effects from the partial derivatives $\frac{\partial x_{it}}{\partial x_{jt}}$. By the chain rule, $\frac{\partial x_{it}}{\partial x_{jt}} = \frac{\partial x_{it}}{\partial H_s}\frac{\partial H_s}{\partial x_{jt}}$, and the partial derivative $\frac{\partial x_{it}}{\partial x_{jt}}$ changes with the values of x_{it} and x_{jt}. We replace x_{it} and $\boldsymbol{H}_s$ with their expected values (Hruschka 2014) and use variational inference for efficient estimation (see Appendix B for details).

The CRBM has been successfully used in video and image processing (Taylor and Hinton 2009, Salakhutdinov et al. 2013). Unlike video and image data, household decision patterns are often very sparse, and large fractions of purchasing decisions across time and product categories are independent of each other. Hence, only a subset of all possible cross-category intertemporal effects should be in the final model. In other words, many of the parameters in the original model should be zero to make the final model parsimonious, interpretable, and accurate. Our proposed model combines a dropout method with regularization, which extends current CRBM models, and captures such sparse data as household decision patterns efficiently as discussed next.

3. Model Estimation

The parameters of the CRBM model are estimated numerically by searching for the values that maximize the log likelihood of the training data. A first-order algorithmic approach to the maximization is to calculate the gradient of the log-likelihood function with respect to the parameters and then, use that gradient to perform a search. The bipartite CRBM structure implies that each component of the gradient has the form

$$\frac{\partial \log P(x_{sit})}{\partial \beta_{ij}} = E(x_{sit}h_{sj})_{data} - E(\hat{x}_{sit}\hat{h}_{sj})_{model}, \tag{8}$$

where β_{ij} is the parameter capturing the interaction between the purchase decision variable for category i and hidden variable h_{sj}. The first term in Equation (8) is the expected value of the product of x_{sit} and h_{sj} over households and time periods in the training data, and the second term is the expected value over the distribution of the same product assumed by the model. The second term could be estimated by brute force by conducting Gibbs sampling to generate samples from the model with random initial values for $\hat{x}_{sit}$ and $\widehat{h_{sj}}$ and then, calculating the average of the product of $\hat{x}_{sit}$ and $\widehat{h_{sj}}$ for each product category. However, such a naïve sampling approach would require a lot of training time, a problem for large-scale data analysis. Although the RBM is differentiable and has a simpler structure than the MVP, we need an efficient model training algorithm for parameter estimation.

3.1. Methods to Improve Efficiency—Contrastive Divergence, Momentum, and Minibatch

Various methods have been proposed for making the inference procedure efficient. Contrastive divergence (Hinton 2002) estimates the data-dependent statistics $E(x_{it}h_j)_{data}$ in Equation (8) by a one-step sampling process. The hidden variables H are drawn conditional on the original training data. For the data-independent term, $E(\hat{x}_{it}\hat{h}_j)_{\text{model}}$, we first sample hidden units from the original data and then, sample simulated data from the sample hidden units. The average value of the product of the simulated data and the hidden variable sample is then used as the data-independent term in the gradient. Although one-step Gibbs sampling cannot yield an accurate estimate of $E(\hat{x}_{it}\hat{h}_j)_{\text{model}}$, the gradient based on the sampling method has been shown to work well in practice (Hinton 2002).

We shortened the time to convergence in parameter estimation by using the *momentum method* and *batch gradient descent* (Salakhutdinov and Larochelle 2010). The momentum method speeds up the learning rate when the direction of the gradient is unchanged on consecutive iterations and slows it down when the gradient changes direction. This process yields the optimal parameter value in far fewer iterations than if we use a constant step size.

The updated equation is as follows (Hinton 2010):

$$\Delta\beta_{ij}(t) = \alpha\Delta\beta_{ij}(t-1) + \varepsilon\frac{\partial \log P(x_{it})}{\partial\beta_{ij}}(t), \tag{9}$$

where t and $t-1$ indicate the current and previous iterations, $\Delta\beta_{ij}(t)$ is the parameter increment at $t-1$, α is the momentum, ε is the learning rate, and $\log P(x_{it})$ is the log-likelihood function being maximized.

Dividing the training data into batches and updating the parameters after each batch can also speed up the model training process, because a group of smaller matrices can typically be processed faster than can a single very large matrix (Cotter et al. 2011). Both gradient descent and stochastic gradient descent are essentially variations of batch gradient descent (Bottou 2010). Stochastic gradient descent uses each single point as a batch, whereas traditional gradient descent uses the whole training data as a batch. In our empirical application, we used each household's shopping data as a batch.

3.2. Methods to Improve Estimation of Sparse Models—Elastic Net and Dropout

Across many product categories and over several time periods, only small fractions of all household purchases are likely to be truly interrelated. Hence, we only need a subset of the modeled effects in the final model to improve its predictive accuracy and interpretability. Without a proper parameter selection process, the model is prone to overfitting. According to statistical learning theory, a complex model can fit any data very well but often fails to separate the idiosyncrasies of the training set from the true general patterns in the data, thus performing poorly on testing data.

Marketing researchers have normally relied on conventional methods, such as stepwise regression or bootstrapping, for model (feature) selection. These methods do not scale easily for large-dimension models and often, do not improve predictive accuracy. Another popular option is to use analytical measures, such as information criteria (Akaike information criterion [AIC] or Bayesian information criterion [BIC]). These criteria require the estimation of the normalizing constant for Markov random field models, such as our model, and also, the estimation of all possible models for comparison, which could be very time consuming.

L1 (*Lasso*) and L2 (*Ridge*) *regularization* methods have been used to achieve sparsity (L1 only), prevent overfitting, and improve prediction accuracy.[6] They place a constraint on the value of parameters during the model training process, encouraging movement of parameter values toward zero. From a Bayesian perspective, we place a Laplace (L1) and a Gaussian (L2) prior on the parameters, both zero centered and symmetric, to reduce the effect of the data-specific idiosyncrasies on posterior estimation.

L2 regularization shrinks the parameter values to prevent overfitting, but no parameters are dropped from the model. Therefore, L2 regularization may improve model prediction accuracy but does not help with model sparsity. In contrast, L1 regularization actually pushes some parameters to be exactly zero, leading to a sparse model with fewer (nonzero) parameters. However, when there are many parameters and some variables are highly correlated, L1 would lead to unstable model estimates. Combining L1 and L2 in model estimation—the *elastic net* (Zou and Hastie 2005)—is an effective way to prevent overfitting and achieve a sparse model in big data analysis. Such mixed regularization methods have been successfully used in linear and logit regression, and they are often based on coordinate descent.

There have been applications of L1 regularization in the RBM and the CRBM (Salakhutdinov et al. 2013). Most have been implemented by using either subgradient methods that ignore the nonsmoothness of the L1 penalty term or truncation methods in which parameter values below an arbitrary threshold are dropped during model training. We have used the elastic net on models with 4 and 100 product categories to examine its effect on model performance. With only four categories, the elastic net does not show any significant effect on model fit or performance. With 100 categories, testing error reduces considerably with 100 of categories. Without the elastic net, the 100-category model fits the *training* data much better than the 4-category model, with negligible training error. However, the *testing* error is 9.37% higher for 100 relative to 4 categories owing to overfitting. Model interpretability is the second goal of the elastic net. We seek to identify a small subset of predictors with the strongest effects in the model to gain meaningful insight (Hastie et al. 2016). In our case, after applying the elastic net to the 100-category model, fewer than two-thirds of the parameters remain.[7]

In estimating our model with the elastic net, we successfully merged the proximal method with the contrastive divergence (CD) training algorithm to achieve model training efficiency and sparsity (see Appendix C for details). The proximal gradient method allows for searching for the optimal path in gradient descent when part of the objective function is nondifferentiable. In this instance, the L1 penalty term is not differentiable at zero, and conventional gradient methods, such as CD training, do not work properly with the regularization term.

For high-dimension CRBM models, the large number of active hidden variables also contributes to overfitting (in addition to the parameters). These variables affect the associations among visible variables. Although we only need a limited number of activated hidden variables to capture the data distribution, the elastic net does not constrain their number. Crossvalidation can determine the optimal number of hidden variables but would be very time consuming for high-dimensional data.

The *dropout method* (Salakhutdinov and Larochelle 2010, Makhzani and Frey 2015) has been shown to effectively prevent overfitting and improve a neural network's predictive performance. This method requires a change to the implementation of the training algorithm. During each batch training phase, some hidden variables are randomly dropped from the model with probability p. After the model training is complete, the parameters are multiplied by p, and all hidden variables are retained in the final model. The dropout procedure prevents hidden variables from trying to fit the idiosyncratic characteristics of the training data, and it can also be viewed as a form of model averaging.

We evaluated the effect of elastic net and dropout on our proposed model. We found that neither elastic net nor dropout provided much benefit when the model is simple (e.g., with only four products and no intertemporal effects). Regularization methods improved model fit when the model becomes complex. With a large number of hidden variables, dropout plays a greater role in improving model fit than the elastic net. Detailed results of our comparative evaluation are available in the online appendix.

4. Illustrative Empirical Application

We reiterate that our empirical study is to illustrate how the model can be applied and to compare its predictive performance and robustness relative to the traditional MVP model, even in the case of relatively small applications with few product categories and/or time periods. We describe the data first (Section 4.1) followed by the comparison with the MVP model (Section 4.2). Next, we report the results of the application of the CRBM to the large-scale model with up to 100 product categories and 12 time periods (Section 4.3) and discuss the implications of marketing mix (Section 4.4) and crossproduct and intertemporal effects (Section 4.5). Finally, we show how an augmented ("deep") CRBM can be used to estimate the model at the UPC rather than product category level (Section 4.6).

4.1. The Data

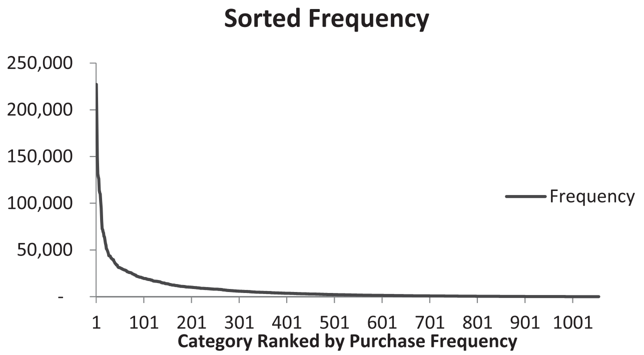
The Nielsen household panel data set (provided for the sole purpose of academic research) contains households' shopping histories with their demographic information for a total of 6,695,833 observations. Each observation is a purchase transaction made by a household between 2004 and 2008. In the data set, the three most frequented stores are Kroger, Walmart, and Meijer. There are 4,000 households' purchasing histories over 208 weeks and across 1,055 product categories (as defined by Nielsen). Each household in the data set averaged 2.74 shopping trips a week. We selected 2,086 households with complete purchasing histories and demographic information for our analysis.

The 1,055 product categories vary significantly in the frequency of their purchase, with milk the most frequently purchased category. Figure 3 plots the category purchase frequencies (based on purchasing histories across households), with the categories sorted in decreasing order of frequency. Note the "long tail" of infrequently purchased product categories. For our illustrative application, we selected the 100 most purchased categories, including products from dairy milk to nutritional supplements.

We converted each household's purchasing history into a binary matrix. The rows represent the 208 weeks recorded in the data set, and the columns represent the 100 product categories ordered in decreasing overall purchase frequency. For example, if a household purchased at least one unit of one item from the dairy milk category in week 1, then the value in row 1, column 1 is one and zero otherwise (no purchase). There are 2,086 such matrices in all, one for each selected household. We randomly selected 1,500 matrices for the training set and used the remaining 586 matrices for testing the model.

Only demographic variables that demonstrated effects on purchasing decisions in previous research

Figure 3. Product Category Purchase Frequencies



(e.g., the presence of children, household income, and household size) were initially included (Manchanda et al. 1999, Hansen et al. 2006). We also trained and tested models with promotion and price indices at the product category level. These indices were constructed following approaches in previous multicategory research (Manchanda et al. 1999, Duvvuri et al. 2007, Duvvuri and Gruca 2010). For the price index, we used the unit price actually paid if the household made a purchase in the category that week; if not, a weighted average based on the household's purchasing history across all brands in that category was computed as the price index. The promotion index was similarly constructed: if there was a promotion in the week of purchase, the index was one; otherwise, it was the weighted average.

We did not include a household inventory variable in the model owing to multiple concerns with the inclusion of inventory variables: collinearity with the intertemporal effects, the strong assumption of a constant rate of consumption, and endogeneity (Manchanda et al. 1999). The inventory effect is implicitly captured by the intertemporal effects. Furthermore, we found that the assumption of a constant rate of purchase does not hold for this data set, even for products such as milk. We examined the data using model-free approaches first, which confirmed the irregularity of households' purchasing behavior. On average, a household buys milk about every 2.6 weeks; the conditional frequencies also show that a household is more likely to buy milk after a milk purchase in the previous week than otherwise. Our final model also captured this interesting, nonobvious intertemporal pattern.

4.2. Model Comparison Vs. Multivariate Probit

4.2.1. RBM Vs. MVP. We first compared the performance of a simple RBM with no time lags with that of the MVP. No household heterogeneity, marketing mix effects, or regularization methods were considered for either model to ensure a fair comparison. Given the MVP's ability to handle only a small number of categories owing to computational constraints, we limited the analysis to four categories—soup, pasta sauce, dog food, and pasta—which have been used in previous multicategory research (pasta and pasta sauce are complements; dog food is considered to be independent of the other three categories). The result was consistent with the finding of Hruschka (2014) that the basic RBM performs at least as well as the MVP (Table 2).

Table 2. Model Comparison: False Positive and Negative Rates for RBM and MVP Models (Four Product Categories)

	False positive	False negative
RBM	0.0679	0.8346
MVP	0.0681	0.8369

4.2.2. CRBM Vs. Hierarchical MVP. Our proposed model is not a basic RBM but a CRBM incorporating intertemporal effects and unobserved heterogeneity. Hence, to be fair, we compared our proposed model with a *hierarchical* multivariate probit (HMVP). The HMVP has a very similar model structure (using a covariance matrix to model the contemporaneous effects instead of hidden variables), and the household-level covariance matrix is sampled from a population prior. Details on the HMVP are provided in the online appendix.

For the comparison, we examined the same four product categories but now, considered intertemporal effects over four weeks. We calculated false positive and false negative rates across all basket combinations, and once again, the results show that, although the HMVP achieves the same false negative rate, the CRBM performs better in terms of the false positive rate (Table 3). We have also included a confusion matrix to visualize the predictive performance of the CRBM relative to the HMVP in the online appendix.

For additional comparison, we used receiver operating characteristics (ROC) curves to measure the performance of both HMVP and CRBM at various thresholds settings (Figure 4). CRBM outperformed HMVP as demonstrated by the area under the ROC, especially when the threshold is in the intermediate range; both models outperformed a random classifier (denoted by the dotted diagonal line in Figure 4).

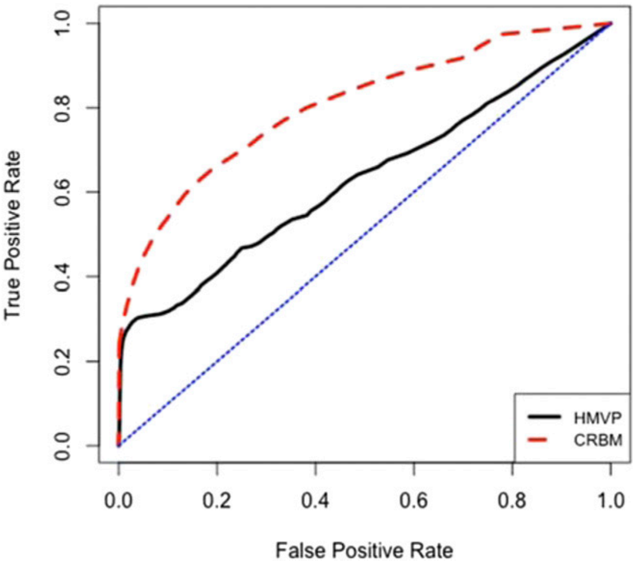
4.2.3. Extension to 10 Categories. We then extended the model to the 10 most frequently purchased product categories, including price and promotion, and found that the CRBM again outperformed the MVP in terms of model fit and predictive accuracy.

The CRBM can be viewed as the nonlinear version of the MVP model with correlated latent factors. We, therefore, also compared our model with an MVP with latent linear factors and found again that the CRBM performs better. Its flexible structure can capture more complex associations between purchasing decisions with fewer hidden variables, reducing dimensionality without sacrificing performance.

Table 3. Model Comparison: False Positive and Negative Rates for CRBM and HMVP Models (Four Categories over Four Weeks)

	False positive	False negative
CRBM	0.0456	0.6919
HMVP	0.0580	0.6920

Figure 4. Model Comparison: ROC Curves for HMVP and CRBM Models



4.2.4. Estimation (Model Training) Time. A huge advantage of the RBM and the CRBM over the MVP is the vastly superior efficiency in terms of the estimation or model training time. The difference in efficiency becomes more pronounced as the dimensionality of the model increases. In our experiments, we compared the training time for the RBM or CRBM with MVP/HMVP (with MCMC estimation at 4 [without and with intertemporal effects] and 10 product categories, respectively) (see rows 2 and 3 in Table 4). With 10 categories, it was not possible to estimate the multiperiod intertemporal HMVP model.

The drastic reduction in training time for CRBM is critical when it comes to big data analysis. Although this efficiency is the most significant advantage of our model, we have shown that, even with small data sets, the CRBM performs at least as well as MVP (actually better using mean square error and false positive rate for comparison). In terms of statistical computation, we used stochastic gradient descent, which has a runtime bound. In contrast, for the MVP, where the sampling method was used, there is no guaranteed time bound (Shai and Ambuj 2011).

Recent developments in variational inference offer efficient alternatives to Bayesian model training by casting posterior inference as an optimization problem to improve estimation efficiency while sacrificing some accuracy (Girolami and Rogers 2006, Blei et al. 2017). Keeping the structure of the HMVP model, we trained the model by automatic differentiation variational inference (ADVI), which uses stochastic gradient descent to find the optimal parameter values (Carpenter et al. 2017, Kucukelbir et al. 2017). Monte Carlo integration is still needed to estimate the gradient in most cases, which slows down the process. In our experiment, it took over 46 minutes to estimate the HMVP model with four time lags with multithread parallel computing for sampling enabled (Carpenter et al. 2017). Although ADVI improved model training efficiency relative to MCMC (see row 4 in Table 4), accuracy declined significantly, with a 100% higher false positive rate. In comparison, our CRBM is considerably more efficient and accurate.

4.3. Large-scale Crosscategory and Intertemporal Effects

Having demonstrated our model's dramatically greater efficiency relative to the MVP with low-dimensional data with comparable (or better) model fit and predictive performance, we next expanded our analysis to include all 100 product categories and a 12-week horizon for intertemporal effects to examine purchasing patterns more comprehensively. In essence, we need to model the associations between 1,300 ($12 \times 100 + 100$) categories × time periods using regular computing power, which to our knowledge, has not been attempted in the marketing literature. Estimation of the more traditional multivariate models is simply not feasible given the size of the data.

We first trained the four-product category model with a four-period time lag (model 4P4T).[8] Price and promotion indices and demographic variables, with possible causal effects on household purchasing decisions, were also incorporated. The joint distribution of decision outcomes is thus conditioned on these variables. We set the range for the regularization penalty term λ to be between 0.0005 and 0.05. The initial value of λ was set at 0.05 and then decreased by 0.00025 after each epoch; α was set at 0.6 to put more weight on the L1 regularization in the elastic net. A momentum term of 0.9 was applied to the training process after five epochs. We used the recommended dropout rate of 0.5 for training (Srivastava et al. 2014). These values are typical for CRBM training (Taylor and Hinton 2009), and we did not find a significant difference in results for different values, consistent with Hinton (2010). Optimal values for the learning rate and penalty terms can be determined by cross-validation; in our case, we did not find it necessary, because the improvement was very limited.[9] Each

Table 4. Model Training Times for HMVP and CRBM Models

Model Number of categories Intertemporal effects?	4 No	4 Yes (four periods)	10 No
RBM/CRBM, seconds	<30	<60	<120
MVP/HMVP (MCMC), hours	>7	>0.50	>90
MVP/HMVP (VB), minutes	>35	>46	>57

Note. VB, Variational Bayes.

Table 5. Own Effects and cross Effects of Promotional Index

	Purchase			
Promotion	Soup	Pasta sauce	Dog food	Pasta
Soup	5.308	−0.321	−0.034	−0.115
Pasta sauce	−0.125	6.728	−0.001	0.434
Dog food	−0.166	−0.030	5.166	−0.003
Pasta	−0.074	0.143	0.000	6.712

household's purchase history over the 208 weeks was used as a minibatch during training.

The own effects and cross effects of promotional index are consistent with the marketing literature (Table 5). The signs of parameter coefficients indicate positive or negative effects, whereas their values capture the effect magnitudes. The own effects are much greater across all four product categories than the cross effects as expected. Pasta sauce shows the strongest own effects. Most cross effects are negative; for example, promotion of dog food negatively affects sales of the other three products. For pasta and pasta sauce, the cross effects are positive, which we would expect between complements.

Table 6 shows the coefficients that represent the intertemporal effects. Most intertemporal effects are positive. Within-category effects seem to be positive for all products and across all time lags. Although dog food purchase does not seem to be affected by purchases of other products in previous weeks, its purchase suggests a positive indication that the household may buy the other products in the following week.

Next, we compared CRBM models with varying numbers of time lags and product categories to examine whether the added complexity could improve model fit and performance. If there are strong interdependencies in purchasing patterns across several categories and if some of the relationships are intertemporal, a model with more product categories and time lags should be able to perform better in capturing the decision pattern and making predictions.

For comparison, we trained a model with no time lags (4P). We then added four time lags (4P4T) to examine if the added time dimension was beneficial. We also calibrated models with 100 product categories and 12 time lags. We compared models with different time lags and discovered that the testing error dropped as the number of lags was expanded to 12 weeks but then, stayed relatively flat. This may be because of most of the products in the data set having a shopping cycle of 12 weeks or less. We used the mean squared error (the Brier score) as the error measure for both testing and training errors. The testing error should be a better metric for model comparison and model selection, because it is a better approximation of the generalization error than other metrics, such as AIC or BIC, which are based on training errors (Friedman et al. 2001). Tables 7 and 8 display the testing and training errors (in terms of both Brier scores, false positives, and false negatives) for the various CRBM-based models.

The results from our illustrative application suggest that there are indeed intertemporal crosscategory associations among a large number of products. In general, we expect that incorporating these associations would improve both fit and prediction. The numbers of both time lags and product categories matter. We also noticed variations in intertemporal crosscategory effects on different product categories. Dog food, which seems to be independent of the other three categories, saw a larger decrease in the training and testing errors than the others as the number of time lags increased. The intuition is that, for purchases in categories that tend to be independent of other categories, past purchases might be more strongly indicative of future purchases of the same product.[10]

4.3.1. Robustness Check.

To examine the robustness of the parameter estimates, we compared models with the top 100 and top 200 product categories. All within-product category effects, such as price and promotion, and most other parameters for the top 100 categories remain nonzero in the model with the top 200 categories. The number of active hidden variables increased as we increased the number of product categories, which however, did not seem to affect the number of parameters that remained nonzero. The categories outside the top 100 have very sparse

Table 6. Intertemporal Effects for Four Products over Four Time Lags

	T			
	Soup	Pasta sauce	Dog food	Pasta
$T-1$				
Soup	0.302	0.013	−0.080	−0.016
Pasta sauce	0.047	0.055	0.000	0.031
Dog food	0.010	0.063	1.023	0.000
Pasta	−0.015	0.009	0.000	0.086
$T-2$				
Soup	0.495	0.164	−0.038	0.155
Pasta sauce	0.149	0.414	0.004	0.170
Dog food	0.010	0.200	1.767	0.147
Pasta	0.061	0.026	0.000	0.344
$T-3$				
Soup	0.479	0.159	−0.019	0.092
Pasta sauce	0.146	0.472	0.007	0.141
Dog food	0.015	0.130	1.712	0.091
Pasta	0.063	0.020	0.000	0.510
$T-4$				
Soup	0.527	0.231	−0.009	0.097
Pasta sauce	0.174	0.515	0.008	0.178
Dog food	0.013	0.224	1.652	0.083
Pasta	0.052	0.129	0.000	0.484

Table 7. Comparison of Errors Across CRBM Models

	Brier score (mean squared error)						
Testing error	Soup	Pasta sauce	Dog food	Pasta	All products	False positive	False negative
4P	0.0947	0.0509	0.0271	0.0402	0.0532	0.0461	0.7244
4P4T	0.0925	0.0506	0.0238	0.0403	0.0518	0.0456	0.6919
4P12T	0.0910	0.0501	0.0230	0.0403	0.0511	0.0464	0.6770
100P4T	0.0908	0.0500	0.0239	0.0400	0.0500	0.0443	0.6775
100P12T	0.0896	0.0495	0.0229	0.0399	0.0490	0.0476	0.6541

Notes. 4P and 100P indicate CRBM models with 4 and 100 product categories, respectively. 4T and 12T represent 4 and 12 time lags in the models, respectively. The all products column shows the mean errors over four product categories in the model.

crosscategory associations, and the original 100 × 100 associations are essentially qualitatively unchanged. The value of the elastic net penalty term does affect the number of nonzero parameters. Nonzero parameters in the model selected by the elastic net do not mean that they are statistically significant from a frequentist perspective; they can be viewed as having nonzero posterior expected values from a Bayesian perspective. Therefore, the results should be seen as exploratory and in need of testing via confirmatory experiments.

4.4. Examination and Discussion of Marketing Mix Effects

Note that the proposed CRBM model is designed to predict shopping behavior of a household on the basis of the marketing mix, past behavior (say, over the last 12 weeks), the hidden variables associated with the household, and key demographics. In this subsection, we discuss the effects of promotion based on the model parameter estimates. In the following plot (Figure 5), each column represents the product category purchasing decision at time T, and each row represents the product promotion index. The categories are ordered in decreasing purchase frequency from left to right and from top to bottom. The value in each cell represents the effect of promotion on the product purchase decision shown as color-coded "pixels"—blue for a positive effect, red for negative effect, and white for no significant effect of one product's promotion on the purchase of the product indicated by the column. The pattern formed by all of the pixels provides an initial "big picture" across 100 categories.

Although we note that the findings are indicative and require far more exhaustive empirical and experimental research for substantive interpretation, they provide some face validity and tentatively suggest the type of less obvious implications that warrant additional research. As expected, the promotional effect is always positive within a category (the blue pixels along the diagonal). Also, in general, promotions of the more frequently purchased products have greater effect on the purchase of other products than promotions of the more rarely purchased products. Unlike within-category promotional effects, the crosscategory effects can be either positive or negative. For example, according to the plot, the promotion of milk implies a greater likelihood of purchasing cereal but a lower likelihood of buying carbonated soft drinks, which makes intuitive sense. Promotion of rarely purchased products (e.g., dog treats) has little or no effect on other product purchases: hence, the many white dots along the rows for such categories.

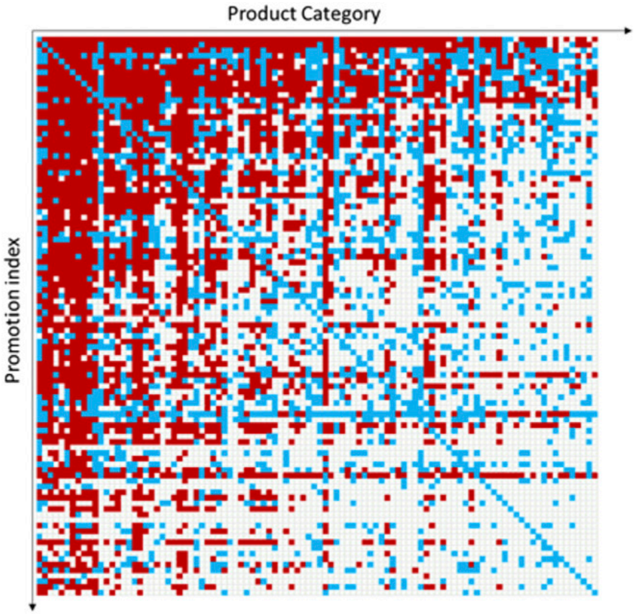
Interestingly, there are also product promotions that are associated with reduced purchases of almost all other products. For example, the promotion of cigarettes coincides with lower sales of most products

Table 8. Comparison of Errors Across CRBM Models

	Brier score (mean squared error)						
Training error	Soup	Pasta sauce	Dog food	Pasta	All products	False positive	False negative
4P	0.0794	0.0334	0.0198	0.0242	0.0392	0.0496	0.5501
4P4T	0.0785	0.0332	0.0166	0.0241	0.0381	0.0486	0.5127
4P12T	0.0776	0.0330	0.0157	0.0238	0.0375	0.0500	0.5002
100P4T	0.0760	0.0328	0.0169	0.0241	0.0385	0.0453	0.519
100P12T	0.0747	0.0325	0.0158	0.0237	0.0375	0.0462	0.513

Notes. 4P and 100P indicate CRBM models with 4 and 100 product categories, respectively. 4T and 12T represent 4 and 12 time lags in the models, respectively. The all products column shows the mean errors over four product categories in the model.

Figure 5. Plot of Promotional Effects



(i.e., the corresponding row has mostly red dots). This may be because smokers are more likely to visit the supermarket just for this promotion compared with their usual shopping trips when cigarettes are not on sale. The price effects show similar patterns as those of promotion. In general, all within-category price effects are negative, and most price effects are concentrated on the most frequently purchased products.

The effect of household size on purchasing decisions is also interesting. As expected, the likelihood of purchase increases in the size of the household but only up to a point—beyond a household size of four, the effects disappear in most cases. Perhaps larger households are constrained in terms of purchasing power. Future research can examine how these effects differ across product categories (e.g., necessities versus nonessentials).

4.5. Examination and Discussion of Intertemporal and Contemporaneous Effects

Although most intertemporal own category effects are positive, the results suggest that including cross effects over multiple categories and over multiple periods can capture a more complete picture of household purchasing patterns, which can be complex, asymmetric, and time varying. Figure 6 plots the intertemporal effects at $T - 1$ as an illustration. Similar to the previous plot of promotional effects, each column represents the product category purchasing decision at time T, and each row represents product category purchasing decisions at $T - 1$. The categories are ordered in decreasing purchase frequency from left to right and from top to bottom. Each cell represents the association between a category purchase in period T and that in $T - 1$ shown as color-coded pixels—blue, red, and white for positive, negative, and no association, respectively.

Most (but not all) within-category intertemporal effects are positive, especially for the most frequently purchased items. For independent products, such as dog treats and cat food, purchases at $T - 1$ do not affect purchases of most other products at T. In contrast, cigarette purchases typically suggest fewer purchases across categories in the next period.

In the CRBM (as in the RBM), the parameters denoting associations between hidden and visible variables representing contemporaneous category purchases are estimated explicitly. However, there are no parameters directly denoting associations between the contemporaneous crosscategory effects, which must be estimated indirectly from the associations between visible and hidden variables (see Section 2.3). The results (Figure 7) seem to confirm previous findings that products with similar purchase frequencies tend to have positive contemporaneous associations with each other (Manchanda et al. 1999). As expected, there are positive associations between complements, such milk and cookies, and negative ones between substitutes, such as cookies and donuts.

4.6. Modeling at the UPC Level

Associations at the UPC level are not only multidimensional but also, hierarchical, presenting a considerable challenge for traditional models, such as the MVP. There are associations both within and between brands, and there are also similar associations at the

Figure 6. Plot of Intertemporal Effects of a One-Week Lag

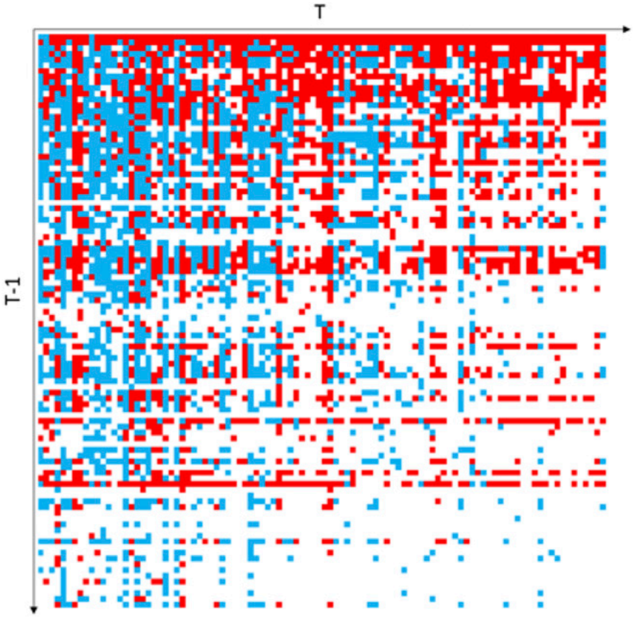
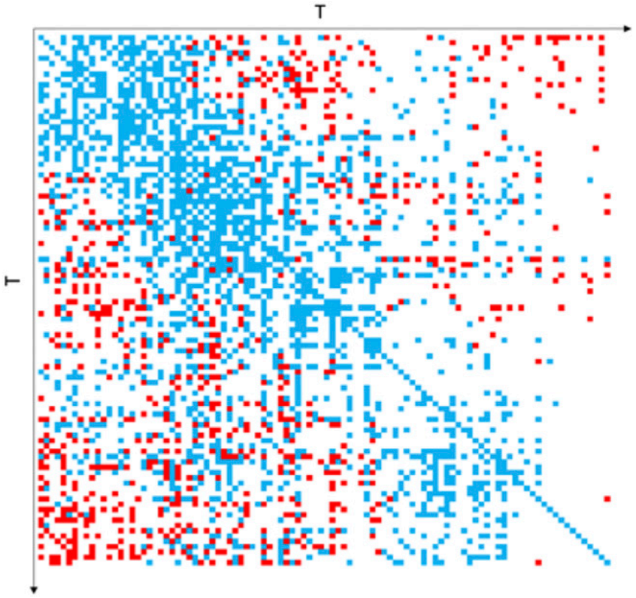


Figure 7. Plot of Contemporaneous Effects



product category level. All of the associations affect purchasing decisions at the UPC level and should be captured separately and simultaneously in a model. The CRBM is well suited to model such complex structures owing to its hierarchical Bayesian nature.

We model the UPC-level associations by welding together two CRBMs (Figure 8). UPCs are first connected to layers representing product categories and brands. A UPC is only connected to the hidden variables that represent its category and brand. A second layer then is added to model the associations among categories and brands.

We trained the proposed model using the same data set with the 200 most frequently purchased UPCs used for the analysis. There is a huge drop in terms of the purchasing frequency past around the 200th UPC (Figure 9).

Apart from the twofold RBMs that capture the contemporaneous associations, the UPC-level model is similar to the proposed model at the product category level, including purchases over 12 weeks. Price and promotion for each UPC are included directly in the model without converting them into indices as in the product category-level analysis.

We trained the model with different starting values of the elastic net penalty terms and the number of hidden variables. The general pattern is stable and tells an interesting but logical story—all associations within a brand are either positive or zero, whereas associations within product categories can be either positive or negative. Purchasing a UPC suggests that the household is more likely to purchase another UPC of the same brand (suggestive of brand loyalty). Purchasing a UPC within a certain product category, however, might either increase or decrease the purchasing probability of another UPC from the same product category.

5. Conclusion

Our proposed CRBM-based model significantly extends the marketing literature on multicategory joint purchasing decision models by expanding the number of product categories and incorporating multiple time periods. It offers all of the benefits of traditional models but with the huge advantage of efficient scale up for big data analysis. Our approach is comprehensive in that it allows for the inclusion of intertemporal multicategory effects via a clearly specified model incorporating explanatory variables without having to impose a priori constraints to ensure model tractability. Furthermore, the clear structure of the model permits meaningful interpretation of the results.

The proposed model should find application in marketing practice, especially in online or mobile marketing. The so-called “retail apocalypse” has increasingly driven businesses online, and personalized marketing has become vital for success in retail. By using the proposed model, retailers can potentially capture and predict each individual consumer’s (or household’s) complex shopping patterns with greater accuracy for personalized marketing. For example, the model can suggest an automated shopping list based on past shopping record, which can then be conveniently incorporated in applications, such as Google shopping list, benefiting both retailers and consumers. Recommendation systems built on RBM have already found success in the movie context (Salakhutdinov et al. 2007). With the abundance of consumer data, the proposed model can even be more useful by analyzing consumer decision patterns across platforms. What a consumer watches on YouTube

Figure 8. Deep Boltzmann Machine for Modeling at the UPC Level

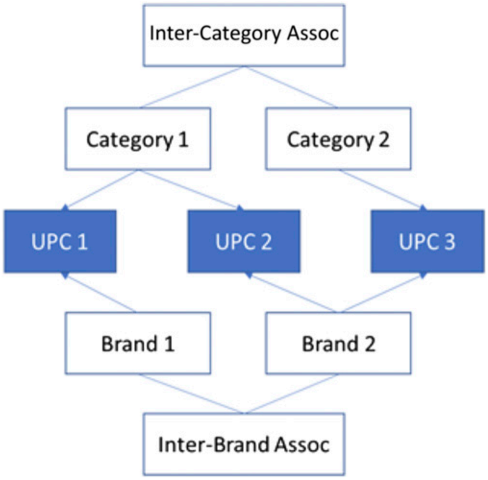
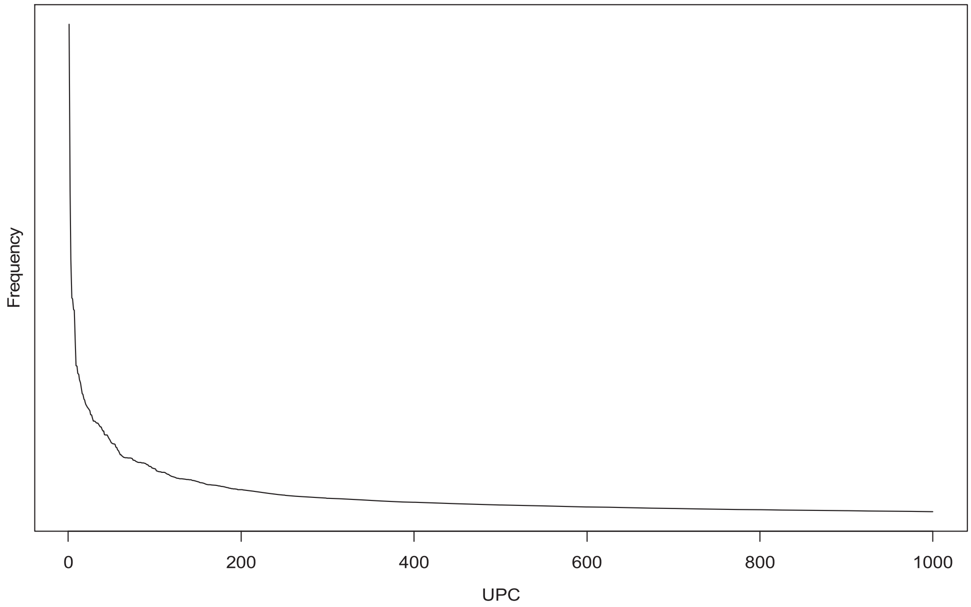


Figure 9. Purchase Frequency at the UPC Level



might be correlated with her movie choices on Netflix and the music that she listens to on Spotify. Our model is able to pick up the associations efficiently among the choices for better personalized recommendation and targeted online advertising, improving conversion rate and consumer satisfaction.

Future research should include more extensive empirical applications of the proposed model in different settings and shopping contexts and perhaps, variants of this model that enjoy the same benefits. There are, of course, several other avenues for extensions to this paper. We modeled consumer purchasing decisions as binary outcomes and did not explore the associations of purchasing decisions in other forms, such as the amount spent in each product category or the quantity purchased. The amount spent may be modeled as a continuous variable, and an extension of the RBM may be used to model the associations (Hinton and Salakhutdinov 2006). The Poisson RBM can be used to model the quantities of products purchased by the consumer. It would be interesting to see if using spending or purchase quantity as the dependent variable (especially at the brand level) leads to new consumer insights.

Acknowledgments

The authors thank the senior editor, the associate editor, and two reviewers for their insightful comments and suggestions throughout the review process.

Appendix A. Glossary of Notation

The symbols used to describe the model are summarized in the following table.

Symbol	Definition
$\boldsymbol{X}_{sit} = (x_{s1t}, x_{s2t}, \cdots, x_{s(1-1)t}, x_{sIt})$	An I-dimensional vector X_{st} representing household S's purchasing decision over I products at time t
$\boldsymbol{H}_s = (h_{s1}, h_{s2}, \cdots, h_{s(J-1)}, h_{sJ})$	A J-dimensional hidden vector H_s for household S
z	The normalizing constant of a probability function
β_i, β_j	Constant terms in probability functions for x_i and h_i
β_{ij}	Associations between product purchase decision x_i and hidden variable h_j
β_{it}	Dynamic terms in probability functions for x_{sit}
$\boldsymbol{D} = (d_1, d_2, \cdots, d_{p-1}, d_p)$	A p-dimensional vector D representing a consumer's demographic background
$\boldsymbol{C} = (c_1, c_2, \cdots, c_{r-1}, c_r)$	An R-dimensional vector C representing price and promotion for a product category
$x_{i'}^{t-m}$	Past purchasing decision for product i' at time $t - m$
β_{ip}	Direct effect of demographic dp on purchasing decisions x_{it}
$\Delta\beta_{ij}(t)$	Increment for parameter update at iteration t
ε	Learning rate
λ	Penalty term for the elastic net
α	Momentum term

Appendix B. Variational (Mean Field) Inference for the Model

According to Kullback–Leibler divergence (Salakhutdinov et al. 2013),

$$\log P(X_t;\beta) \geq \sum_{\boldsymbol{H}} \boldsymbol{P}(\boldsymbol{H}_t \mid \boldsymbol{X}_t)\mathbf{log}\,\boldsymbol{P}(\boldsymbol{X}_t, \boldsymbol{H}_t; \boldsymbol{\beta}) - \sum_{\boldsymbol{H}} \boldsymbol{P}(\boldsymbol{H}_t|\boldsymbol{X}_t)\mathbf{log}\,\boldsymbol{P}(\boldsymbol{X}_t|\boldsymbol{V}_t).$$

We can use a factorized distribution Q to approximate the posterior of H_s as follows:

$$Q(\boldsymbol{H}_s \mid \boldsymbol{X}_{st}; \mu) = \prod_J q(h_{sj} \mid \boldsymbol{X}_{st});$$
$$q(h_{sj} = 1 \mid \boldsymbol{X}_t) = \mu_{sj}.$$

Then, the lower bound becomes

$$\log P(\boldsymbol{X}_{st};\beta) \geq \sum_H \boldsymbol{Q}(\boldsymbol{H}_s \mid \boldsymbol{X}_{st}; \mu_s)\log P(\boldsymbol{X}_{st}, \boldsymbol{H}_s; \beta) - \sum_H Q(\boldsymbol{H}_s \mid \boldsymbol{X}_{st}; \mu_s)\log Q(\boldsymbol{H}_s \mid \boldsymbol{X}_{st}; \mu_s).$$

To maximize the lower bound, we set the derivative to zero, and then,

$$\mu_{sj} = \frac{1}{1+\exp(-\Sigma_I \beta_{ji} x_{sit} - \beta_j)}.$$

To predict $\boldsymbol{X}_{st}$, we set $h_{sj} = \mu_{sj}$ and estimate the expected value $E(\boldsymbol{X}_{st})$ by solving simultaneously the above equation and the conditional probability

$$E(x_{sit}) = P(x_{sit} = 1 \mid \boldsymbol{H}_s) = \frac{1}{1+\exp(-\beta_i - \Sigma_J \beta_{ji} h_{sj})}.$$

We compared the prediction results from variational inference with those of sampling methods. Both methods have very similar performance in terms of prediction accuracy.

Appendix C. Training CRBM with Elastic Net

For clarity, we use β to represent any of the parameters in the model, because they are estimated by following the same training algorithm. We use X to represent the training data, which can be either the entire training data or a small batch of them. To train the model with L1 and L2 regularization using stochastic gradient descent, we first start with the unconstrained objective function:

$$\min_\beta \{-\log(P(X;\beta))\}.$$

The objective function is the negative likelihood of the RBM or the CRBM. We update just one parameter at each iteration. When training the model with the gradient-based method, we attempt to search iteratively for the optimal value. During each step, we look for the next optimal value around a point β'. The initial parameter value can be random. We first perform a Taylor series expansion around the starting point β':

$$\min_\beta \left\{ -\log(P(X;\beta')) - \nabla\log(P(X;\beta'))(\beta - \beta') - \frac{\nabla^2 \log(P(X;\beta'))}{2}(\beta - \beta')^2 \right\}.$$

If we replace $\nabla^2 \log(P(X;\beta'))$ with t, the objective function becomes

$$\min_\beta \left\{ \frac{1}{2t}(\beta - (\beta' + t\nabla\log(P(X;\beta')))^2 \right\},$$

where the term $(\beta' + t\nabla\log(P(X;\beta'))$ performs a one-step gradient descent with step size t. We then can add both regularization terms to the objective function:

$$\min_\beta \left\{ \frac{1}{2t}\left(\beta - (\beta' + t\nabla\log(P(X;\beta')))^2 + \lambda(\alpha|\beta| + \frac{(1-\alpha)\beta^2}{2}\right)\right\}.$$

Because the L1 regularization term is differentiable, the only issue with the optimization problem here is the L1 regularization term:

$$\min_\beta \left\{ \frac{1}{2t}(\beta - (\beta' + t\nabla\log(P(X;\beta')))^2 + \lambda\alpha|\beta| \right\}.$$

The next optimal value around β' that minimizes the objective function is as follows, which is called the soft-thresholding operator:

$$S_{\lambda t}(\beta) = \begin{cases} \beta' + t\nabla\log(P(X;\beta') - \lambda\alpha t & \text{if } \beta' + t\nabla\log(P(X;\beta') > \lambda\alpha t \\ \beta' + t\nabla\log(P(X;\beta') + \lambda\alpha t & \text{if } \beta' + t\nabla\log(P(X;\beta') < -\lambda\alpha t \\ 0 & \text{if } -\lambda\alpha t \leq \beta' + t\nabla\log(P(X;\beta') \\ & \quad \leq \lambda\alpha t. \end{cases}$$

After finding the optimal value for the parameter, we add the second penalty term, the L2 regularization, to the objective function. The final optimal value for this iteration is

$$\beta^* = \frac{S_{\lambda t}(\beta)}{1+\lambda(1-\alpha)t}.$$

We also include a momentum term for each iteration to accelerate the training process.

Tuning parameters, such as λ, t, and α, need to be determined before model training. λ determines the size of the penalty. If λ is too large, the elastic net will push all parameter values toward zero. If λ is too small, the regularization terms will have little or no effect on the parameter values. Although it is challenging to find the optimal value for λ, we can define an optimal range for λ, start the model training with a large λ, and decrease λ after each run through the data (Hastie et al. 2016). The parameter estimates may not reach the optimal value if t is too big. If t is too small, the convergence rate could be slow, and the estimates might stick in local optimal. We can estimate the magnitude of the parameter value and choose a step size that is appropriate for estimates of that magnitude. For RBM training, the recommended step size t can be found in the tutorial by Hinton (2010). α determines the weight put on L1 and L2 regularizations. It takes on a value within the range [0, 1]. If α is one, then the elastic net just becomes L1 regularization. If α is zero, it is L2 regularization. More weight should be put on L1 regularization if the majority of the parameters in the model are insignificant.

Endnotes

[1] If there are n product categories, the correlation matrix grows at the rate of $O(n^2)$. If we also include the crosscategory intertemporal effects over t time periods, the number of associations that we need to model is $O(n^2t^2)$, which further increases the running time for model estimation.

[2] MCMC requires a large number of samples for relatively accurate estimation, the time to convergence is not guaranteed, and it is difficult to run in parallel. Although some methods have been proposed for running MCMC in parallel, these are mostly approximations with limited accuracy.

[3] Of course, the assumption of latent causes does not imply that we are able to uniquely identity them.

[4] Because our data are at the household level (as is typically the case), we henceforth use a "household" rather than a "consumer" as the buying unit.

[5] We can still use the basic RBM structure to model both intertemporal and contemporaneous effects—by connecting all visible variables (past and current purchasing decisions) to the hidden variables. The latter captures the associations among visible variables indirectly, including both intertemporal effects and contemporaneous effects. However, this is an incorrect model specification.

[6] These methods were originally proposed for predictive variable selection in linear models and then extended to a wide variety of statistical models, including conventional neural networks.

[7] Although this is still a large number of parameters, these are also regularized by the L2 penalty.

[8] The usual practice is to set the number of hidden variable to slightly below the number of visible variables (100 in our case). Dropout training (which we use) essentially prevents overfitting from an initial specification of too many hidden variables.

[9] For example, we conducted a coarse-grid search for the optimal learning rate with a 0.0002 interval and found that the optimal learning rate would reduce the average error by 0.00002.

[10] We also tested the predictive performance of the CRBM against a null model that excludes crosscategory intertemporal effects. Specifically, 12 weeks' past purchases and 100 categories were included in the comparison between the CRBM and a regression model that ignores crosscategory purchases in predicting the probability of purchase in a given week. The CRBM outperforms this null model.

References

Blei DM, Kucukelbir A, McAuliffe JD (2017) Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* 112(518):859–877.

Bottou L (2010) Large-scale machine learning with stochastic gradient descent. *Proc. COMPSTAT'2010* (Physica-Verlag, Berlin, Heidelberg, Germany), 177–186.

Boztug Y, Hildebrandt L (2008) Modeling joint purchases with a multivariate MNL approach. *Schmalenbach Bus. Rev.* 60(4): 400–422.

Burnap A, Ren Y, Lee H, Gonzalez R, Papalambros PY (2014) Improving preference prediction accuracy with feature learning. *Proc. ASME 2014 Internat. Design Engrg. Tech. Conf. Comput. Inform. Engrg. Conf.* (American Society of Mechanical Engineers, New York), V02AT03A012.

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A (2017) Stan: A probabilistic programming language. *J. Statist. Software* 76(1):1–32.

Cotter A, Shamir O, Srebro N, Sridharan K (2011) Better mini-batch algorithms via accelerated gradient methods. *Adv. Neural Inform. Processing Systems* 24:1647–1655.

Duvvuri SD, Gruca TS (2010) A Bayesian multi-level factor analytic model of consumer price sensitivities across categories. *Psychometrika* 75(3):558–578.

Duvvuri SD, Ansari A, Gupta TS (2007) Consumers' price sensitivities across complementary categories. *Management Sci.* 53(12): 1933–1945.

Edwards YD, Allenby GM (2003) Multivariate analysis of multiple response data. *J. Marketing Res.* 40(3):321–334.

Fader PS, Hardie BG, Lee KL (2005) "Counting your customers" the easy way: An alternative to the Pareto/NBD model. *Marketing Sci.* 24(2):275–284.

Friedman J, Hastie T, Tibshirani R (2001) *The Elements of Statistical Learning*, Springer Series in Statistics (Springer, Berlin).

Girolami M, Rogers S (2006) Variational Bayesian multinomial probit regression with Gaussian process priors. *Neural Comput.* 18(8): 1790–1817.

Hansen K, Singh V, Chintagunta P (2006) Understanding store-brand purchase behavior across categories. *Marketing Sci.* 25(1):75–90.

Hastie T, Tibishirani R, Wainwright M (2016) *Statistical Learning with Sparsity* (CRC Press, Boca Raton, FL).

Hinton G (2010) A practical guide to training restricted Boltzmann machines. *Momentum* 9(1):926.

Hinton G (2002) Training products of experts by minimizing contrastive divergence. *Neural Comput.* 14(8):1771–1800.

Hinton G, Salakhutdinov R (2006) Reducing the dimensionality of data with neural networks. *Science* 313(5786):504–507.

Hruschka H (2014) Analyzing market baskets by restricted Boltzmann machines. *OR Spectrum* 36(1):209–228.

Keenan DM (1982) A time series analysis of binary data. *J. Amer. Statist. Assoc.* 77(380):816–821.

Kucukelbir A, Tran D, Ranganath R, Gelman A, Blei DM (2017) Automatic differentiation variational inference. *J. Machine Learn. Res.* 18(1):430–474.

Lawrence ND (2004) Gaussian process latent variable models for visualisation of high dimensional data. *Adv. Neural Inform. Processing Systems* 16(3):329–336.

Makhzani A, Frey BJ (2015) Winner-take-all autoencoders. *Adv. Neural Inform. Processing Systems* 28:2791–2799.

Manchanda P, Ansari A, Gupta S (1999) The "shopping basket": A model for multicategory purchase incidence decisions. *Marketing Sci.* 18(2):95–114.

Mehta N (2007) Investigating consumers' purchase incidence and brand choice decisions across multiple product categories: A theoretical and empirical analysis. *Marketing Sci.* 26(2): 196–217.

Miles S (2013) 7 strategies for boosting digital coupon conversions. Retrieved March 3, 2014, http://streetfightmag.com/2013/10/03/7-strategies-for-boosting-digital-coupon-conversions/.

Pearl J (2014) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Elsevier, Amsterdam).

Russell GJ, Petersen A (2000) Analysis of cross category dependence in market basket selection. *J. Retailing* 76(3):367–392.

Salakhutdinov R, Larochelle H (2010) Efficient learning of deep Boltzmann machines. *Proc. 13th Internat. Conf. Artificial Intelligence Statist.*, vol. 9 (Microtome Publishing, Brookline, MA), 693–700.

Salakhutdinov R, Mnih A, Hinton G (2007) Restricted Boltzmann machines for collaborative filtering. *Proc. 24th Internat. Conf. Machine Learn.* (ACM, New York), 791–798.

Salakhutdinov R, Tenenbaum JB, Torralba A (2013) Learning with hierarchical-deep models. *IEEE Trans. Pattern Anal. Machine Intelligence* 35(8):1958–1971.

Seetharaman PB, Chib S, Ainslie A, Boatwright P, Chan T, Gupta S, Mehta N, Rao V, Strijnev A (2005) Models of multi-category choice behavior. *Marketing Lett.* 16(3–4):239–254.

Shai S, Ambuj T (2011) Stochastic method for L1 regularized loss minimization. *J. Machine Learn. Res.* 12(June):1865–1892.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: A simple way to prevent neural networks from overfitting. *J. Machine Learn. Res.* 15(1):1929–1958.

Shocker AD, Bayus BL, Kim N (2004) Product complements and substitutes in the real world: The relevance of "other products." *J. Marketing* 68(1):28–40.

Talhouk A, Doucet A, Murphy K (2012) Efficient Bayesian inference for multivariate probit models with sparse inverse correlation matrices. *J. Comput. Graphical Statist.* 21(3):739–757.

Taylor GW, Hinton G (2009) Factored conditional restricted Boltzmann machines for modeling motion style. *Proc. 26th Annual Internat. Conf. Machine Learn.* (ACM, New York), 1025–1032.

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B. Statist. Methodology* 67(2):301–320.

CORRECTION

In this article, "Using Conditional Restricted Boltzmann Machines to Model Complex Consumer Shopping Patterns" by Feihong Xia, Rabikar Chatterjee, and Jerrold H. May (first published in *Articles in Advance*, July 12, 2019, *Marketing Science*, DOI: 10.1287/mksc.2019.1162), the authors note that the performance of the basic restricted Boltzmann machine versus multivariate probit is consistent with the findings of Hruschka (2014). While the cited paper references both the multivariate probit and multinomial logit models, it compares the restricted Boltzmann machine to the multinomial logit model, not the multivariate probit model as stated in this paper. The authors apologize for this error.