Proceedings of the ASME 2014 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference
IDETC/CIE 2014
August 17-20, 2014, New York, United States

DETC2014-35440

IMPROVING PREFERENCE PREDICTION ACCURACY WITH FEATURE LEARNING

Alex Burnap*
Design Science
University of Michigan
Email: aburnap@umich.edu

Yi Ren*
Mechanical Engineering
University of Michigan
Email: yiren@umich.edu

Honglak Lee
Computer Science and Engineering
University of Michigan
Email: honglak@eecs.umich.edu

Richard Gonzalez
Psychology
University of Michigan
Email: gonzo@umich.edu

Panos Y. Papalambros
Mechanical Engineering
University of Michigan
Email: pyp@umich.edu

ABSTRACT

Motivated by continued interest within the design community to model design preferences, this paper investigates the question of predicting preferences with particular application to consumer purchase behavior: How can we obtain high prediction accuracy in a consumer preference model using market purchase data? To this end, we employ sparse coding and sparse restricted Boltzmann machines, recent methods from machine learning, to transform the original market data into a sparse and high-dimensional representation. We show that these 'feature learning' techniques, which are independent from the preference model itself (e.g., logit model), can complement existing efforts towards high-accuracy preference prediction. Using actual passenger car market data, we achieve significant improvement in prediction accuracy on a binary preference task by properly transforming the original consumer variables and passenger car variables to a sparse and high-dimensional representation.

1 Introduction

Within the design community, preference modeling has been investigated extensively in the area of design for market systems, synthesizing engineering and marketing models to improve decision making in product design [1, 2]. A large number of studies have demonstrated the importance of consumer preference in engineering design in applications, such as vehicle engines [3], packaging [4], silhouettes [5], bathroom weight scales [6], and hand saws [7].

A persisting challenge in modeling consumer preferences is to improve prediction accuracy based on a given set of stated or revealed preference observations. The design and marketing research communities have spent significant effort primarily in three directions: (1) Developing sophisticated statistical models to capture better the heterogeneous and stochastic nature of consumer preferences; examples include mixed and nested logit models [8, 9], consideration sets [10], and kernel-based methods [11, 12]; (2) creating adaptive questionnaires to obtain stated information more efficiently using a variety of active learning methods [13, 14]; and (3) ways of acquiring useful "covariates" that better explain the acquired observations [15].

In this study, we take a different approach: taking cues from recent advances in machine learning, we "learn" *features* from the existing data that are more representative of the consumer's decision-making process. These features can be abstract data entities and, preferably but not necessarily, have some interpretable value. Such *feature learning* is different from simply finding new data as these features are functions of the original data. A key point is that feature learning uses the same set of data and the same preference prediction algorithm (e.g., logit model

*Both authors contributed equally.

or support vector machine), yet performs prediction within the feature space instead of the original data space to achieve significant improvements in prediction accuracy [16]. Moreover, these methods do not require domain-specific knowledge and thus can be adapted to various types of data, i.e., we do not need to have physical insight into the data-generating process underlying the observed data. Accordingly, feature learning methods have been successful for various learning tasks, including image classification [17–19], speech recognition [20–22], information retrieval [23, 24], and natural language processing [25, 26].

The contribution of this paper is in being a first investigation on the use of learning features to improve consumer preference prediction for a heterogeneous market data set, i.e., data consisting of a variety of units such as real-valued, binary, and categorical. We examine two feature learning methods, sparse coding [27, 28] and sparse restricted Boltzmann machines (RBM) [29,30], and apply them to the problem of predicting passenger car purchase behavior using real car market data. Results indicate 10% or more improvement in prediction accuracy relative to a baseline logit model approach commonly used in the marketing and design communities.

The paper is structured as follows: Section 2 briefly covers background information on feature learning advances in the machine learning community. Section 3 sets up the preference prediction task, namely, predicting which passenger car a consumer will purchase as a binary classification problem. Section 4 details the two feature learning algorithms, as well as their tailoring to suit market data. Section 5 discusses the methods for data processing and the experimental setup of the preference prediction task. Section 6 discusses results and directions for further study. Section 7 gives the conclusion to this work.

2 Related Work

Feature learning methods can capture factors implicit in the original data by encoding the data in a new representation. With this new representation, we may use the same supervised model (e.g., logit model) as before to achieve higher predictive performance. The general idea is to: (1) map the original data points to a higher dimensionality space made up of features, in which the dimensionality of the feature space is a factor larger than the dimensionality of the original data space as defined by an overcompleteness factor γ; and (2) induce "sparsity" within this higher-dimensional space, such that only a proportion of the new space is "activated" by a given data point as governed by a sparsity penalty β. Both the mapping and the new representation in that space are determined by minimizing some objective function describing the reconstruction error between the original variables and their new feature representation, as well as the sparsity penalty on the activation of features. More technical details will be given in Section 4.

The first method we examined is sparse coding. Sparse coding is similar to principal component analysis (PCA) in that it embeds the original data points in a new basis as shown in Figure 1(a). However, unlike PCA, only the "decoding" from the feature representation to the original data is linear, whereas the "encoding" process is done using L1-norm sparsity penalization on a linear reconstruction error, resulting in coefficients that are real-valued and sparse [31]. Since the basis vectors are no longer orthogonal, sparse coding can be applied to learning overcomplete basis sets, in which the number of bases is greater than the input dimension. The advantage of having an overcomplete basis is that it can effectively capture the underlying statistical distribution of the data, leading to better coding efficiency. Readers are referred to [32] for more discussion on the advantages of overcomplete representations.

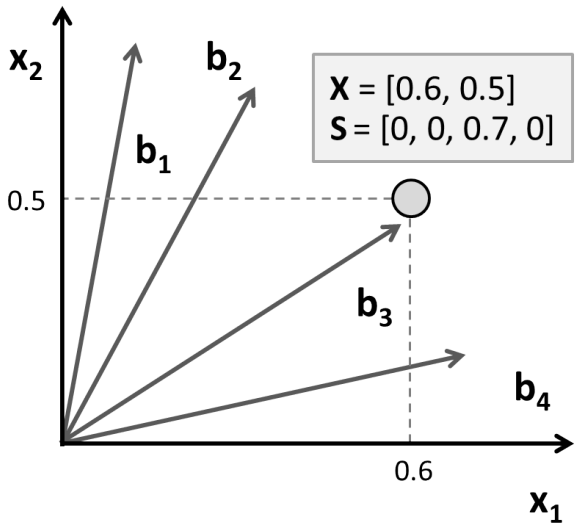

FIGURE 1. The concept of (a) sparse coding and (b) restricted Boltzmann machine. The original data is represented by $[x_1, x_2]$, while the new features of the same data are represented by $[s_1, s_2, s_3, s_4]$ (informally speaking, coefficients for the basis vectors b_1, b_2, b_3, b_4). Note the sparsity of the new feature representation; for example, with sparse coding, although the original data has only non-zero coefficients in the original basis x_1, x_2, the new feature representation has most coefficients equal to zero.

The second method we examined is the sparse restricted Boltzmann machine (RBM) [29, 33]. This feature learning technique is a special case of the more general Boltzmann machine, an undirected graphical model in which the energy associated with a state space defines the probability of finding the system in that state [29]. In the RBM, each state is determined by both visible and hidden nodes, where each node corresponds to a random variable. The hidden nodes, which are stochastic binary, are added to increase the expressiveness of the system model, and are the features that capture the new higher-dimensional rep-

TABLE 1. Consumer variables and their unit types

Consumer Variable	Unit Type	Consumer Variable	Unit Type
Income level	Categorical	Number of older children	Categorical
Age	Real	Number of children	Categorical
Gender	Binary	Resident location	Categorical
Number in house	Categorical	Education level	Categorical
Number of young children	Categorical	Income-State Income ratio	Real
Numer of medium children	Categorical	Income-State Living cost ratio	Real

TABLE 2. Passenger car variables and their unit types

Car Variable	Unit Type	Car Variable	Unit Type
Invoice	Real	MSRP	Real
Curb Weight	Real	Net HP	Real
MPG	Real	Length	Real
Width	Real	Height (mean)	Real
Wheelbase	Real	Final drive ratio	Real
AWD/4WD	Binary	Turbo	Binary
Supercharger	Binary	Class	Categorical
Make	Categorical	Pass. Capacity	Categorical
Engine size	Real	Hybrid	Binary
Luxury	Binary		

resentation. The "restricted" portion of the RBM refers to the restriction on visible-visible connection and hidden-hidden connection as shown in Figure 1(b).

3 Preference Prediction as Binary Classification

We consider the following binary classification problem: Given a consumer represented by a set of heterogeneous variables and two passenger cars, each represented by a set of car variables, which passenger car will the consumer purchase? We use an actual database of consumers and their passenger car purchase decisions as detailed below [34].

3.1 Car purchase data from 2006

In order to have a data set with both consumer variables as well as passenger vehicle variables, we synthesized the Maritz car purchase survey from 2006 [34], the Chrome car specification database [35], and the 2006 estimated US state income and living cost data from U.S. Census Bureau [36]. The combined database contains a list of purchase records, with each row describing the consumer's variables, e.g., income level, age, gender, and the variables of the corresponding car he or she bought.

From this original data set, we focus only on the consumer group who bought passenger cars of classes between mini-compact and large cars, excluding purchase records for station wagons, trucks, minivans, and utility vehicles. In addition, records for consumers who did not consider other cars before their purchases were removed, as well those records for which consumers purchased cars for another party. Finally, we removed the passenger car variables regarding whether the passenger car used diesel or gasoline fuel and whether it had an automatic or manual transmission, due to lack of information during the synthesis of the three data sets.

In the end, the database contained 212 unique passenger car models bought by 6556 unique consumers. The full list of consumer variables and passenger car variables can be found in Tables 1 and 2. Note that the variables in these tables are grouped into three formats: Real, binary and categorical, based on the nature of the variables. More discussion on the impact of correctly specifying data formats will be presented in Section 4.

3.2 Choice set generation

We next converted the database of purchase records into a choice set by generating a four sets of pairwise comparisons for each consumer, with the chosen car in each pair being the car that was actually purchased. While previous studies have shown the impact on prediction performance given different generations of choice sets (see [37] for example) we will show that the proposed feature learning method improves prediction performance consistently for arbitrarily generated choice sets. This work is therefore complementary to studies on developing appropriate choice set generation schemes, such as [38].

We thus assume that every consumer considers five alternatives before settling on a final purchase choice. These four cars besides the purchased one are selected according to their relative frequency in the entire data set, i.e., a consumer is more likely to consider a Honda Accord or Toyota Camry over a Volvo S40 since the former cars have a larger market share than the latter.

3.3 Training and testing data

Following conventions in marketing research [10, 12], three-fourths of the pairwise choices from each consumer are randomly chosen and used for training and validating the predictive model (model selection), while the rest were used for testing (model assessment). Three random splits of training and testing data are used to bootstrap the average prediction performance.

3.4 Bilinear utility

A pairwise comparison record for consumer r with cars p and q consists of the consumer's original variables $\mathbf{x}_u^{(r)}$ for $r \in \{1, \ldots, 6556\}$ and specifications of the two cars, $\mathbf{x}_c^{(p)}$ and $\mathbf{x}_c^{(q)}$ for $p, q \in \{1, \ldots, 212\}$. We adopt the conventions of utility theory for the measure of consumer preference for a given product. In particular, this study assumed a bilinear utility model for consumer r and car p:

$$U_{rp} = \left[\text{vec}\left(\mathbf{x}_u^{(r)} \otimes \mathbf{x}_c^{(p)}\right), \mathbf{x}_c^{(p)}\right]^T \mathbf{\Omega}, \tag{1}$$

where $\otimes$ is an outer product for vectors, vec $(\cdot)$ is vectorization of a matrix, $[\cdot, \cdot]$ is concatenation of vectors, and $\mathbf{\Omega}$ is the part-worth vector. While we do not consider main effects from consumer variables in this formulation of the utility, we discuss potentially better utility models in Section 6.

3.5 Classification models

The classification model refers to the predictive algorithm used to capture the relationship between a consumer's purchase decision, variables describing the consumer, and variables describing the car. While the choice of classification model is not the focus of this paper, we pilot tested popularly used models including L1 and L2 logit model, naïve Bayes, L1 and L2 linear and kernelized support vector machine, and random forests.

Based on these pilot results, we chose the L2 logit model due to its widespread use in the design and marketing communities. In particular, we used the primal form of the logit model and stochastic gradient descent for parameter optimization. Equation (2) captures how the logit model describes the probabilistic relationship between consumer r's preference for either car p or car q as a function of their associated utilities given by Equation (1). Note that η are Gumbel-distributed random variables accounting for noise over the underlying utility of the consumer r's preference for either car p or car q.

$$P\left(U_{rp} + \eta_{rp} > U_{rq} + \eta_{rq}\right) = \frac{e^{U_{rp}}}{e^{U_{rp}} + e^{U_{rq}}} \tag{2}$$

4 Feature Learning

We present two feature learning methods as introduced in Section 2, and discuss their extensions to better fit the market data described in Section 3.

4.1 Sparse coding

For a set of input vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)} \in \mathbb{R}^K$, sparse coding finds basis vectors $\mathbf{b}_1, \ldots, \mathbf{b}_N \in \mathbb{R}^K$ and corresponding sparse vectors of weights (or "activations") $\mathbf{h}^{(1)}, \ldots, \mathbf{h}^{(M)} \in \mathbb{R}^N$ such that $\mathbf{x}^{(m)} \approx \sum_n \mathbf{b}_n h_n^{(m)}$ [28]. These weights act as the features for our new transformed representation from the original data space. Further, the reconstruction error $\mathbf{x}^{(m)} - \sum_n \mathbf{b}_n h_n^{(m)}$ is assumed to be Gaussian distributed with zero means and covariance $\sigma^2\mathbf{I}$. Here we take $\sigma = 1$ and apply this Gaussian assumption to both consumer variables and car variables. The log-likelihood of the mth observed data (e.g., variables of a car or variables of a consumer) can be expressed as $-\frac{1}{2\sigma^2}||\mathbf{x}^{(m)} - \sum_{n=1}^N \mathbf{b}_n h_n^{(m)}||^2$. Further, we impose sparsity on the activations $\mathbf{h}^{(m)}, \ldots, \mathbf{h}^{(M)}$ by applying an epsilon-L_1 penalty $\phi(h_n) = (h_n^2 + \epsilon)^{\frac{1}{2}}$, with $\epsilon = 10^{-6}$.

Based on these settings, feature learning using sparse coding can be formulated into the following optimization problem where the optimal basis and weights can be learned:

$$\begin{aligned} \min_{\{\mathbf{b}_n\}, \{\mathbf{h}^{(m)}\}} \quad & \sum_{m=1}^M \frac{1}{2}||\mathbf{x}^{(m)} - \sum_n \mathbf{b}_n h_n^{(m)}||^2 \\ & + \beta \sum_{m=1}^M \sum_{n=1}^N \phi(h_n^{(m)}) \\ \text{subject to} \quad & ||\mathbf{b}_n||^2 \leq 1, \ \forall n = 1, \ldots, N. \end{aligned} \tag{3}$$

where β is the sparsity penalty that controls the proportion of zero-valued features learned by sparse coding.

As Tables 1 and 2 detail, both consumer and car data contain variables of three formats: real-valued, binary, and categorical. The Gaussian assumption is less reasonable for the reconstruction errors of the latter two forms of data than for the real-valued ones. Therefore, we refine the assumption to let the errors of binary features to be Bernoulli distributed and those of categorical features to be categorically distributed. For example, in consumer profiles, the error for "age" is treated as Gaussian, "gender" as Bernoulli, and "region" as categorical.

Under this refined assumption, consider that we have K_G Gaussian, K_B Bernoulli and K_C categorical variables. We will now decompose the basis $\mathbf{B} := [\mathbf{b}_1, \ldots, \mathbf{b}_N]$ by rows to be $\mathbf{B}^T = [\mathbf{B}_G^T, \mathbf{B}_B^T, \mathbf{B}_C^T]$ where $\mathbf{B}_G$, $\mathbf{B}_B$ and $\mathbf{B}_C$ are bases corresponding to features with Gaussian, Bernoulli and the categorical input variables, respectively. Further, denote D_k as the number of categories for the kth categorical variable. Feature learning can than be formulated as the following optimization problem.

$$\begin{aligned} \min_{\mathbf{B}, \{\mathbf{h}^{(m)}\}} \quad & f := \sum_{m=1}^M -{\mathbf{h}^{(m)}}^T \mathbf{B}^T \mathbf{x}^{(m)} \\ & + {\mathbf{h}^{(m)}}^T \mathbf{B}_G^T \mathbf{B}_G \mathbf{h}^{(m)}/2 \\ & + \sum_{k=1}^{K_B} \log\left(1 + \exp((\mathbf{B}_B \mathbf{h}^{(m)})_k)\right) \\ & + \sum_{k=1}^{K_C} \log\left(1 + \sum_{d=1}^{D_k} \exp\left((\mathbf{B}_{C,k} \mathbf{h}^{(m)})_d\right)\right) \\ & + \beta \sum_{m=1}^M \sum_{n=1}^N \phi(h_n^{(m)}) \\ \text{subject to} \quad & ||\mathbf{b}_n||^2 \leq c, \ \forall n = 1, \ldots, N \end{aligned} \tag{4}$$

Training To train this model, the optimization procedure in Lee et al. [39] is adopted. We start by initializing randomized bases $\mathbf{B}$ and finding corresponding optimal $\mathbf{h}^{(1)}, \ldots, \mathbf{h}^{(M)}$ by iteratively solving quadratic approximations of the problem in Equation (4) using the conjugate gradient algorithm [28]. Then for fixed $\mathbf{h}^{(1)}, \ldots, \mathbf{h}^{(M)}$, we use a projected gradient search to optimize $\mathbf{B}$. This iteration terminates when convergence is achieved.

4.2 Restricted Boltzmann machine

The restricted Boltzmann machine (RBM) is an energy-based model where each energy state is defined by a layer of K visible nodes corresponding to an input datum $\mathbf{x}$, a layer of N hidden nodes denoted as $\mathbf{h}$, a weight matrix $\mathbf{W}$ connecting the visible and hidden nodes, and biases for both the hidden nodes and visible nodes, $\mathbf{a}$ and $\mathbf{b}$ respectively. From our original data set $\mathbf{x}^{(1)} \ldots \mathbf{x}^{(M)} \in \mathbb{R}^K$, the RBM provides a high-dimensional feature representation $\mathbf{h}^{(1)} \ldots \mathbf{h}^{(M)} \in \mathbb{R}^N$ that may be more discriminative that the original data representation for the preference learning task described in Section 3.

The probability of a state with energy $E(\mathbf{x}, \mathbf{h}; \theta)$, where θ are the energy functions parameters, is defined by the Boltzmann distribution.

$$P(\mathbf{x}, \mathbf{h}) = \frac{e^{-E(\mathbf{x},\mathbf{h};\theta)}}{\sum_{\mathbf{x}} \sum_{\mathbf{h}} e^{-E(\mathbf{x},\mathbf{h};\theta)}} \tag{5}$$

The "restriction" on the RBM is to disallow visible-visible and hidden-hidden node connections. This restriction results in conditional independence of each individual hidden unit h given the vector of inputs $\mathbf{x}$, and each visible unit v given the vector of hidden units $\mathbf{h}$.

$$P(\mathbf{h}|\mathbf{x}) = \prod_{n=1}^{N} P(h_n|\mathbf{x}) \tag{6}$$

$$P(\mathbf{x}|\mathbf{h}) = \prod_{k=1}^{K} P(x_k|\mathbf{h}) \tag{7}$$

Though we assume the hidden nodes are binary, i.e., $h_n \in \{0, 1\}$, the input nodes require K_G Gaussian, K_B binary, or K_C categorical distributions to match the respective heterogeneous variables types described in Table 1 and Table 2. These distributions are explicitly defined by the energy function over the joint distribution of visible and hidden nodes.

Real-valued random variables (e.g., vehicle curb weight) are modeled using the Gaussian density. The energy function for Gaussian inputs and binary hidden nodes is:

$$\begin{aligned} E_G(\mathbf{x}, \mathbf{h}; \theta) = &- \sum_{k=1}^{K_G} \sum_{n=1}^{N} h_n w_{nk} x_k \\ &+ \frac{1}{2} \sum_{k=1}^{K_G} (x_k - b_k)^2 - \sum_{n=1}^{N} a_n h_n \end{aligned} \tag{8}$$

where the variance term is clamped to unity under the assumption that the input data are standardized.

Binary random variables (e.g., gender) are modeled using the Bernoulli density. The energy function for Bernoulli nodes in both the input layer and hidden layer is:

$$\begin{aligned} E_B(\mathbf{x}, \mathbf{h}; \theta) = &- \sum_{k=1}^{K_B} \sum_{n=1}^{N} h_n w_{nk} x_k \\ &- \sum_{k=1}^{K_B} x_k b_k - \sum_{n=1}^{N} a_n h_n \end{aligned} \tag{9}$$

Categorical random variables (e.g., vehicle manufacturer) are modeled using the categorical density. The energy function for categorical inputs with D_k classes for k-th categorical input variable (e.g., Toyota, General Motors, etc.) is given by:

$$\begin{aligned} E_C(\mathbf{x}, \mathbf{h}; \theta) = &- \sum_{k=1}^{K_C} \sum_{n=1}^{N} \sum_{d=1}^{D_k} h_n w_{nkd} \delta_{kd} x_{kd} \\ &- \sum_{k=1}^{K_C} \sum_{d=1}^{D_k} \delta_{kd} x_{kd} b_{kd} - \sum_{n=1}^{N} a_n h_n \end{aligned} \tag{10}$$

where $\delta_{kd} = 1$ if $x_{kd} = 1$ and 0 otherwise.

The conditional density for a single binary hidden unit given the combined K_G Gaussian, K_B binary, and K_C categorical input variables is then:

$$\sigma(a_n + \sum_{k=1}^{K_G} w_{nk} x_k + \sum_{k=1}^{K_B} w_{nk} x_k + \sum_{k=1}^{K_C} \sum_{d=1}^{D_k} w_{nk} \delta_{kd} x_k) \tag{11}$$

where $\sigma(s) = \frac{1}{1+\exp(-s)}$ is a sigmoid function.

For an input data point $\mathbf{x}^{(m)}$, the new high-dimensional representation $\mathbf{h}^{(m)}$ is given by the "activations" of the hidden nodes.

$$\mathbf{h}^{(m)} = \mathbb{E}[\mathbf{h}|\mathbf{x}^{(m)}] \tag{12}$$

$$= [P(h_1 = 1|\mathbf{x}, \theta), \ldots, P(h_N = 1|\mathbf{x}, \theta)] \tag{13}$$

Training To train the model, we optimize the weight and bias parameters $\theta = \{\mathbf{W}, \mathbf{b}, \mathbf{a}\}$ by minimizing the negative log-likelihood of the data $\left\{\mathbf{x}^{(1)} \ldots \mathbf{x}^{(M)}\right\}$ using gradient descent. The gradient of the log-likelihood is:

$$\begin{aligned} \frac{\partial}{\partial \theta} \sum_{m=1}^{M} \log P\left(\mathbf{x}^{(m)}\right) &= \frac{\partial}{\partial \theta} \sum_{m=1}^{M} \log \sum_{\mathbf{h}} P\left(\mathbf{x}^{(m)}, \mathbf{h}\right) \\ &= \frac{\partial}{\partial \theta} \sum_{m=1}^{M} \log \sum_{\mathbf{h}} \frac{e^{-E(\mathbf{x}^{(m)},\mathbf{h})}}{\sum_{\mathbf{x},\mathbf{h}} e^{-E(\mathbf{x}^{(m)},\mathbf{h})}} \\ &= \sum_{m=1}^{M} \mathbb{E}_{\mathbf{h}|\mathbf{x}^{(m)}} \left[\frac{\partial}{\partial \theta} E\left(\mathbf{x}^{(m)}, \mathbf{h}\right)\right] \\ &\quad - \mathbb{E}_{\mathbf{h},\mathbf{x}} \left[\frac{\partial}{\partial \theta} E(\mathbf{x}, \mathbf{h})\right] \end{aligned} \tag{14}$$

The gradient is the difference of two expectations, the first of which is easy to compute since it is "clamped" at the input datum x, but the second of which requires the joint density over the entire $\mathbf{x}$ space for the model.

In practice, this second expectation is approximated using the Contrastive Divergence algorithm by Gibbs sampling the hidden nodes given the visible nodes, then the visible nodes given the hidden nodes, and iterating a sufficient number of steps for the approximation [40]. During training, we induce sparsity of the hidden layer by setting a target activation p_n, fixed to 0.1, for each hidden unit h_n [33]. The overall objective to be minimized is then the negative log-likelihood from Equation (14) and a penalty on the deviation of the hidden layer from the target activation. Since the hidden layer is made up of sigmoid densities, the overall objective function is:

$$\begin{aligned} &\sum_{m=1}^{M} \log \sum_{\mathbf{h}} P\left(\mathbf{x}^{(m)}, \mathbf{h}\right) \\ &+ \beta \sum_{n=1}^{N} p_n^{(m)} \log h_n + \left(1 - p_n^{(m)}\right) \log\left(1 - h_n\right) \end{aligned} \tag{15}$$

where β is the hyperparameter trading off the sparsity penalty with the log-likelihood.

5 Experiment and Methods

In this experiment, the L2 logit model was used to predict the car purchase preferences of consumers from the data set under two cases. The first case acted as a "baseline" using only the normalized consumer and car variables to predict consumer preference. The second case used the features learned from sparse coding or sparse RBM concatenated with the original consumer and car variables to predict consumer preference.

In addition, we performed an analysis of two hyperparameters common to both sparse coding and sparse RBM on how they affected preference prediction accuracy. The first was the sparsity penalty β, found in Equation (4) and Equation (15), which controls the number of features activated for a given input datum. The second was the overcompleteness factor γ, which defines by what factor the dimensionality of the feature space is larger than the dimensionality of the original data space.

The detailed experiment flow is summarized below and illustrated in Figure 5:

1. The full data set was sampled according to the procedure in Section 3.3 to create three new data sets with each new data set made up of 75% of the original data for training and validation, and 25% of the original data set for testing. The training and validation subsets were then randomly split into training and validation subsets at a ratio of 66% to 33% into training and validation sets, for a total of three new cross validation data sets.

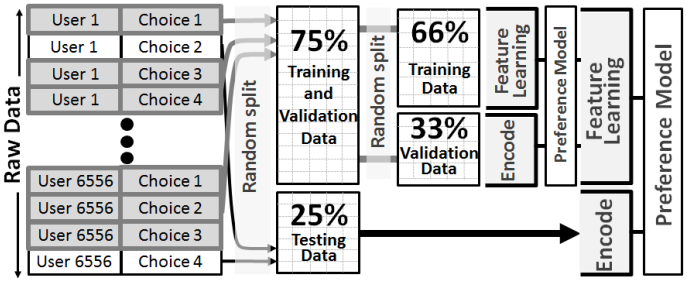

FIGURE 2. Data processing, training, validation, and testing flow.

2. The L2 logit model was trained on each training set, in which the training datum was either represented by only the original variables, or a new representation defined by the original variables concatenated with features learned using sparse coding or sparse RBM. Specifically, for some consumer variables $\mathbf{x}_u$ and corresponding learned features $\mathbf{h}_u$, we used $\mathbf{x}_{u'}^T := [\mathbf{x}_u^T, \mathbf{h}_u^T]$ to define the new representation of the consumer; likewise, for some car $\mathbf{x}_c$ and corresponding learned features $\mathbf{h}_c$, we used $\mathbf{x}_{c'}^T := [\mathbf{x}_c^T, \mathbf{h}_c^T]$ to define the new representation of the consumer. Combined with Equation (1), a single data point used for training is the difference in utilities between car p and car q for a given consumer r.

$$\left[\mathbf{x}_{u'}^{(r)} \otimes \left(\mathbf{x}_{c'}^{(p)} - \mathbf{x}_{c'}^{(q)}\right), \mathbf{x}_{c'}^{(p)} - \mathbf{x}_{c'}^{(q)}\right] \tag{16}$$

Note that the choice data generated in Section 3.2 can be considered as belonging to a single class, i.e., each data point represents one car being chosen over another. In order to properly train a one-class classifier using commonly available supervised learning algorithms for binary-class data, we follow suggestions from [41], by concatenating the training data with a copy of itself but with signs flipped for both data points and labels.
Both sparse coding and restricted Boltzmann machine were swept over a full-factorial grid of the hyperparameters γ (for overcompleteness) and β (for sparsity); with γ ranging from 2 to 5 in steps of 0.5, and β ranging from 0.4 to 1.0 in steps of 0.1. It is important to note that though the feature representations were learned on just the training set, they were used to "encode" the held out validation set during cross validation. This encoding was performed using either the sparse coding objective function from Equation (4) with the learned basis, or in the case of the RBM, inferred using Equation (11) with the learned weights and biases.
3. The cross validation prediction accuracy of the L2 logit model was assessed on the held out validation set, and the cross-validated hyperparameter set with the best preference prediction accuracy was saved. These best hyperparame-

Preference Model	Data Representation	Prediction Accuracy (std.)
L2 Logit	Baseline	71.0% (0.4%)
L2 Logit	Exponential Family Sparse Coding	76.4% (0.8%)
L2 Logit	Exponential Family Sparse RBM	**81.4%** (1.1%)

TABLE 3. Averaged preference prediction accuracy on held-out test data for the L2 logit model trained on the baseline (original variables) and feature learning representations.

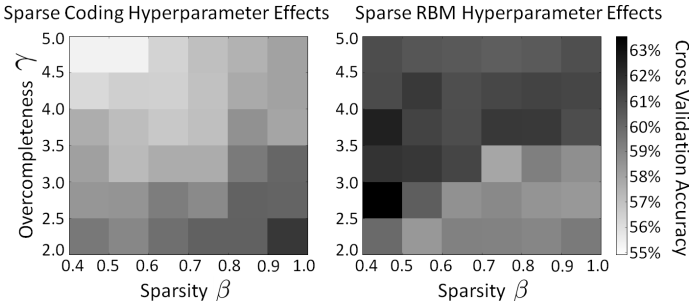

FIGURE 3. Effects of hyperparameters for sparse coding and sparse RBM with respect to the overcompleteness factor γ and sparsity penalty β.

ters, namely, the L2 regularization hyperparameter for the L2 logit model, the overcompleteness factor γ, and the sparsity penalty β, were used to refit the L2 logit model on the combined training and validation set. Likewise, both sparse coding and sparse RBM were retrained on the combined training and validation set using the hyperparameters obtained through the earlier cross validation.

4. Finally, the predictive accuracy of the newly trained L2 logit model was assessed on the held out test set, in which the test set was again encoded from the combined training and testing for the feature learning learning cases. Note that the encoding of the test set was assumed the same as the full data set due to the choice set generation procedure detailed in Section 3.2, i.e., each 75% split was chosen from the each unique user, as well as the number of unique cars being much less than a combined train and validation set.

6 Results and Discussion

Table 3 shows the averaged prediction accuracy of the trained model on the baseline and on the learned features from sparse coding or sparse RBM. The latter achieve up to 81.4% predictive accuracy, while the baseline achieved 71.0%. Features learned using sparse RBM resulted in better prediction accuracy than features learned using sparse coding.

Figure 3 shows the influence of the overcompleteness factor γ and the sparsity penalty β on sparse coding and sparse RBM, assessed during cross validation, and thus with lower prediction accuracy than the test results due to less data used for learning.

The experiment shows that significant increases in consumer preference prediction accuracy are possible using sparse coding or the restricted Boltzmann machine for transforming the original data variables to a more representative feature space. Moreover, the prediction performance is highly dependent on the sparsity penalty β and overcompleteness factor γ chosen for both sparse coding and the restricted Boltzmann machine.

These findings underpin the major contribution in this paper; namely, to show that unsupervised feature learning methods may be applied to marketing data sets characterized by heterogeneous data types for a design preference prediction task. These findings are novel in the sense that these feature learning methods have primarily been applied on homogeneous data sets made up on variables of the same distribution, as well as for non-preference prediction tasks. Notable exceptions include work done in the area of collaborative filtering [30].

Consequently, there is much room for improvement within this area. For example, by stacking feature learning layers (often referred to as "deep learning") and fine-tuning further, researchers have recently shown impressive results by breaking previous records in image recognition by large margins [42]. Although we have shown that even a single layer of feature learning can significantly increase predictive accuracy for preference modeling, such deep learning methods may prove to further increase preference prediction accuracy. Future work may also account for the inherent conditional statistical structure between consumers and products, as well as by further accounting for heterogeneous variable types commonly associated with design preference and marketing data sets.

Better understanding of the limitations of using feature learning methods for design and marketing research should also be investigated. For example, the large number of parameters associated with feature learning methods results in greater computational cost when performing model selection. In addition to the cross validation techniques used in this paper, model selection metrics such as BIC and AIC may give further insight along these lines.

There also lie opportunities to develop utility models more representative of the consumer preference task. Our method of defining utility did not incorporate consumer variables as main effects, instead only incorporating consumer variables in interaction terms as detailed in Section 3.4. Furthermore, the use of a descriptive utility model versus our use of a normative utility model may capture a consumer's purchase preferences better.

7 Conclusion

We have presented two feature learning methods, sparse coding and sparse restricted Boltzmann machines, applied to a market data set characterized by heterogeneous unit type (e.g., gender, age, number of cylinders). Our results show that feature learning methods can significantly improve prediction accuracy on a consumer preference prediction task. Our work highlights the key point that these methods are complementary to the choice of classification model. Specifically, all results presented use the same L2 logit model classification algorithm, yet classification is performed in either the original data space or in learned feature space.

The reported results indicate potential for further design research. More advanced feature learning methods already developed by the machine learning community may be used for market data, especially those that stack multiple layers of the single-layer feature learning methods presented in this paper. The pursuit of feature learning methods tailored to market data may result in useful developments for the machine learning community as well, especially given the heterogeneous nature of the data types encountered with market data.

Acknowledgments

The authors would like to thank Bart Frischknecht and Kevin Bolon for the assistance in coordinating data sets, Clayton Scott for useful suggestions, and Maritz Research Inc. for generously making the use of their data possible. This work has been supported by the National Science Foundation under Grant No. CMMI-1266184. This support is gratefully acknowledged.

REFERENCES

[1] Wassenaar, H. J., and Chen, W., 2003. "An approach to decision-based design with discrete choice analysis for demand modeling". *Journal of Mechanical Design,* **125**, p. 490.

[2] Lewis, K. E., Chen, W., and Schmidt, L. C., 2006. *Decision making in engineering design*. American Society of Mechanical Engineers.

[3] Wassenaar, H., Chen, W., Cheng, J., and Sudjianto, A., 2005. "Enhancing discrete choice demand modeling for decision-based design". *Journal of Mechanical Design,* **127**(4), pp. 514–523.

[4] Kumar, D., Hoyle, C., Chen, W., Wang, N., Gomez-Levi, G., and Koppelman, F., 2007. "Incorporating customer preferences and market trends in vehicle package design". *ASME International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*, pp. 571–581.

[5] Reid, T. N., Frischknecht, B. D., and Papalambros, P. Y., 2012. "Perceptual attributes in product design: Fuel economy and silhouette-based perceived environmental friendliness tradeoffs in automotive vehicle design". *Journal of Mechanical Design,* **134**, p. 041006.

[6] Michalek, J., Feinberg, F., and Papalambros, P., 2005. "Linking marketing and engineering product design decisions via analytical target cascading". *Journal of Product Innovation Management,* **22**(1), pp. 42–62.

[7] He, L., Chen, W., Hoyle, C., and Yannou, B., 2012. "Choice modeling for usage context-based design". *Journal of Mechanical Design,* **134**, p. 031007.

[8] McFadden, D., and Train, K., 2000. "Mixed MNL models for discrete response". *Journal of Applied Econometrics,* **15**(5), pp. 447–470.

[9] Berkovec, J., and Rust, J., 1985. "A nested logit model of automobile holdings for one vehicle households". *Transportation Research Part B: Methodological,* **19**(4), pp. 275–285.

[10] Hauser, J. R., Toubia, O., Evgeniou, T., Befurt, R., and Dzyabura, D., 2010. "Disjunctions of conjunctions, cognitive simplicity, and consideration sets". *Journal of Marketing Research,* **47**(3), pp. 485–496.

[11] Chapelle, O., and Harchaoui, Z., 2004. "A machine learning approach to conjoint analysis". *Advances in Neural Information Processing Systems*, pp. 257–264.

[12] Evgeniou, T., Pontil, M., and Toubia, O., 2007. "A convex optimization approach to modeling consumer heterogeneity in conjoint estimation". *Marketing Science,* **26**(6), pp. 805–818.

[13] Toubia, O., Simester, D. I., Hauser, J. R., and Dahan, E., 2003. "Fast polyhedral adaptive conjoint estimation". *Marketing Science,* **22**(3), pp. 273–303.

[14] Abernethy, J., Evgeniou, T., Toubia, O., and Vert, J.-P., 2008. "Eliciting consumer preferences using robust adaptive choice questionnaires". *IEEE Transactions on Knowledge and Data Engineering,* **20**(2), pp. 145–155.

[15] Lee, T. Y., and Bradlow, E. T., 2007. Automatic construction of conjoint attributes and levels from online customer reviews. Tech. rep., University Of Pennsylvania, The Wharton School Working Paper.

[16] Coates, A., Ng, A. Y., and Lee, H., 2011. "An analysis of single-layer networks in unsupervised feature learning". *International Conference on Artificial Intelligence and Statistics*, pp. 215–223.

[17] Bengio, Y., and LeCun, Y., 2007. "Scaling learning algorithms towards AI". *Large-Scale Kernel Machines,* **34**, pp. 1–41.

[18] Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y., 2009. "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations". *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 609–616.

[19] Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y., 2011. "Unsupervised learning of hierarchical representations with

convolutional deep belief networks". *Communications of the Association for Computing Machinery,* **54**(10), pp. 95–103.

[20] Mohamed, A.-R., Sainath, T. N., Dahl, G., Ramabhadran, B., Hinton, G. E., and Picheny, M. A., 2011. "Deep belief networks using discriminative features for phone recognition". *2011 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5060–5063.

[21] Lee, H., Largman, Y., Pham, P., and Ng, A. Y., 2009. "Unsupervised feature learning for audio classification using convolutional deep belief networks". pp. 1096–1104.

[22] Dahl, G., Mohamed, A.-R., Hinton, G. E., et al., 2010. "Phone recognition with the mean-covariance restricted Boltzmann machine". *Advances in Neural Information Processing Systems*, pp. 469–477.

[23] Salakhutdinov, R., and Hinton, G., 2009. "Semantic hashing". *International Journal of Approximate Reasoning,* **50**(7), pp. 969–978.

[24] Torralba, A., Fergus, R., and Weiss, Y., 2008. "Small codes and large image databases for recognition". *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8.

[25] Collobert, R., and Weston, J., 2008. "A unified architecture for natural language processing: Deep neural networks with multitask learning". *Proceedings of the 25th International Conference on Machine Learning*, pp. 160–167.

[26] Mnih, A., and Hinton, G., 2009. "A scalable hierarchical distributed language model". *Advances in Neural Information Processing Systems,* **21**, pp. 1081–1088.

[27] Olshausen, B. A., et al., 1996. "Emergence of simple-cell receptive field properties by learning a sparse code for natural images". *Nature,* **381**(6583), pp. 607–609.

[28] Lee, H., Battle, A., Raina, R., and Ng, A., 2006. "Efficient sparse coding algorithms". *Advances in Neural Information Processing Systems*, pp. 801–808.

[29] Smolensky, P., 1986. "Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1". MIT Press, Cambridge, MA, USA, ch. Information Processing in Dynamical Systems: Foundations of Harmony Theory, pp. 194–281.

[30] Salakhutdinov, R., Mnih, A., and Hinton, G., 2007. "Restricted Boltzmann machines for collaborative filtering". *Proceedings of the 24th International Conference on Machine Learning*, pp. 791–798.

[31] Olshausen, B. A., and Field, D. J., 1997. "Sparse coding with an overcomplete basis set: A strategy employed by V1?". *Vision Research,* **37**(23), pp. 3311–3325.

[32] Lewicki, M. S., and Sejnowski, T. J., 2000. "Learning overcomplete representations". *Neural Computation,* **12**(2), pp. 337–365.

[33] Lee, H., Ekanadham, C., and Ng, A. Y., 2008. "Sparse deep belief net model for visual area V2". *Advances in Neural Information Processing Systems 20*, pp. 873–880.

[34] Maritz Research Inc., 2007. Maritz Research 2006 new vehicle customer satisfactions survey. Information online at: `http://www.maritz.com`.

[35] Chrome Systems Inc., 2008. Chrome New Vehicle Database. Information inline at: `http://www.chrome.com`.

[36] United States Census Bureau, 2006. 2006 U.S. Census estimates. Information online at: `http://www.census.gov`.

[37] Shocker, A. D., Ben-Akiva, M., Boccara, B., and Nedungadi, P., 1991. "Consideration set influences on consumer decision-making and choice: Issues, models, and suggestions". *Marketing Letters,* **2**(3), pp. 181–197.

[38] Kumar, D., Hoyle, C., Chen, W., Wang, N., Gomez-Levi, G., and Koppelman, F., 2013. "Predicting Consumer Choice Set Using Product Association Network and Data Analytics". *ASME International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*.

[39] Lee, H., Raina, R., Teichman, A., and Ng, A. Y., 2009. "Exponential family sparse coding with applications to self-taught learning". *Proceedings of the 21st International Joint Conference on Artifical Intelligence (IJCAI)*, pp. 1113–1119.

[40] Hinton, G. E., 2002. "Training products of experts by minimizing contrastive divergence". *Neural computation,* **14**(8), pp. 1771–1800.

[41] Scott, C., 2012. The equivalence between the one-class and paired support vector machines for nonseparable data. Tech. rep., Communications and Signal Processing Laboratory, University of Michigan.

[42] Krizhevsky, A., Sutskever, I., and Hinton, G. E., 2012. "Imagenet classification with deep convolutional neural networks". In *Advances in Neural Information Processing Systems 25*. pp. 1097–1105.