Learning Predictive Choice Models for Decision Optimization

Waheed Noor, Matthew N. Dailey, *Senior Member, IEEE,* and Peter Haddawy

Abstract—Probabilistic predictive models are often used in decision optimization applications. Optimal decision making in these applications critically depends on the performance of the predictive models, especially the accuracy of their probability estimates. In this paper, we propose a probabilistic model for revenue maximization and cost minimization across applications in which a decision making agent is faced with a group of possible customers and either offers a variable discount on a product or service or expends a variable cost to attract positive responses. The model is based directly on optimizing expected revenue and makes explicit the relationship between revenue and the customer's response behavior. We derive an expectation maximization (EM) procedure for learning the parameters of the model from historical data, prove that the model is asymptotically insensitive to selection bias in historical decisions, and demonstrate in a series of experiments the method's utility for optimizing financial aid decisions at an international institute of higher learning.

Index Terms—Predictive choice models, decision optimization, sparse data, em algorithm, sample selection bias, imbalance data

1 Introduction

Many applications that attempt to make or recommend optimal decisions exploit predictive models. As a simple example, imagine a company that is able to offer a product or service at a variable discount across a group of potential customers. If the company wants to maximize its revenue, it should offer to each customer the discount that maximizes expected revenue. Optimizing the expected revenue requires predicting, for each possible discount rate, the probability that the customer will accept the offer and purchase the product.

Here we focus on a large class of such revenue and cost optimization problems in which a decision making agent is faced with a group of possible customers and either offers each member of the group a variable discount on a specific product or service or expends a variable cost to attract a positive response. These problems occur across many domains, including loan finance, insurance, direct marketing, network admission, and financial aid for education.

In much of the existing work on predictive optimization for such applications [1]–[4], the predictive model takes as input an individual customer's attributes and a possible offer. The model outputs the probability of a positive response to the offer. The model parameters are selected based on optimization of some intermediate objective function such as classification error or squared error between the predicted probability and the actual outcome on a training set, and then this putatively optimal predictive model is used to optimize the decision variable of interest. We call this the "predictive classification" approach to predictive modeling in decision optimization.

Since predictive classification models are trained in a different mode of operation than they are used in, i.e., they minimize objective functions that do not directly correspond to the way the models are used at decision-making time, they may be suboptimal. A good classifier does not necessarily produce accurate probabilities; for instance, the naive Bayes classifier achieves good classification results even when the conditional independence assumption is violated and the probability estimates contain large errors [5]. Another possible limitation of this approach is that it treats the customer historical responses as being deterministic. It is more common in decision theory to treat customer responses as being probabilistic choices among multiple alternatives governed by some unknown probability distribution.

Predictive classification methods divide customers into positive and negative classes and do not explicitly encode common-sense constraints, such as the fact that in many applications, positive response probabilities should increase or decrease monotonically in the attractiveness of the decision variable. Domain knowledge such as monotonicity constraints, if incorporated into learning algorithms, can be beneficial. As one example, Altendorf, Restificar, and Dietterich [6] incorporate monotonicity by constraining estimates of class conditional probabilities and demonstrate improved generalization performance on standard data

- *W. Noor is with the Computer Science and Information Management, Asian Institute of Technology, Klongluang 12120, Thailand, and also with the Department of Computer Science, University of Balochistan, Quetta 87300, Pakistan. E-mail: waheed.noor@ait.asia; waheed.noor@uob.edu.pk.*
- *M. N. Dailey is with the Computer Science and Information Management, Asian Institute of Technology, Klongluang 12120, Thailand. E-mail: mdailey@ait.asia*
- *P. Haddawy is with the Faculty of ICT, Mahidol University, Nakhon Pathom 73170, Thailand. E-mail: peter.had@mahidol.ac.th.*

Manuscript received 6 June 2012; revised 26 Aug. 2013; accepted 26 Oct. 2013. Date of current version 10 July 2014.
Recommended for acceptance by A. Gionis.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier 10.1109/TKDE.2013.173

sets. When encoded by an explicit choice model, such constraints would arise naturally rather than being forced through constraints on parameter estimates during learning, and choice probabilities could be interpolated in or extrapolated to ranges of the decision variable where training data is insufficient.

A last limitation of predictive classification methods is that many learning methods are sensitive to various kinds of bias in the training set such as imbalanced data and selection bias [7]. For instance, the naive Bayes classifier used by [1] and [3] is well known to be sensitive to imbalanced data and selection bias [8].

In this paper, we formulate a probabilistic model that makes explicit the customer's choice behavior. At model selection time, rather than minimizing an intermediate cost function such as training set classification error or squared error, we instead perform maximum likelihood estimation of the parameters of the predictive choice model. At decision-making time, we can select the decision that maximizes expected revenue with respect to the estimated choice model. Our approach makes explicit the relationship between revenue and the customer's choice behavior. We prove that the model is asymptotically insensitive to the type of selection bias that characterizes historical training data, and we also show empirically that it effectively handles imbalanced data in practice.

The novel aspect of the model is that, based on a customer's individual characteristics, it provides estimates of the parameters of an item response function predicting, for each level of the decision variable, the acceptance probability for the corresponding offer. Doing so is difficult because historical training data are sparse; for each record, we have only the individual's characteristics, a single offer, and a single sample from the unknown probability distribution governing the individual's response. In order to estimate choice models tailored to the customer's individual characteristics despite the problem of sparse training data, we collapse the training data for similar historical customers using a mixture model. Architecturally, one may characterize the model as a variant of the mixture of experts [9], but it is trained using a very different objective function: to make a series of decisions that maximize revenue under the double uncertainty of the choice model and the mixture model.

The model is general enough to be used across many domains, but we are specifically motivated by the problem of optimizing financial aid offer decisions by an international institute of higher learning. The objective is to determine the financial aid offer for each student that would maximize the institute's expected revenue for that student. Although in this domain, other factors, such as an applicant's academic merits and the institute's faculty and facility capacity, will strongly influence the decision maker's final decision, the revenue-maximizing financial aid offer makes an excellent starting point for that decision.

In a first series of experiments with synthetic data, we find that the method is capable of learning accurate predictive models, even in the presence of highly imbalanced data sets and selection bias. In a second series of experiments with real-world data from the admissions department at the authors' institute, we find that the model accurately predicts the decision-making behavior of a pool of applicants for admission to master and doctoral programs in 2008. Based on offers the institute actually made to the pool of applicants, the model predicts an expected enrollment of 205 students compared to an actual enrollment of 202 students. As might be predicted from the experiments with synthetic data, the predictive choice model outperforms LogitBoost (boosted logistic regression models) [10] and logistic model trees (LMT) the model indicates that by modifying its offers, the institute could have increased its revenue from the pool of 2008 applicants by 22%.

We conclude that predictive choice models provide an effective and elegant method for accurate estimation of conditional probabilities in the presence of a decision variable under training set selection bias. They output accurate response probability and expected revenue curves tuned to particular subsets of customers. This makes them extremely effective tools for giving decision makers insight into their customers' behavior and for improving the quality of decision making across a wide variety of applications.

2 Mathematical Model

Let $\boldsymbol{X}$ be an M-dimensional random vector over the general population space $\mathcal{X}$ of features representing the characteristics of individuals, and let $\boldsymbol{x}_i, i = 1, 2, \ldots, N$ represent feature vectors sampled independently and identically distributed (i.i.d.) from distribution $P(\boldsymbol{X} = \boldsymbol{x})$. We write $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iM})$, where x_{im} is the value of the m^{th} feature of the i^{th} example in the sample.

Let Y be a random variable over the set of class labels $\mathcal{Y} = \{0, 1\}$ representing individual outcomes. We write $y_i = 1$ and $y_i = 0$ to represent a positive and negative outcome for the i^{th} example, respectively.

Let $H = \{(\boldsymbol{x}_i, y_i)\}_{i \in 1,\ldots,N} \subseteq \mathcal{X} \times \mathcal{Y}$ denote a set of labeled training examples (historical data). We assume each example is sampled i.i.d. from some unknown distribution over $\mathcal{X} \times \mathcal{Y}$. Let $U = \{\boldsymbol{x}_i\}_{i \in 1,\ldots,K} \subseteq \mathcal{X}$ denote a set of unlabeled examples (future data). We assume U is drawn i.i.d. from some unknown distribution over $\mathcal{X}$.

Then the goal of predictive model estimation is, given H, to learn or approximate an unknown function $f: \mathcal{X} \mapsto \mathcal{Y}$ or an unknown distribution $P(Y = y \mid \boldsymbol{X} = \boldsymbol{x})$ minimizing some cost function. Taking the probabilistic approach, using Bayes' theorem, we can express these probabilities as $P(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{P(\boldsymbol{X}=\boldsymbol{x}|Y)P(Y)}{P(\boldsymbol{X}=\boldsymbol{x})}$. The predictive model generates posterior probabilities $P(Y \mid \boldsymbol{X} = \boldsymbol{x})$ for each class label.[1]

If the end goal was maximum a posteriori (MAP) classification, i.e., to assign the most probable class label, we would simply choose the class label with the highest posterior probability. However, in the case of learning predictive models for optimization, our end goal is to make optimal decisions maximizing some utility function. For simplicity we will use the more specific term *revenue* rather than utility. Revenue may not be constant for all examples and may depend on some decision variable D. Let d represent a particular but arbitrary value the decision variable D can take on. If we write $R_k(d)$ to indicate the revenue obtained when

1. For simplicity we will use $\boldsymbol{x}$ in place of $\boldsymbol{X} = \boldsymbol{x}$ in our probability expressions.

the decision or offer is d and the outcome or response is k, the *expected revenue* obtained when individual $\boldsymbol{x}$ is offered decision d would be $\sum_k R_k(d)P(Y = k \mid \boldsymbol{x}, d)$. Now our goal is to find, for each example $\boldsymbol{x}_i \in U$, the optimal decision or offer

$$d_i^* = \underset{d_i}{\operatorname{argmax}} \sum_k R_k(d_i)P(Y = k \mid \boldsymbol{x}_i, d_i). \quad (1)$$

For simplicity, we assume that the decision variable D is continuous with $D \in [0..1]$, but extensions to ordinal or discrete decision variables are also possible. We further assume that for a specific $\boldsymbol{x}_i$, when $y_i = 1$, the revenue R is a decreasing linear function of D, i.e., $R_1(d) = 1 - d$, and that when $y_i = 0$, the revenue is 0, i.e., $R_0(d) = 0$. Under these assumptions, (1) for the optimal decision specializes to

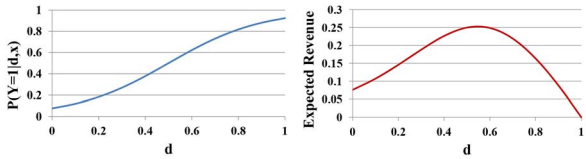
$$d_i^* = \underset{d_i}{\operatorname{argmax}}\, P(Y = 1 \mid \boldsymbol{x}_i, d_i)[1 - d_i]. \quad (2)$$

Therefore, to make an optimal offer to individual i using Equation 2, one might first estimate $P(Y = 1 \mid d_i, \boldsymbol{x}_i)$ as a function of d_i and then select the d_i at which the expression is maximized. Importantly, under this approach, accurate prediction of the optimal decision depends on the accuracy of the actual probabilities output by the predictive model $P(Y = 1 \mid d, \boldsymbol{x}_i)$ rather than its effectiveness in classifying examples into positive and negative classes.

In order to model the positive response probability curve $P(Y = 1 \mid d_i, \boldsymbol{x}_i)$ as a function of d_i, we will make use of random utility theory, which forms the basis of discrete choice models [12], [13]. A discrete choice model provides the probability that individual i will choose alternative j over other alternatives $j', j' \in J, j' \neq j$, where J is total set of alternatives, assuming that each individual will choose the alternative maximizing his or her utility. In the predictive optimization context, once the decision maker has committed to an offer d_i to individual i, the customer is faced with a two-alternative discrete choice: to accept or reject the offer. Let U_i be the utility that individual i obtains by accepting the offer as opposed to rejecting it. In the random utility model, U_i is assumed to take the form $U_i = V_i + \epsilon_i$, where V_i is deterministic and ϵ_i is a random variable over some distribution. We assume that V_i is linear in d_i, parameterized by the form $V_i = k_i(d_i - \eta_i)$, so that η_i represents the offer level at which acceptance and rejection are equally attractive for individual i and k_i represents individual i's sensitivity to the offer level. If we further assume that ϵ_i is distributed according to the logit distribution, the acceptance probability, which we will denote as $f(d_i; \eta_i, k_i)$, can be given by

$$P(Y = 1 \mid d_i, \boldsymbol{x}_i) \approx f(d_i; \eta_i, k_i) = \frac{1}{1 + e^{-k_i(d_i - \eta_i)}}. \quad (3)$$

$f(d_i; \eta_i, k_i)$ is the well-known logistic sigmoid function[14], where η_i is the inflection point at which the slope is maximal and k_i determines the slope of the function at the inflection point. Besides its natural derivation from a random utility model, Equation 3 has attractive common-sense characteristics such as smooth monotonicity of the response probability as a function of the decision variable.



(a) (b)

Fig. 1. Response probability example. (a) Probability of a positive response according to the sigmoid model of Equation 3. (b) Expected revenue for the same individual according to the model of (2).

TABLE 1
General Form of Historical Data Used for Predictive Optimization

X_1	X_2	X_3	$\cdots$	X_M	D	Y
x_{11}	x_{12}	x_{13}	$\cdots$	x_{1M}	d_1	y_1
x_{21}	x_{22}	x_{23}	$\cdots$	x_{2M}	d_2	y_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{N1}	x_{N2}	x_{N3}	$\cdots$	x_{NM}	d_N	y_N

Fig. 1 gives a sample of a typical positive response probability curve $P(Y = 1 \mid d_i, \boldsymbol{x}_i)$ according to (3) and the corresponding expected revenue curve according to the assumptions made thus far. The optimal decision for this individual would be to make an offer of $d = 0.5470$, where the expected revenue is maximal. We will see in Section 3.6 how this optimal offer can be calculated.

3 Model Estimation

The mathematical model introduced in Section 2 is appropriate for many real world applications such as item response, credit scoring, enrollment management, and bid pricing, in which we must estimate values for some control parameters (in our case the decision variables D_i) to optimize cost or revenue.

In this section, we present a method for estimating such models from historical data. Table 1 shows the form of the historical data given in these domains. Each d_i represents the value for the decision variable, D_i, that was historically selected by an expert as the offer for individual i based on features $\boldsymbol{x_i}$.

Setting D_i to a certain value d_i for each individual i is a decision making task that is followed by a response from individual i. Unfortunately, unlike many traditional supervised learning problems in which the training data are complete, the historical data(refer again to Table 1), for a particular individual i, only provides a single acceptance or rejection for an arbitrary (and not necessarily optimal) decision. The data do not provide any information about what individual i's response would be for any offer other than d_i; these data are censored.

Despite the difficulty that this censorship introduces, we would like to find, for each input case $\boldsymbol{x}_i$, the optimal decision d_i^* maximizing expected revenue per (2). We propose to do so by modeling the individual's response behavior as being probabilistic with $P(Y = 1 \mid \boldsymbol{x}_i, d_i)$ being a sigmoid function with parameters η and k, per (3).

To make it practical to learn the predictive model $P(Y = 1 \mid \boldsymbol{x}_i, d_i)$ from a sparse training set, we will

assume that individuals with similar attributes, according to some similarity function, *have the same response behavior*. This assumption holds in many domains of predictive choice modeling, so we exploit it in our model by effectively pooling the response behavior of similar individuals. When the assumption does not hold, other simple models such as naive Bayes that treat attribute levels independently might perform as well or better than our model.

Pooling historical data for similar individuals means that the choice model parameters are only personalized to a particular cluster of individuals, not to each individual. But it is important to realize that the cluster an individual falls into is determined by his or her individual attributes, so there is indeed an extremely strong dependency of the choice model parameters on the customer's attributes. As we shall see, our model will be set up to not only pool historical records for similar individuals (as would a standard clustering algorithm), but to do so in a way that separates groups of individuals with dissimilar choice behavior into different clusters.

We will use a mixture density model to represent clusters of similar individuals. Each component density in the mixture could be an arbitrary density, but we will assume it to be Gaussian, and each $\boldsymbol{x}_i$ is generated by one of the components in the mixture, although the identity of the component that generated $\boldsymbol{x}_i$ is unknown. Of course, this model requires continuous variables or separate models for different levels of discrete variables. For convenience, in the experiments reported here, we will cast ordinal variables as continuous and avoid the use of non-ordinal discrete variables. When we use a mixture model to combine a collection of binary logit choice models, we will obtain a model that has the flavor of the mixed logit model, which is known to be sufficiently powerful to approximate any random utility model [15].

We are thus faced with a learning problem under two types of uncertainty: the mixture component assignment for each individual (the *mixture model*) and the actual values of η and k for each mixture component (the *choice model*). We use the expectation maximization (EM) algorithm [16] to find a maximum likelihood solution to this learning problem. The mixture model and an explicit customer choice model to approximate response probabilities will enable the predictive choice model to interpolate or extrapolate response probabilities to regions where the training data are insufficient or nonexistent, thus reducing the effect of censoring.

3.1 Preliminaries

Assume that there is a finite number of components J in the mixture model. Let $\mathbb{X} = (X, D, Y)$ be the observed (training) data set of N instances, where $X = \{\boldsymbol{x}_i\}_{i\in 1\ldots N}$, $D = \{d_i\}_{i\in 1\ldots N}$ and $Y = \{y_i\}_{i\in 1\ldots N}$. Let $\boldsymbol{\theta} = \{\pi_j, \boldsymbol{\mu}_j, \Sigma_j, \eta_j, k_j\}_{j\in 1\ldots J}$ be the set of parameters to be estimated. π_j, $\boldsymbol{\mu}_j$ and Σ_j represent the mixing coefficients, mean vector, and covariance matrix for the j^{th} component in the mixture model. Let $\boldsymbol{Z} = \{\boldsymbol{z}_i\}_{i\in 1\ldots N}$ be the set of assignments, which are hidden variables. $\boldsymbol{z}_i$ is a J-dimensional binary vector in which a particular element z_{ij} is equal to 1 and all others are equal to 0, i.e., for all $i \in 1, \ldots, N$ and all $j \in 1, \ldots, J$, $z_{ij} \in \{0, 1\}$, and $\sum_{j=1}^{J} z_{ij} = 1$).

3.2 Complete Data Likelihood

Now, we can define the complete data as $\{\mathbb{X}, \boldsymbol{Z}\}$ and its likelihood as

$$P(\mathbb{X}, \boldsymbol{Z} \mid \boldsymbol{\theta}) = P(\boldsymbol{Z} \mid \boldsymbol{\theta}) P(\mathbb{X} \mid \boldsymbol{Z}, \boldsymbol{\theta}). \tag{4}$$

Using the product rule, the second term can be factored to obtain

$$\begin{aligned} P(\mathbb{X} \mid \boldsymbol{Z}, \boldsymbol{\theta}) &= P(X, D, Y \mid \boldsymbol{Z}, \boldsymbol{\theta}) \\ &= P(X \mid \boldsymbol{Z}, \boldsymbol{\theta}) P(D \mid X, \boldsymbol{Z}, \boldsymbol{\theta}) P(Y \mid X, D, \boldsymbol{Z}, \boldsymbol{\theta}). \end{aligned} \tag{5}$$

Based the assumption that similar individuals have the same decision behavior and correspond to a component (cluster) in the mixture model, we can assume that X is independent of Y given Z, i.e., $X \perp Y \mid \boldsymbol{Z}$. (5) can now be written as

$$P(\mathbb{X} \mid \boldsymbol{Z}, \boldsymbol{\theta}) = P(X \mid \boldsymbol{Z}, \boldsymbol{\theta}) P(D \mid X, \boldsymbol{Z}, \boldsymbol{\theta}) P(Y \mid D, \boldsymbol{Z}, \boldsymbol{\theta}). \tag{6}$$

When the decisions d_i are given for historical data, it is reasonable to assume that they were made as a deterministic function of X, according to some (possibly sub-optimal) policy. Therefore, we can write

$$P(D \mid X, \boldsymbol{Z}, \boldsymbol{\theta}) = P(D \mid X) = 1,$$

so (6) can be re-written as

$$P(\mathbb{X} \mid \boldsymbol{Z}, \boldsymbol{\theta}) = P(X \mid \boldsymbol{Z}, \boldsymbol{\theta}) P(Y \mid D, \boldsymbol{Z}, \boldsymbol{\theta}), \tag{7}$$

and the likelihood in (4) becomes

$$P(\mathbb{X}, \boldsymbol{Z} \mid \boldsymbol{\theta}) = P(\boldsymbol{Z} \mid \boldsymbol{\theta}) P(X \mid \boldsymbol{Z}, \boldsymbol{\theta}) P(Y \mid D, \boldsymbol{Z}, \boldsymbol{\theta}). \tag{8}$$

To expand $P(Y \mid D, \boldsymbol{Z}, \boldsymbol{\theta})$, we use the choice model from (3). Taking into account positive and negative choices, we can write

$$P(y_i \mid \boldsymbol{x}_i, d_i) = \begin{cases} f(d_i; \eta_j, k_j), & \text{if } y_i = 1, \\ 1 - f(d_i; \eta_j, k_j), & \text{if } y_i = 0. \end{cases}$$

With the assumption that similar individuals have the same decision behavior under the mixture model and that individual choices are conditionally independent, we can write $P(Y \mid D, \boldsymbol{Z}, \boldsymbol{\theta})$ as

$$\prod_{i=1}^{N} \prod_{j=1}^{J} \left[f(d_i; \eta_j, k_j)^{y_i} (1 - f(d_i; \eta_j, k_j))^{1-y_i} \right]^{z_{ij}}. \tag{9}$$

From (9) and following Bishop [14] for defining mixing coefficients π_j as marginal probabilities over $\boldsymbol{Z}$, and conditional distribution of X, the logarithm of the likelihood (in (8)) takes the form

$$\begin{aligned} \ln P(\mathbb{X}, \boldsymbol{Z} \mid \boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{j=1}^{J} z_{ij} \Big[&\ln \pi_j + \ln \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_j, \Sigma_j) + \\ &y_i \ln f(d_i; \eta_j, k_j) + (1 - y_i) \ln(1 - f(d_i; \eta_j, k_j)) \Big]. \end{aligned} \tag{10}$$

3.3 Expectation of the Likelihood

The log likelihood in (10) is only directly useful when we know the values of the latent variables (the z_{ij}). We instead consider the expectation of the complete data likelihood with respect to the posterior distribution of the latent variables [14]. This posterior distribution can be written as

$$P(z_{ij} = 1 \mid \boldsymbol{x}_i, y_i, d_i, \boldsymbol{\theta}) = \frac{\pi_j \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_j, \Sigma_j) f(d_i; \eta_j, k_j)^{y_i} (1 - f(d_i; \eta_j, k_j))^{1-y_i}}{\sum_{l=1}^{J} \pi_l \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_l, \Sigma_l) f(d_i; \eta_j, k_j)^{y_i} (1 - f(d_i; \eta_j, k_j))^{1-y_i}}. \quad (11)$$

Since z_{ij} is a binary variable, we have

$$\mathbb{E}[z_{ij}] = P(z_{ij} = 1 \mid \boldsymbol{x}_i, y_i, d_i, \boldsymbol{\theta}).$$

We shall denote this expectation, also known as the "responsibility" of component j for observation i, by $\gamma(z_{ij})$. Using (11) to obtain the expectation of the complete data log likelihood (10) gives us

$$\mathbb{E}_{\boldsymbol{Z}}[\ln P(\mathbb{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})] = \sum_{i=1}^{N} \sum_{j=1}^{J} \gamma(z_{ij}) \Big[\ln \pi_j + \ln \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_j, \Sigma_j) + y_i \ln f(d_i; \eta_j, k_j) + (1 - y_i) \ln(1 - f(d_i; \eta_j, k_j)) \Big]. \quad (12)$$

Our EM algorithm consists of iteratively maximizing Equation 12 with respect to $\boldsymbol{\theta}$ and recalculating the expectations $\gamma(z_{ij})$. We now show how to perform the maximization step.

3.4 Maximization of the Expected Likelihood

To maximize (12) with respect to $\boldsymbol{\theta}$, we can differentiate with respect to each parameter in $\boldsymbol{\theta}$ and solve the system of resulting equations. This leads to an analytical solution for the mixture model and an iterative optimization procedure for the choice model.

3.4.1 Mixture Model Parameters

Because of the factorial structure of our likelihood, the derivatives with respect to the mixture model parameters do not involve the choice model parameters, and vice versa. Due to this, the derivatives with respect to π_j, $\boldsymbol{\mu}_j$, and Σ_j and the resulting parameter update equations in the EM maximization step are identical to those obtained for the standard Gaussian mixture model [14].

3.4.2 The Choice Model

To maximize (12) with respect to the choice model parameters η_j and k_j, when we take the partial derivatives and set them to 0, we obtain a nonlinear system with no closed-form solution. This means that on each iteration of the EM algorithm, the maximization step itself requires iterative optimization of the choice model parameters. We use conjugate gradient ascent to find the parameters η_j and k_j maximizing $\mathbb{E}_{\boldsymbol{Z}}$, keeping the responsibilities $\gamma(z_{ij})$ fixed. For the optimization, we define the weight vector as $\boldsymbol{w} = (\eta_j, k_j)$.

For simplicity, we denote the likelihood function $\mathbb{E}_{\boldsymbol{Z}}[\ln P(\mathbb{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})]$ (12) as $\mathbb{E}_{\boldsymbol{Z}}(\boldsymbol{w})$ and its gradient with respect to $\boldsymbol{w}$ as $\nabla \mathbb{E}_{\boldsymbol{Z}}(\boldsymbol{w})$, where

$$\frac{\delta \mathbb{E}_{\boldsymbol{Z}}(\boldsymbol{w})}{\delta \eta_j} = k_j \sum_{i=1}^{n} \frac{\gamma(z_{ij})}{1 + e^{-k_j(d_i - \eta_j)}} \left(1 - y_i - y_i e^{-k_j(d_i - \eta_j)}\right),$$

$$\frac{\delta \mathbb{E}_{\boldsymbol{Z}}(\boldsymbol{w})}{\delta k_j} = \sum_{i=1}^{N} \frac{\gamma(z_{ij})(d_i - \eta_j)}{1 + e^{-k_j(d_i - \eta_j)}} \left(y_i + y_i e^{-k_j(d_i - \eta_j)} - 1\right).$$

Our procedure initializes the choice model parameters for component i with positive random values constrained as $0 \geq \eta_i \leq 1$ and $0 > k_i \leq 30$. Here we use $k_i \leq 30$, because values greater than 30 have little effect on the slope of the curve. Although random initialization works well in practice, more accurate initial values could be obtained, for example, through weighted linear regression.

3.5 Complete Model Estimation Procedure

Up to now, we have assumed that the number of components J in the mixture model is fixed and given a priori. In practice, of course, J is not known in advance. We thus repeat the estimation procedure for multiple levels of J and select the best model using the *minimum description length* (MDL) criterion [17]. We find that minimizing the MDL criterion leads to a good balance between model complexity and model fit. There do exist more sophisticated techniques such as the variational Bayesian framework for optimal model selection for mixture models [18], [19], but we find the MDL criterion to be adequate.

The EM procedure for our model, like any EM procedure, increases the likelihood function monotonically on each iteration, so it is guaranteed to converge to a local maximum, but the specific maximum depends on the initial values of the parameters [16]. To avoid suboptimal local maxima, in our experiments, we simply perform multiple restarts with different initial random parameter settings. It would be straightforward to adopt a more sophisticated approach such as the deterministic annealing method for EM to reduce the computational complexity of the model estimation procedure [20], [21].

We can now specify the complete estimation procedure (Algorithm 1). The algorithm returns the mixture model and choice model parameters for the optimal number of components. The estimated model can then be used for predicting response probabilities for new cases and calculating optimal decisions (offers).

3.6 Optimal Decision Making

Once the model $\boldsymbol{\theta}_{J^*}$ has been learned from training data, faced with a new case $\boldsymbol{x}$, we would like to find the decision d^* maximizing expected revenue. We first use the mixture model to find the most likely cluster j for $\boldsymbol{x}$. Then, from (2), we can obtain

$$d^* = \operatorname*{argmax}_d f(d; \eta_j, k_j)[1 - d].$$

Differentiating the expression $f(d; \eta_j, k_j)[1 - d]$ with respect to d, setting to 0, and simplifying, we obtain

$$e^{k_j - k_j \eta_j - 1} = (k_j - d k_j - 1) e^{k_j - d k_j - 1}. \quad (13)$$

Algorithm 1: Predictive Choice Model Estimation

Input: $\mathbb{X}$, $J_{\max}$, $K_{\max}$
Output: Optimal parameters $\boldsymbol{\theta}_{J^*}$
for $J = 1$ **to** J_{max} **do**
  for $K = 1$ **to** K_{max} **do**
    Initialization: $\boldsymbol{\theta}_{JK} \leftarrow$ random initial parameters
    while *not converged* **do**
      E-step: calculate posterior probabilities
      $\gamma(z_{ij})$ from (11) using $\boldsymbol{\theta}_{JK}$
      M-step: re-estimate
      $\boldsymbol{\theta}_{JK}^{\text{new}} = \underset{\boldsymbol{\theta}_{JK}}{\operatorname{argmax}} \mathbb{E}_{\boldsymbol{Z}}[\ln P(\mathbb{X}, \boldsymbol{Z} \mid \boldsymbol{\theta}_{JK})]$ keeping
      $\gamma(z_{ij})$ fixed
      $\boldsymbol{\theta}_{JK} \leftarrow \boldsymbol{\theta}_{JK}^{\text{new}}$
    end
  end
  $\boldsymbol{\theta}_J = \boldsymbol{\theta}_{JK}$, where $K = \underset{K}{\operatorname{argmax}} \mathbb{E}_{\boldsymbol{Z}}[\ln P(\mathbb{X}, \boldsymbol{Z} \mid \boldsymbol{\theta}_{JK})]$
  Calculate MDL($\boldsymbol{\theta}_J$)
end
$\boldsymbol{\theta}_{J^*} \leftarrow \underset{\boldsymbol{\theta}_J}{\operatorname{argmin}}\, \text{MDL}(\boldsymbol{\theta}_J)$

Next, letting $\mathbb{T} = k_j - dk_j - 1$ and letting $\mathbb{Y} = e^{k_j - k_j\eta_j - 1}$, we can rewrite the equation in the form

$$\mathbb{Y} = \mathbb{T}e^{\mathbb{T}}.$$

Such equations can be solved using Lambert's W [22]:

$$\mathbb{T} = W(\mathbb{Y}),$$

giving us the optimal decision

$$d_j^* = \frac{k_j - 1 - W(e^{k_j - k_j\eta_j - 1})}{k_j}. \tag{14}$$

3.7 Bias Sensitivity

In this section, we consider the sensitivity of the predictive choice model to two kinds of bias, in comparison to the generative modeling approach, in which we train a predictive model based on observations of individuals, offers, and responses. The two types of bias are imbalance and selection bias.

Imbalance occurs in classification or supervised learning problems where at least one class is under-represented in the training data relative to other classes. Many classification algorithms require balanced data at training time to perform well. Others may explicitly account for imbalance by estimating prior probabilities from the training data, but uneven class distributions in the training data may or may not be same as in the general population of interest [23]. There is a large body of literature from the machine learning and data mining communities aiming to address and overcome this problem [23]–[28]. Most of these methods are based either on resampling or reweighting. However, a recent method [29] explicitly accounts for imbalance in its formulation to overcome the problem of class imbalance for logistic regression models.

Imbalance is quite common in predictive optimization applications. For example, in financial aid for education, due to competition and insufficient scholarship funds, the proportion of applicants with positive responses to an offer might be small compared to the negatively responding applicants. Under such circumstances, it is important that statistical models used for predictive optimization handle imbalance properly. Fortunately, in the proposed predictive choice model, keeping the decision (offer) fixed, since we are directly estimating the parameters, η and k, of $P(y \mid \boldsymbol{x})$ using maximum likelihood estimation rather than estimating separate models for positive and negative classes, each historical record of a response to that offer for members of a particular group of similar customers is equally informative about $P(y \mid \boldsymbol{x})$, regardless of the balance or imbalance of the responses. Therefore, the predictive choice model overcomes imbalance naturally; we will demonstrate this empirically in Section 4.1.

Now we consider selection bias. Most learning algorithms assume that the training dataset and the test dataset are drawn i.i.d. from the same probability distribution. This is often known as the *stationary distribution* or *no bias distribution* assumption [30]. But in real world applications such as drug testing, direct marketing, loan approval, school enrollment, credit scoring, spam filtering, and species habitat modeling, the training set and test set distributions are often biased. The available data may be biased due to the data selection process for the underlying model [7] or due to not having complete control over the data collection process [8]. For example, in loan approval applications, the training sample would consist of examples with repay/default labels for only those applicants who were approved by the experts for a loan in the first place. However, the actual goal is to model the repay/default behavior of all applicants in general. Such bias violates the stationary distribution assumption and is often known as *sample selection bias*, i.e. if $\mathcal{P}(\boldsymbol{x}, y)$ is the training distribution and $\mathcal{Q}(\boldsymbol{x}, y)$ is the test distribution, then $\mathcal{P}(\boldsymbol{x}, y) \neq \mathcal{Q}(\boldsymbol{x}, y)$. Models trained on biased samples result in poor estimates of expected generalization error [7].

Zadrozny [8] and Fan *et al*. [31] formally characterize different types of selection bias and categorize different inductive learners based on their sensitivity to those types of bias. Based on the notation of [8], let S be a binary random variable that controls the selection of an example $(\boldsymbol{x}, y)$ for the training set, where $S = 1$ means that the example is sampled and $S = 0$ means that the example is not sampled. We assume that the training set $\mathcal{T}$ consists of examples $(\boldsymbol{x}, y, s)$ that are drawn independently from an unknown distribution $\mathcal{D}$ over domain $\mathcal{X} \times \mathcal{Y} \times \mathcal{S}$. Depending on the problem and data set, we may or may not have access to the examples for which $S = 0$. Under this model, different types of selection bias can be identified based on the conditional dependencies between $\boldsymbol{x}$, y, and s. There are four possible patterns for the conditional dependencies, but here we only consider the very common *feature bias*.

Feature bias, also known as *missing at random* (MAR) in the missing data context, is the case when only the observable features $\boldsymbol{x}$ influence the sample selection. s in this case is dependent on $\boldsymbol{x}$ but conditionally independent of y given $\boldsymbol{x}$, or, $P(s = 1 \mid \boldsymbol{x}, y) = P(s = 1 \mid \boldsymbol{x})$ and $P(y \mid \boldsymbol{x}, s = 1) = P(y \mid \boldsymbol{x})$. For example, in direct marketing applications, the decision maker's decision to send an

offer to an individual depends on the individual's characteristics $\boldsymbol{x}$, but not on his or her response, y. Similarly, in medical applications, a treatment is given based on the symptoms $\boldsymbol{x}$. Therefore, in both cases, the selected sample in the historical data is not a random sample from the general population. Feature bias changes the marginal distribution over $\boldsymbol{x}$ in the training set; although $\mathcal{P}(y \mid \boldsymbol{x}) = \mathcal{Q}(y \mid \boldsymbol{x})$, the training and test data distributions are associated via $\mathcal{P}(\boldsymbol{x}) \propto \mathcal{Q}(\boldsymbol{x})P(s = 1 \mid \boldsymbol{x})$.

Depending on the way learning algorithms model $P(y \mid \boldsymbol{x})$, the specific data set under consideration, and the specific form of the bias, different learning algorithms may or may not be sensitive to feature bias. Zadrozny and colleagues categorize learning algorithms as "local" for a specific scenario if they are not sensitive to feature bias and "global" otherwise. For example, general Bayesian classifiers are often local, and naive Bayes classifiers are often global.

Researchers from econometrics, statistics, and machine learning have made substantial progress in sample selection bias correction for different statistical and machine learning models [7], [8], [31]–[35]. Many of the methods attempt some form of reweighting or resampling in regions where the training data are sparse, in the hope of improving model's performance in those sparse regions.

Another group of methods relies on maximum likelihood estimates of the parameters to avoid feature bias [32]. For example, in loan approval applications, if the repayment model parameters are estimated via maximum likelihood from the accept cases, it is possible to extrapolate the repayment probability model to reject cases, thereby eliminating the feature bias.

Predictive optimization data suffer from feature bias. In the generative modeling approach, in which the predictive model is a classifier, it is very common to include the decision variable D as a part of the feature vector $\boldsymbol{X}$ at training time but as a control variable at test time. When this is the case, the sample selection variable S depends directly on $\boldsymbol{X}$ (in particular the element D of $\boldsymbol{X}$), but not on Y. For example, in areas such as direct marketing, bid pricing, financial aid allocation, and bandwidth allocation in network admission, we only observe customer responses for the values/levels of the decision variable that were selected by decision makers in the past.

Here we formalize the selection process for the predictive optimization problem. We are given a particular set of training customers $\mathcal{T}$. Since our ultimate goal is to find the optimal offer to make to each eligible customer, we first assume that customers ineligible for an offer have been removed by a pre-process. For example, in financial aid allocation, we presume that applicants are pre-screened for admission eligibility and then offers are made to each eligible applicant. After eligibility screening, for each customer $\boldsymbol{x}_i \in \mathcal{T}$, a specific value d_i for D_i is selected according to some (unknown, possibly suboptimal) policy. At this point, since every customer is given exactly one offer, for the selected decision d_i, we can write $P(d_i \mid \boldsymbol{x}_i) = 1$ and, for all $d'_i \neq d_i$, $P(d'_i \mid \boldsymbol{x}_i) = 0$. Finally, we observe the customer's response, y_i. We assume that customer i's response is drawn i.i.d. from a binary probability distribution in which the probability of acceptance is given by the logistic sigmoid function $f(d_i; \eta_i^*, k_i^*)$ in d_i with unknown control parameters η_i^* and k_i^*. The customer's attributes $\boldsymbol{x}_i$, the actual decision d_i, and the actual response y_i, over all i, form the given *biased training set* $\mathbb{X}$.

Under this model, we can show that under certain circumstances, our estimation procedure for the predictive choice model is asymptotically insensitive to feature bias. That is, we must show that our estimate of $P(Y \mid d, \boldsymbol{x})$ for any d is not affected by the selection process. Our estimate of $P(Y \mid d, \boldsymbol{x})$ is based on maximum likelihood estimation of model parameters $\boldsymbol{\theta}$ according to the EM algorithm described in Section 3.5. To prove insensitivity to feature bias, we posit the existence of an *unbiased* training set $\mathbb{X}'$ corresponding to $\mathbb{X}$ in which, rather than the biased decisions d_i, we are given unbiased decisions d'_i drawn i.i.d. from an unknown distribution $P(d \mid \boldsymbol{x})$ and unbiased responses y'_i drawn i.i.d. from the distribution in which the probability of acceptance is given by $f(d'_i; \eta_i^*, k_i^*)$. If $\boldsymbol{\theta}$, a set of parameters locally maximizing the expected log likelihood of $\mathbb{X}$, asymptotically also locally maximizes the expected log likelihood of $\mathbb{X}'$, then our estimation procedure is necessarily insensitive to feature bias.

The main assumption we must make in order to prove unbiasedness of the estimator is that for each customer i, the cluster assignment $\boldsymbol{z}_i$ induced by the estimation procedure is a sufficient statistic for $\boldsymbol{x}_i$. A further needed assumption is that over the set of customers mapped to a specific cluster j, the probability of any particular decision level d is nonzero. The formal theorem statement and proof follow.

Theorem 1. *Let M be a fixed positive integer. Let $\mathbb{X} = \{(\boldsymbol{x}_i, d_i, y_i)\}_{i \in 1 \ldots n}$ be an observed predictive optimization data set with $\boldsymbol{x}_i \in \mathbb{R}^M$, $d_i \in [0..1]$ and $y_i \sim P(y \mid \boldsymbol{x}_i, d_i)$, where the d_i are selected through some arbitrarily biased procedure and $P(y \mid \boldsymbol{x}, d)$ is a logistic sigmoid $f(d; \eta_i^*, k_i^*)$ in d with unknown control parameters η_i^* and k_i^*.*

Let $\mathbb{X}' = \{(\boldsymbol{x}_i, d'_i, y'_i)\}_{i \in 1 \ldots n}$ be a corresponding unbiased data set, with $d'_i \sim P(d \mid \boldsymbol{x}_i)$ for some arbitrary unknown $P(d \mid \boldsymbol{x})$ such that for all d, $\boldsymbol{x}$, $P(d \mid \boldsymbol{x}) > 0$ and $y'_i \sim P(y \mid \boldsymbol{x}_i, d'_i) = f(d'; \eta_i^*, k_i^*)$.

Let $\boldsymbol{\theta}$ be a set of parameters locally maximizing $\mathbb{E}_{\boldsymbol{Z}}[\ln P(\mathbb{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})]$, where $\boldsymbol{Z} = \{\boldsymbol{z}_i\}_{i \in 1 \ldots n}$ is a set of binary random vectors with $\boldsymbol{z}_i \in \{0, 1\}^J$ and for all i, $\sum_j z_{ij} = 1$.

If, for all $i \in 1 \ldots n$, $\boldsymbol{z}_i$ is a sufficient statistic for $\boldsymbol{x}_i$, that is, for any $t \in 1 \ldots n$, $\boldsymbol{z}_i = \boldsymbol{z}_t$ implies that $(\eta_i^*, k_i^*) = (\eta_t^*, k_t^*)$, and if as n approaches infinity, for all $d \in [0..1]$ and all $j \in 1 \ldots J$, $P(d \mid z_{\cdot j}) > 0$, then $\boldsymbol{\theta}$ asymptotically maximizes $\mathbb{E}_{\boldsymbol{Z}}[\ln P(\mathbb{X}', \boldsymbol{Z} \mid \boldsymbol{\theta})]$.

Proof. In (12), the expectation of the complete data log likelihood is taken with respect to the posterior distribution of $\boldsymbol{Z}$, i.e.,

$$\mathbb{E}_{\boldsymbol{Z}}[\ln P(\mathbb{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})] = \ln P(\mathbb{X}, \boldsymbol{Z} \mid \boldsymbol{\theta}) P(\boldsymbol{Z} \mid \mathbb{X}, \boldsymbol{\theta}),$$

where $\ln P(\mathbb{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})$ is the complete data log likelihood (8) and $P(\boldsymbol{Z} \mid \mathbb{X}, \boldsymbol{\theta})$ can be written as

$$P(\boldsymbol{Z} \mid \mathbb{X}, \boldsymbol{\theta}) = \frac{P(\mathbb{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})}{\sum_{\boldsymbol{Z}} P(\mathbb{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})}.$$

To show that the maximum likelihood estimate of $\boldsymbol{\theta}$ is unaffected by selection bias, we need to show that

asymptotically, if $\boldsymbol{\theta}$ maximizes $P(\mathbb{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})$ then $\boldsymbol{\theta}$ also maximizes $P(\mathbb{X}', \boldsymbol{Z} \mid \boldsymbol{\theta})$.

Expanding and rewriting the complete data likelihood in (8) under selection bias, we obtain

$$P(\mathbb{X}, \boldsymbol{Z} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{j=1}^{J} P(z_{ij} = 1 \mid \boldsymbol{\theta}) \\ P(X = \boldsymbol{x}_i \mid z_{ij} = 1, \boldsymbol{\theta}) P(y_i \mid d_i, z_{ij} = 1, \boldsymbol{\theta}).$$

Since the factors $P(z_{ij} = 1 \mid \boldsymbol{\theta})$ and $P(X = \boldsymbol{x}_i \mid z_{ij} = 1, \boldsymbol{\theta})$ do not involve d_i or d_i', they are identical to the corresponding factors in the expansion of $P(\mathbb{X}', \boldsymbol{Z} \mid \boldsymbol{\theta})$.

We are left with the expression $P(y_i \mid d_i, z_{ij} = 1, \boldsymbol{\theta})$. For a specific cluster j, we model $P(y_i \mid d_i, z_{ij} = 1, \boldsymbol{\theta})$ by the sigmoid function $f(d; \eta_j, k_j)$, and, in each M-step of Algorithm 1, the parameters η_j, and k_j are estimated independently of the parameters for other clusters based on maximum likelihood over the training set. This means that if $\boldsymbol{\theta}$ locally maximizes $\mathbb{E}_{\boldsymbol{Z}}[\ln P(\mathbb{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})]$ for any fixed distribution $\boldsymbol{Z}$, it must be the case that for each j, η_j, and k_j locally maximize $\prod_i P(z_{ij} = 1 \mid \boldsymbol{\theta}) P(X = \boldsymbol{x}_i \mid z_{ij} = 1, \boldsymbol{\theta}) f(d_i; \eta_j, k_j)^{y_i} f(d_i; \eta_j, k_j)^{(1-y_i)}$.

Since $\boldsymbol{z}_i$ is by assumption a sufficient statistic for $\boldsymbol{x}_i$, and since as n approaches infinity, for all d, $P(d \mid z_j) > 0$, and since each y_i is an unbiased sample from $P(y \mid \boldsymbol{x}_i, d_i)$, as n increases, the estimator for the parameters η_j, and k_j will converge to the true parameters η_i^*, k_i^*. The same is true for the estimator based on $\mathbb{X}'$.

We can thus say that if $\boldsymbol{\theta}$ locally maximizes $\prod_i \prod_j P(z_{ij} = 1 \mid \boldsymbol{\theta}) P(X = \boldsymbol{x}_i \mid z_{ij} = 1, \boldsymbol{\theta}) f(d_i; \eta_j, k_j)^{y_i} f(d_i; \eta_j, k_j)^{(1-y_i)}$, asymptotically, it must also locally maximize $\prod_i \prod_j P(z_{ij} = 1 \mid \boldsymbol{\theta}) P(X = \boldsymbol{x}_i \mid z_{ij} = 1, \boldsymbol{\theta}) f(d_i'; \eta_j^*, k_j^*)^{y_i'} f(d_i'; \eta_j^*, k_j^*)^{(1-y_i')}$. Since asymptotically, $\boldsymbol{\theta}$ locally maximizes both $\mathbb{E}_{\boldsymbol{Z}}[\ln P(\mathbb{X}, \boldsymbol{Z} \mid \boldsymbol{\theta})]$ and $\mathbb{E}_{\boldsymbol{Z}}[\ln P(\mathbb{X}', \boldsymbol{Z} \mid \boldsymbol{\theta})]$, the predictive choice model is insensitive to feature bias. □

With Theorem 1, we have shown that the predictive choice model, under certain conditions, is a local learner, i.e., it is asymptotically insensitive to feature bias. In practice, of course, the main assumptions are quite strong. We require that N is large, that cluster identity is a sufficient statistic for the individual's attributes in the choice model, and that within each cluster, the probability of any particular decision d being selected by the decision maker in the historical training data must be nonzero. The last two requirements are in opposition to each other: for the sufficient statistic requirement to be true, the clusters should generally be small, but for the selection probability requirement to be true, the clusters should generally be large. When the training data are sparse for a particular cluster, selection bias could certainly affect the resulting choice model estimates. However, we shall see in the forthcoming experimental evaluation that in practice, even when the assumptions are violated, our method nonetheless exhibits good performance.

The above analysis of bias sensitivity of PCM is only applicable to problems with binary responses to a single decision variable. There could be applications in which the customer is faced with multiple choices, under the influence of multiple attributes of the offer. For such complex offer/response scenarios, a more sophisticated choice model and analysis of the effect of selection bias and imbalance would be required.

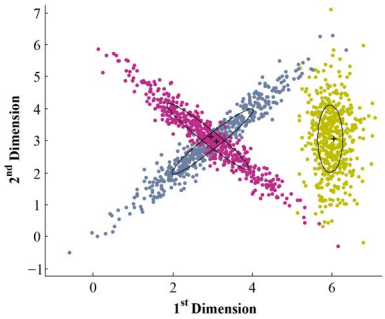


Fig. 2. Synthetic data for Experiment 1. Data were generated according to three 2D Gaussians. Ellipses show the unit-variance isocontours of each Gaussian.

4 Experiments

To empirically evaluate the predictive choice model (PCM), we performed two experiments, one on synthetic data and one on real admissions data from our institute.

4.1 Experiment 1: Synthetic Data

In Experiment 1, we performed an in-depth exploration of the properties of the proposed predictive optimization model on synthetic data. We generated two datasets, the first with three mixture components and the second with 18 mixture components. Our results on the three-component dataset led to some interesting hypotheses about the relative performance of PCM and standard methods, so we followed up on those hypotheses with the second, larger, dataset.

4.1.1 *Dataset 1: Three Components*

We started our experiments by generating two-dimensional synthetic data from a mixture of three Gaussian distributions. We generated 500 points according to each Gaussian component; the 1500 samples are shown in Fig. 2. For each sample i, we then generated a random decision d_i from a uniform distribution. For each component in the mixture model, we set the respective choice model parameters η_j and k_j to arbitrary positive values and then generated the probability of acceptance for each instance in the dataset using $f(d_i; \eta_j, k_j)$. Finally, we set the label y_i for each case to 1 if a uniformly distributed random value was less than $f(d_i; \eta_j, k_j)$; otherwise, we set $y_i = 0$. Each component in this three-component synthetic dataset has different characteristics, particularly in the balancedness of the response variable distribution, the choice model parameters η and k, and the amount of overlap between the mixture components.

After generating the data, we used them to estimate PCM with Algorithm 1. The PCM minimized the MDL cost at $J = 3$, which is the actual number of clusters the data were generated from. The Kullback-Leibler (KL) distance [14] between the true mixture model distribution and the best estimated mixture model distribution is 0.003, indicating a very strong match.

We fit PCM models to the synthetic data using both hard and soft assignment to mixture components. Then, for each of the samples i on that sample's training decision d_i, we

TABLE 2
Summary of Results on Synthetic Data in Experiment 1

Method	RMSE	
	(3 comp.)	(18 comp.)
PCM (Hard Assignment)	0.1020	0.0885
PCM (Soft Assignment)	0.0911	0.0805
LR	0.2264	0.1996
LMT	0.2143	0.1760

compared the predicted probability of acceptance to the actual probability. Since the actual positive response probabilities in the synthetic data are known, and our goal is to accurately predict these probabilities for each case over the range of the decision variable, we use root mean squared error (RMSE) to measure precisely how well the model accomplishes that goal. We use RMSE = $\sqrt{\sum_{i=1}^{N}(\hat{p}_i - p_i)^2/N}$, where $\hat{p}_i$ and p_i are the predicted and actual response probabilities, respectively, and N is total number of examples in the dataset (1500 in our case). For hard and soft assignment, we obtained RMSEs of 0.102 and 0.091, respectively, again indicating a very good match. These results indicate that the method is accurate at identifying the choice model parameters for each mixture model component. The additional error in predictive model based on hard assignment is due to overlap between distributions of different clusters.

How can we evaluate these results? For purposes of comparison, we followed a more standard approach using a logistic regression (LR) model and a logistic model tree (LMT) model [11]. LR models are able to provide explicit class probabilities with low variance but possibly high bias. The LR model we use in this paper is LogitBoost [10], which fits additive logistic regression models by maximum likelihood. A LMT model is a decision tree with logistic regression functions at the leaves. This combination is believed to provide a better bias-variance trade off than LR. LMT uses LogitBoost to fit the logistic regression models at the leaves, and it is also able to prune the generated tree to a simple logistic model when the simple logistic model provides better performance than LMT. We have also experimented with other methods such as naive Bayes, decision trees, and Bayesian networks, but these methods perform more poorly than LogitBoost and LMT.

We fit LR and LMT models to the same three-component synthetic data, then we compared the predicted and actual probabilities in the same way just described for the PCM models. The RMSE over all training samples by LR and LMT was 0.226 and 0.214 respectively, more than double the RMSE for the PCM. The comparison is summarized in the first results column of Table 2. It is clear from these numbers that our model is substantially better than LR and LMT at predicting acceptance probabilities, so long as the data actually conform to the proposed choice model. Put another way, incorporating domain knowledge in the form of the choice model into the predictive model appears to dramatically improve prediction accuracy.

Towards understanding the difference in performance between the PCM, LR, and LMT models, we performed a further analysis of how the characteristics of each component in the mixture model might be related to the level of estimation error made by each model. In Table 3, we show a set of features of each component in the mixture model and each predictive model's performance (measured by RMSE) over the data generated by that component. The features are the imbalance of the data with respect to the response variable Y, which we define as $(0.5 - P(Y = 1))^2$, the amount of overlap between components, and the choice model parameters.

TABLE 3
Component Features, Choice Model Parameters, and Predictive Models RMSE on the Three-Component Synthetic Dataset in Experiment 1

Comp.	Imbalance	η	k	RMSE		
				PCM	LR	LMT
1^{VS}	0.1225	0.15	8	0.080	0.146	0.146
2^{H}	0.1459	0.9	15	0.117	0.435	0.435
3^{H}	0.0001	0.5	5	0.092	0.055	0.055

^{VS}: Very small overlap, i.e. Less than 5%.
^{H}: Highly overlap , i.e. More than 30%.

The PCM is able to automatically identify the components with different choice parameters underlying the data set. The LMT and LR, on the other hand, do not have this capability. To make the comparative analysis more fair, we gave LMT and LR knowledge of the different clusters in the synthetic data set by manually training a separate model on each of the clusters.

The results in the table are provocative in that the LR and LMT models perform relatively well on the balanced, low-overlap component (component 3) but more poorly on the imbalanced components. It is also worth noting that the logistic models perform much better on the positive majority imbalanced component (component 1) than the negative majority imbalanced component (component 2). It is well known that imbalance hampers the performance of classifier learning algorithms [25], [29] and that imbalance is particularly troublesome when minority classes have higher error costs than other classes [26]. Our PCM, on the other hand, does not seem to exhibit this correlation, since the RMSE over the three components is 0.080, 0.117, and 0.092, respectively.

Although these results with three components are indicative of the factors that might contribute to the improved performance of the PCM over the traditional methods, our experiment has insufficient power to test for statistically significant correlations. In order to further test hypotheses about the relationship between mixture component characteristics and the relative performance of the predictive choice model, LR, and LMT, we repeated Experiment 1 on a data set containing a larger number of mixture components.

4.1.2 Dataset 2: 18 Components

We created a new synthetic data set similar to the three-component data set but with 18 components. As before, we generated 500 2D points per component, and we set the choice model parameters for each component to arbitrary values, with the goal of ensuring good coverage of the parameter space. As before, we first estimated the PCM from the synthetic data. Fig. 3 shows the log likelihood and MDL cost of the best model for each model order. We found that the MDL-based model order selection algorithm was again able to identify the true model order, i.e., 18. We

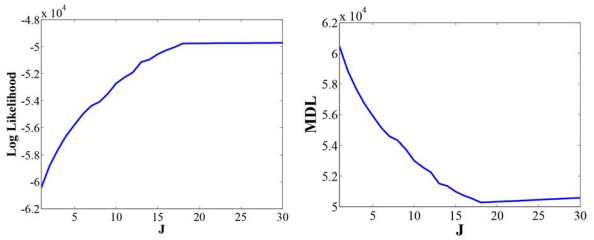


Fig. 3. Log likelihood and MDL cost as a function of J, for the 18-component synthetic dataset in Experiment 1. (a) Log likelihood. (b) MDL cost.

TABLE 4
Component Features, Choice Model Parameters, and Predictive Model RMSE on the 18-Component Synthetic Dataset in Experiment 1

Comp.	Imbalance	η	k	RMSE		
				PCM	LR	LMT
1^{S}	0.09986	0.22	5.0	0.123	0.151	0.151
2^{H}	0.11424	0.18	11.0	0.089	0.175	0.175
3^{H}	0.00078	0.41	7.0	0.106	0.077	0.077
4^{S}	0.00040	0.47	14.0	0.065	0.068	0.165
5^{M}	0.00006	0.44	10.0	0.090	0.068	0.068
6^{H}	0.12390	0.15	13.0	0.086	0.069	0.069
7^{H}	0.05290	0.27	15.0	0.138	0.049	0.107
8^{VS}	0.00384	0.62	5.0	0.056	0.052	0.052
9^{S}	0.10112	0.75	10.0	0.106	0.130	0.130
10^{VS}	0.00292	0.5	9.0	0.056	0.042	0.167
11^{S}	0.08644	0.82	8.0	0.056	0.380	0.392
12^{S}	0.14440	0.89	16.0	0.086	0.438	0.438
13^{S}	0.01166	0.37	13.0	0.021	0.054	0.054
14^{S}	0.01124	0.35	6.0	0.087	0.056	0.056
15^{VS}	0.02756	0.3	10.0	0.033	0.124	0.124
16^{VS}	0.08880	0.71	7.0	0.174	0.177	0.177
17^{VS}	0.10890	0.85	6.0	0.054	0.363	0.363
18^{VS}	0.01210	0.6	13.0	0.025	0.021	0.021

^{S} : Small overlap, i.e. $5-15\%$
^{H} : Highly overlap, i.e. $31-40\%$
^{M} : Moderate overlap, i.e. $16-25\%$
^{VS} : Very Small overlap, i.e. $0-4\%$

again built LR and LMT models using the same methodology, and as expected, the PCM once again outperformed LR and LMT models, as can be seen in the third column of Table 2.

In Table 4, we show the features, choice model parameters, and predictive model performance (RMSE) for the 18-component synthetic data. In order to identify which features each model's performance depends on, we performed a separate multiple linear regression for each model. For the PCM, we used component overlap (shown as superscript with component identification), columns 2 to 4 of Table 4 as independent variables and RMSE (column 5) as the dependent variable. For LR and LMT, we used columns 2, 3 and 4 of Table 4 as independent variables and RMSE (column 6 and 7) as the dependent variable (it is not possible for component overlap to affect these models, because as previously described, we train a separate model on each component). We include the predictive choice model parameters η and k in addition to the mixture component features because they are key factors in generating the response variable Y. The null hypothesis was that all regression coefficients are 0, and the alternative hypothesis was that one or more regression coefficients are non-zero.

TABLE 5
Regression Coefficients, p-Values, and 95% Confidence Intervals for the Predictive Choice Model (PCM) on the 18-Comopnent Synthetic Dataset of Experiment 1

Variable	Coefficient	p-value	95% Confidence Interval	
			Lower Bound	Upper Bound
Imbalance	0.294	0.262	−0.189	0.640
Overlap	0.474	0.150	−0.007	0.040
η	0.091	0.765	−0.093	0.124
k	−0.331	0.196	−0.010	0.002

TABLE 6
Regression Coefficients, p-Values, and 95% Confidence Intervals for the Logistic Regression (LR) Models on the 18-Component Synthetic Dataset of Experiment 1

Variable	Coefficient	p-value	95% Confidence Interval	
			Lower Bound	Upper Bound
Imbalance	0.602	0.001	0.735	2.196
η	0.503	0.003	0.109	0.429
k	−0.049	0.725	−0.012	0.009

TABLE 7
Regression Coefficients, p-Values, and 95% Confidence Intervals for the Logistic Model Trees (LMT) on the 18-Component Synthetic Dataset of Experiment 1

Variable	Coefficient	p-values	95% Confidence Interval	
			Lower Bound	Upper Bound
Imbalance	0.503	0.008	0.363	2.002
η	0.534	0.005	0.096	0.455
k	0.041	0.803	−0.011	0.013

Table 5 shows the estimated coefficients in the full multiple regression model for the PCM along with the significance of each coefficient and 95% confidence intervals for the coeffients. With a Type-I error probability fixed at $\alpha = 0.05$, we cannot reject the null hypothesis for the PCM. All of the coefficients' 95% confidence intervals contain 0 in their range. We therefore conclude that in this experiment, the predictive choice model is not significantly affected by the characteristics of the mixture components.

Tables 6 and 7 show the same data for the LR and LMT models. The r^2 values for the full multiple regression models based on the PCM, LR, and LMT algorihms' performance are 0.32, 0.744 and 0.579, respectively, indicating that the LR and LMT models strongly depend on the characteristics of the components they are being trained upon. For LR and LMT, we must reject the null hypothesis; in particular, the coefficient of the "imbalancedness" and "η" variables are significantly different from 0. We conclude that increased imbalance, in particular negative imbalance, leads to more error for LR and LMT models.

TABLE 8
Attributes Used for Predictive Optimization on Admissions Data

Attribute	Description
AGE	Age of the applicant at the time of application submission.
GPA	Standardized (to a 4.0 scale) grade point average from last degree program.
GNI	Gross national income of the applicant's home country, based on World Bank data.
IRANK	Previous university/institute ranking, on a scale of 0 to 10. IRANK reflects historical data on the difference between students' GPAs at the previous institute and their eventual GPAs at our institute.

TABLE 9
Results on Real Admission Data

Method	Enrollment Actual: 202.00	Revenue Actual: 140.96
PCM[a] + Hard Assignment	205.60	133.90
PCM[a] + Soft Assignment	200.31	134.98
LR[a]	188.92	130.86
LMT[a]	218.13	152.71
PCM[b]	256.79	172.43

[a] : Actual Offer (d_i)
[b] : Optimal Offer (d_i^*)

4.2 Experiment 2: Real Admissions Data

Next, we evaluated our model on real admissions data from our institute. It is a postgraduate research and education institute with about 2,000 master's and doctoral students. Approximately 1/3 of the applicants apply for open admission for postgraduate studies without any particular scholarship. With each offer of admission to applicants in this pool, the institute grants a variable amount of merit-based and need-based financial aid. In Experiment 2, we used admissions data from this pool of applicants to train and test our predictive optimization model. We first created a training set from the admissions data for 2005, 2006, and 2007. We then created a test set from the admissions data for 2008. We selected the four available attributes from the candidates' application records that are continuous or ordinal (shown in Table 8). The decision d_i for each individual i is the actual level of financial aid offered to the applicant in question, and the response y_i indicates whether the applicant accepted the offer and enrolled in the institute or rejected the offer. The training dataset contains 2852 instances, and the test set contains 803 instances. Note that we excluded applicants who applied with external scholarships or did not request financial aid.

We ran Algorithm 1 on the training data and found an optimal number of clusters $J^* = 20$. Using the 20-cluster model, we carried out either hard or soft cluster assignment for each applicant i in the test set. Then for each applicant i, we used the choice model $f(D_i; \eta_j, k_j)$ to either find the predicted optimal decision d_i^* or predict the probability $P(Y_i = 1 \mid \boldsymbol{x}_i, d_i)$ that the candidate would accept the actual historical offer d_i.

We evaluate the results in two ways. The main test of the model's accuracy is to compare, on the one hand, the predicted overall enrollment and revenue under the predictive model for the actual offers d_i to, on the other hand, the actual enrollment and revenue obtained historically with those offers. As another (unverifiable) evaluation of the approach, we also compute the enrollment and revenue the institute would obtain if it were to use the predicted optimal decisions rather than the actual historical decisions.

The results for both evaluations are shown in Table 9. The predicted enrollment according to the model based on hard and soft assignment is off by 1.8% and 0.8% respectively, whereas the predicted revenue is off by 5% and 4.2% respectively. As we would expect, the predicted revenue according to the optimal offer is substantially higher than the results achieved by the institute. In this case, the model predicts that the institute could have increased its revenue by 22% while increasing enrollment by 27%. Note that in other cases it might actually be possible to increase revenue without increasing enrollment.

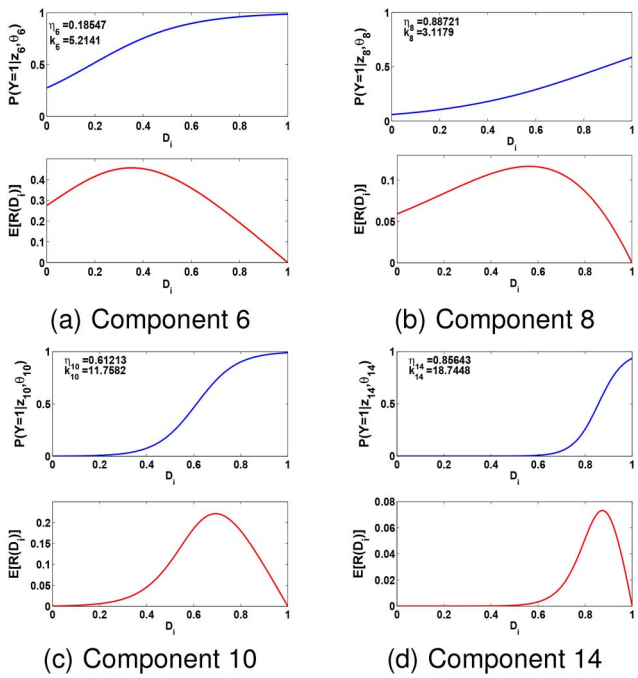


Fig. 4. Probability of positive response and respective expected revenue according to the choice model parameters estimates for different components.

As a baseline for comparison, we also built LR and LMT models for the same data. It is clear from Table 9 that the PCM outperforms both baseline models; the expected enrollment deviates respectively by 6.2% and 8% from the actual enrollment, and the expected revenue deviates by 7.2% and 8.3%, respectively, from the actual revenue.

In Fig. 4, we show the response probability curves for components 6, 8, 10 and 14. We see immediately that different groups of applicants have very different response curves. Choice model parameter η juxtaposes two reasons an applicant might require a higher fellowship offer before accepting: financial need and competitiveness with other institutes' financial aid packages. k characterizes how sensitive the applicant's choice behavior is; higher values yield steeper response curves. From Fig. 4, we see that the applicants mapped to components 6 and 8 have relatively low values for k and those mapped to components 10 and 14 have larger values, resulting in more sensitivity to the

offer. Increased sensitivity may indicate availability of other opportunities such as a job, a competing offer, or inability to pay even heavily discounted tuition. Examining the characteristics of the applicants in component 6 more closely, we find that they come from lower quality institutes (average IRANK 1.5) and from lower-middle income countries (average GNI 1580). Component 8 contains individuals from higher quality institutes (average IRANK 8.5), higher GPAs (average 3.3), and lower-middle income countries (average GNI 3854). Component 10 contains applicants with above-average GPAs (average 3.0) coming from good quality institutes (average IRANK 7.9) and from upper-middle income countries (average GNI 4880). Finally, component 14 contains individuals from lower-middle income countries (average GNI 1407) with higher GPAs (average 3.3) that studied at high quality institutes (average IRANK 8.4). The response curve shows that as one might expect, the last group of applicants is the most sensitive to the institute's financial aid offer. Clearly, the applicant response profiles, when combined with the characteristic attributes of the applicants in a particular cluster, give decision makers a great deal of insight into their applicants' behavior.

5 Discussion and Conclusion

We present a new probabilistic predictive choice model for decision optimization problems where the goal is to either maximize revenue or minimize cost. These problems are faced in many domains such as credit scoring, loan finance, direct marketing, and financial aid for education. In predictive optimization, the response probabilities are oftentimes critically important, since they are directly or indirectly used in decision making. It is very common in these types of applications that the positive response probabilities are monotonically increasing or decreasing functions of some control or decision variable. In some of the previous applied research in this area, classifiers are used as predictive models, model selection is based on optimizing some intermediate quantity such as classification error, and the selected model is then used for optimal decision making. This approach has important limitations: the models are trained and used in different implementation environments, customer responses in the training data are considered as deterministic, the models are sensitive to imbalance and selection bias, the models cannot encode common-sense monotonicity constraints, and despite being good for classification, such models are not guaranteed to produce accurate probability estimates. Unfortunately, besides these limitations, the available training data in these domains are often sparse and incomplete, making it difficult to learn a predictive optimization model effectively using these techniques. As one example, in the financial aid domain, the historical training data are only available for one level of the decision variable, and the other levels are missing (censored).

Our approach makes explicit the relationship between revenue and the customer's choice behavior. We use a mixture model to address the issue of sparse training data for the estimate of a customer's choice model, and we formulate an EM procedure for maximum likelihood estimation of the parameters of the probabilistic choice model. The output model, for a specific customer, is a *function* specifying the probability of a positive response for each level of the decision variable. We can then select the decision that maximizes expected revenue with respect to the estimated choice model. The model is also able to encode domain knowledge constraints, particularly monotonicity, naturally, and it is able to interpolate or extrapolate choice probabilities when training data is insufficient in some ranges of the decision variable. The mixture model gives us a principled way to model the individual choice behavior of each customer without having hundreds of examples of the response behavior of customers with the same profile. By assuming that a customer's cluster is a sufficient statistic for the customer's real attributes, learning the response probability function becomes tractable from a sparse training set. Finally, we show that the predictive choice model is asymptotically insensitive to feature bias and it's performance is not significantly affected by imbalance and censored data.

Our experiments on synthetic data show that the model is quite accurate in identifying the true number of mixture components from which the data are generated, estimating the parameters of each component's choice model, and predicting each individual's response probabilities. The predictive choice model substantially outperforms LR and LMT models, on both an individual example basis and on a per-component basis, even when the component information is provided to the LR and LMT models. The synthetic data experiments also show that our model is less sensitive than LR and LMT models to imbalance in the response variable.

When applied on real admissions data of graduate students who applied for financial aid from our institute, the model is able to accurately estimate expected enrollments and revenue on a test data set based on actual historical financial aid offers and ground truth responses. The predictive choice model dramatically outperforms LR and LMT models in this evaluation. The superior performance of the model justifies the approach of clustering historical data and learning tailored response models for each cluster of similar individuals. As a final, provocative, test, we found that the choice model predicts that the institute could have substantially increased revenue if it made financial aid offers optimally according to the model. While decision makers are expected to use other criteria besides revenue optimization to make financial aid decisions, still, by using the model, they have more insight into prospective students' response behavior. The model could in fact be helpful in increasing the quality of the student body, considering that attracting strong students may require more attractive offers, and it could also be used as a tool to maintain diversity in the student body, considering that targeted groups of students may require more financial help in order to accept admissions offers. Application-specific constraints such as the quality of the incoming class and target student-faculty ratios can be easily incorporated into the decision model. For example, [1] provide a greedy algorithm to find an optimal set of decisions (offers) subject to capacity constraints.

Finally, besides university admissions optimization, our model covers a wide variety of problems in related domains such as finance, education, and health. Many applications

at the intersection of machine learning and decision theory are ripe for further research.

ACKNOWLEDGMENTS

This work was supported by graduate fellowships from the University of Balochistan, the Higher Education Commission of Pakistan, and the Asian Institute of Technology, Thailand.

REFERENCES

[1] L. V. Thanh and P. Haddawy, "Deriving financial aid optimization models from admissions data," in *Proc. ASEE/IEEE Frontiers Education Conf.*, vol. 2. Milwaukee, WI, USA, 2007, pp. 7–12.
[2] D. Pardoe and P. Stone, "Bidding for customer orders in TAC SCM: A learning approach," in *Proc. 3rd Int. Joint Conf. AAMAS*. New York, NY, USA: Springer-Verlag, 2004, pp. 52–58.
[3] R. D. Lawrence, "A machine learning approach to optimal bid pricing," in *Computational Modeling and Problem Solving in the Networked World:(Interfaces in Computer Science Operations Research)*, H. K. Bhargava and N. Ye, Eds. Boston, MA, USA: Springer, 2003, pp. 97–118.
[4] C. Hueglin and F. Vannotti, "Data mining techniques to improve forecast accuracy in airline business," in *Proc. 7th ACM SIGKDD Int. Conf. KDD*, New York, NY, USA, 2001, pp. 438–442.
[5] P. Domingos and M. J. Pazzani, "Beyond independence: Conditions for the optimality of the simple Bayesian classifier," in *Proc. 13th Int. Conf. Machine Learning*, 1996, pp. 105–112.
[6] E. Altendorf, A. Restificar, and T. Dietterich, "Learning from sparse data by exploiting monotonicity constraints," in *Proc. 21st Annu. Conf. UAI*, Arlington, VA, USA, 2005, pp. 18–26.
[7] J. Heckman, "Sample selection bias as a specification error," *Econometrica*, vol. 47, no. 1, pp. 153–161, Jan. 1979.
[8] B. Zadrozny, "Learning and evaluating classifiers under sample selection bias," in *Proc. 21st Int. Conf. Machine Learning*, New York, NY, USA, 2004, p. 114.
[9] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Comput.*, vol. 3, pp. 79–87, Mar. 1991.
[10] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting," *Ann. Statist.*, vol. 38, no. 2, pp. 337–374, 2000.
[11] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," *Mach. Learn.*, vol. 59, no. 1–2, pp. 161–205, 2005.
[12] D. L. McFadden, "Economic choices," *Amer. Econ. Rev.*, vol. 91, no. 3, pp. 351–378, 2001.
[13] K. E. Train, *Discrete Choice Methods with Simulations*, 2nd ed. New York, NY, USA: Cambridge University Press, 2009.
[14] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
[15] D. L. McFadden and K. Train, "Mixed MNL models for discrete response," *J. Appl. Econom.*, vol. 15, no. 5, pp. 447–470, 2000.
[16] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood for incomplete data via the EM algorithm," *J. Roy. Statist. Soc., B*, vol. 39, no. 1, pp. 1–38, 1977.
[17] J. Rissanen, "Universal prior for integers and estimation by minimum description length," *Ann. Statist.*, vol. 11, no. 2, pp. 417–431, 1983.
[18] A. Corduneanu and C. M. Bishop, "Variational Bayesian model selection for mixture distributions," in *Proc. 8th Int. Conf. Artificial Intelligence Statistics*, 2001, pp. 27–34.
[19] M. J. Beal and Z. Ghahramani, "The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures," *Bayesian Statist.*, vol. 7, pp. 453–464, Jun. 2002.
[20] N. Ueda and R. Nakano, "Deterministic annealing EM algorithm," *Neural Netw.*, vol. 11, no. 2, pp. 271–282, 1998.
[21] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems," *Proc. IEEE*, vol. 86, no. 11, pp. 2210–2239, Nov. 1998.
[22] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert W function," *Adv. Comput. Math.*, vol. 5, no. 1, pp. 329–359, 1996.
[23] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, Oct. 2002.
[24] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *Proc. 14th Int. Conf. Machine Learning*, 1997, pp. 179–186.
[25] N. Japkowicz, "Learning from imbalanced dat sets: A comparison of various strategies," in *Proc. Amer. Assoc. Artif. Intell. (AAAI) Workshop Learn. Imbalanced Data Sets*, 2000, pp. 10–15.
[26] M. A. Maloof, "Learning when data sets are imbalanced and when costs are unequal and unknown," in *Proc. ICML Workshop Learning from Imbalanced Data Sets II*, 2003.
[27] H. Guo and H. L. Viktor, "Learning from imbalanced data sets with boosting and data generation: The databoost-IM approach," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 30–39, 2004.
[28] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sept. 2009.
[29] A. B. Owen, "Infinitely imbalanced logistic regression," *J. Mach. Learn. Res.*, vol. 8, pp. 761–773, May 2007.
[30] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 1995.
[31] W. Fan, I. Davidson, B. Zadrozny, and P. S. Yu, "An improved categorization of classifier's sensitivity on sample selection bias," in *Proc. 5th IEEE Int. Conf. Data Mining*, 2005, pp. 605–608.
[32] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 2nd ed. New York, NY, USA: Wiley, 2002.
[33] J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Scholkopf, "Correcting sample selection bias by unlabeled data," in *Advances in Neural Information Processing Systems*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA, USA: MIT Press, 2007, pp. 601–608.
[34] W. Fan and I. Davidson, "On sample selection bias and its efficient correction via model averaging and unlabeled examples," in *Proc. 2007 SIAM Int. Conf. Data Mining*, pp. 320–331.
[35] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

Waheed Noor received the Post Graduate Diploma in Computer Science (PGD-CS) in 2002 and the Master degree in Computer Science (MCS) from the University of Balochistan (UoB), Pakistan in 2004. In 2005, he joined the University of Balochistan as a Lecturer in the Department of Computer Science. He received a graduate fellowship from UoB, the Higher Education Commission (HEC) of Pakistan, and the Asian Institute of Technology (AIT) and received the Ph.D. degree in Computer Science from AIT in 2013. His research interests lie in machine learning, data mining and knowledge engineering, artificial intelligence, and decision-theoretic problem solving.

Matthew N. Dailey received the B.S. and M.S. degrees in Computer Science from North Carolina State University in 1992 and 1995, respectively, and the Ph.D. degree in Computer Science and Cognitive Science from the University of California, San Diego, in 2002. He spent two years as a Research Scientist with Vision Robotics Corporation, San Diego, CA, USA, and two years as a Lecturer in the Computer Science and Information Technology programs at Sirindhorn International Institute of Technology, Thammasat University, Thailand. In 2006, he joined the Computer Science and Information Management Department at the Asian Institute of Technology, Thailand, where he is now an Associate Professor. His research interests lie in machine learning, machine vision, robotics, systems security, and high performance computing. He is a senior member of the IEEE.

Peter Haddawy received a BA degree in Mathematics from Pomona College in 1981 and MSc and PhD degrees in Computer Science from the University of Illinois-Urbana in 1986 and 1991, respectively. He was a tenured Associate Professor in the Department of Electrical Engineering and Computer Science at the University of Wisconsin-Milwaukee, and Director of the Decision Systems and Artificial Intelligence Laboratory until 2002. Subsequently, he served as Professor of Computer Science and Information Management at the Asian Institute of Technology (AIT) through 2010 and as the Vice President for Academic Affairs from 2005 to 2010. He served as Director of UNU-IIST in Macau from 2010 through 2013. He is currently a faculty member in the Faculty of ICT at Mahidol University in Thailand. Professor Haddawy's research has concentrated on the use of decision-theoretic principles to build intelligent systems, and he has done pioneering work in the areas of decision-theoretic planning and probability logic. His current research interests include decision-theoretic problem solving, intelligent medical training systems, and bibliometric techniques supporting evidence-based research policy.