Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computational Statistics & Data Analysis 47 (2004) 775–790
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Isotonic single-index model for high-dimensional database marketing

Prasad A. Naik, Chih-Ling Tsai*

Graduate School of Management, University of California at Davis, Davis, CA 95616-8609, USA

Received 1 August 2003; accepted 19 November 2003

Abstract

While database marketers collect vast amounts of customer transaction data, its utilization to improve marketing decisions presents problems. Marketers seek to extract relevant information from large databases by identifying significant variables and prospective customers. In small databases, they could calibrate logistic regression models via maximum-likelihood methods to determine significant variables and assess customer's response probability. For large databases, however, this approach becomes computationally too intensive to implement in real-time, and so marketers prefer estimation methods that are *scalable* to high-dimensional databases. In addition, database marketing is practiced in diverse product-markets, and so marketers prefer probability models that are *flexible* rather than restrict to specific distributions (e.g., logistic).

To incorporate scalability and flexibility, we propose isotonic single-index models for database marketing. It furnishes the first projective approximation to a general p-variate function. Its link function is order-preserving (i.e., isotonic), thus encompassing all proper distribution functions. We develop a direct approach for its estimation: we first estimate the orientation of high-dimensional parameter vector without specifying the link function (via sliced inverse regression), and then estimate the non-decreasing link function (via isotonic regression). We illustrate its practical use by analyzing a high-dimensional customer transaction database. This approach yields dimension reduction both column- and row-wise; that is, we not only discover significant variables in a large transaction database, but also prioritize customers into a few distinct groups based on estimated response probability (to enable direct mailing of catalogs).

Keywords: Customer selection; Direct marketing; Large datasets; Dimension-reduction; Logistic regression; Constrained non-parametric estimation; Single-index models

* Corresponding author. Tel.: +1-5307528565; fax: +1-5307522924.
E-mail address: cltsai@ucdavis.edu (C.-L. Tsai).

0167-9473/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2003.11.023

1. Introduction

Direct marketing is one of the oldest techniques for marketing products, and is growing at twice the growth rate of the US economy (Statistical Fact Book, 2000, p. 254). Direct marketers collect detailed information on customers' purchase histories and combine it with other information (e.g., financial and geo-demographic data) to create high-dimensional transaction databases, which contain hundreds of explanatory variables. The analysis of large transaction databases is a challenging task for two reasons.

First, direct marketers traditionally apply the maximum-likelihood method to fit logistic regression models using a few variables such as *recency* (when did the customer last buy), *frequency* (how often does the customer buy) and *monetary value* (the dollar value of items purchased). Given hundreds of explanatory variables, however, marketers find the maximum-likelihood method computationally intensive. For example, Balasubramanian et al. (1998, p. 320) state: "... the sheer size of available data places severe demands on computing power... This has revived interest in speeding up traditional numerical and statistical estimation methods..." In other words, marketers need a *scalable approach* to tackle high-dimensionality of customer transaction databases. By scalability, we mean that a method applies to datasets ranging in size from small to large without a corresponding increase in computational effort (Berry and Linoff, 1997, p. 373, 426).

Second, database marketing is practiced across diverse product markets such as books, music and video, auto parts and accessories, gardening, jewelry, sports and outdoors, and stationery (Statistical Fact Book, 2000, p. 93). Customer response probability may be predicted well by logistic regression in some—but not every—product market. Hence, marketers need a *flexible approach* that encompass a broad class of distribution functions rather than restrict to the logistic distribution. By flexibility, we mean that a method can estimate response probability without making specific parametric assumptions.

To incorporate scalability and flexibility, we propose the isotonic single-index model to analyze high-dimensional customer transaction databases. Single-index models are useful in high-dimensional data analyses because they serve as the first projective approximation to a general p-variate function $f(x_1,\ldots,x_p)$. Specifically, the standard single-index model (see, e.g., Brillinger, 1983; Stoker, 1986; Horowitz, 1998; Naik and Tsai, 2000, 2001) is given by $E[y]=F(x'\beta)$, where y denotes the dependent variable, $F(\cdot)$ represents the link function, x includes the p explanatory variables, and β is a conformable parameter vector. In database marketing, we expect the number of variables to be large; that is, $p \approx 100$. In standard single-index models, the link function $F(\cdot)$ is neither known, nor restricted to any specific shape. However, for database marketing, the response variable is binary (i.e., $y \in \{0,1\}$), and so $F(\cdot)$ is a non-decreasing function with $0 < F(z) < 1$ for all $z = x'\beta$. Formally, we let $F(\cdot)$ belong to the function space Φ, which includes all proper distribution functions. Consequently, parametric probability models are special cases; for example, we get the logistic regression when $F(z) = e^z/(1+e^z)$, and we get the probit model when $F(z) = \int_{-\infty}^{z} \phi(t)\,\mathrm{d}t$, where ϕ denotes the standard normal density function.

To calibrate isotonic single-index models, we need to estimate $F(\cdot)$ and β in the joint space Γ: $\Phi \times \Re^p$. Two existing approaches include iterative methods (e.g., Cosslett, 1983) and non-iterative methods (e.g., Powell et al., 1989; Ichimura, 1993), which we briefly review and then outline our direct approach for estimating isotonic single-index models.

Consider first the iterative methods to obtain the estimates of F and β by maximizing the likelihood function $L(F, \beta)$ as follows. For a given $\beta^{(k)}$ in the kth iteration, we maximize the functional $L_1(F) = L(F, \beta^{(k)})$ with respect to a distribution function $F(\cdot)$. Then, using the conditional solution $\hat{F}^{(k)}(\cdot \mid \beta)$, we maximize the concentrated likelihood function $L_2(\beta) = L(\hat{F}^{(k)}(\cdot \mid \beta),\ \beta)$ with respect to β. As Cosslett noted (1983, p. 774), the concentrated likelihood L_2 is a step function over the set $\{\beta\}$, and so this discontinuity rules out the use of gradient-based optimization methods. One has to select a random set of orthogonal directions in p-dimensional space, and then perform a linear search for a maximum along each of these directions (e.g., via simulated annealing). Using the resulting β^{k+1}, we re-iterate this procedure and obtain the sequence $\{(\beta^{(0)}, \hat{F}^{(0)}),\ (\beta^{(1)}, \hat{F}^{(1)}), \ldots, (\beta^{(k)},\ \hat{F}^{(k)}), \ldots\}$, which eventually converges to $(\beta^*, \hat{F}^*)$. Thus, because $L_2(\beta)$ is not differentiable, this iterative method requires substantial computational effort when the number of variables $p \approx 100$. It possesses an additional drawback: as Bult and Wansbeek (1995, p. 388) note, "... the asymptotic distribution is unknown. Therefore, we are not able to give standard errors ... not able to test whether the coefficients ... are significantly different ..."

Consider now the non-iterative methods for estimating β without knowing F (e.g., Powell et al., 1989). These methods require non-parametric density estimation to obtain the derivatives of the joint density of explanatory variables. As dimensionality increases, however, the estimation of joint density function encounters severe computational difficulties (Silverman, 1986, Chapter 4). This curse of dimensionality induces an "empty space phenomenon" in which local neighborhoods are almost surely empty, and neighborhoods that are not empty are almost surely not local (Simonoff, 1996, p. 101). Consequently, non-parametric density estimation is usually not applied to datasets with five or more dimensions. Hence, non-iterative approaches are not easily scalable for database marketing with hundreds of explanatory variables.

To overcome these limitations, we develop a direct approach to estimate F and β by ensuring monotonicity of F without requiring kernel and bandwidth selections. We first obtain $\hat{\beta}$ without the knowledge of $F(\cdot)$ via sliced inverse regression (Li, 1991; Duan and Li, 1991). We then compute the scalar index $z = x'\hat{\beta}$, and apply isotonic regression (Barlow et al., 1972, p. 13) to estimate the non-decreasing $F(\cdot)$ over the range (0, 1). Statistical theory assures that $\hat{\beta}$ is a root-n consistent estimator of β even if F is not known (Li, 1991; Hsing and Carroll, 1992), and that isotonic regression yields the consistent maximum-likelihood estimator of F belonging to proper distribution functions in Φ (Ayer et al., 1955; Kiefer and Wofowitz, 1956; Wald, 1949).

Our direct approach is both scalable and flexible. Because the estimation of β does not depend on $F(\cdot)$, this approach dramatically reduces computational effort compared to iterative methods such as Cosslett's. In contrast to non-iterative methods (e.g., Powell et al., 1989), we do not need to estimate a non-parametric density function in p dimensions (a non-trivial task for $p \approx 100$). Rather, we solve an eigenvalue problem

to obtain parameter estimates and standard errors rapidly, thus making this approach scalable to large databases. Moreover, because $F(\cdot)$ is a non-parametric function, we effectively characterize customer response probability in diverse product-markets via its flexible shape.

We illustrate the application of isotonic single-index models by analyzing real data from a catalog company. Specifically, we combine three customer databases: purchase transactions, credit-history and geo-demographic data. The resulting dataset contains 2424 customers, 166 explanatory variables, and one binary response variable. We find that only 16 of the 166 variables are significant in this application. In addition, individual customers are classified into 15 distinct groups based on estimated response probability, prioritizing customers for mailing catalogs. Thus, we attain column- and row-wise dimension reduction: for a database with N customers and p variables, we find p^* significant variables ($p^* < p$) and M^* customer groups ($M^* < N$). Finally, single-index models represent new statistical developments, and are not applied yet in marketing science (Leeflang et al., 2000, p. 403; Naik and Raman, 2003, p. 385). This study not only marks its first application, but also enables database marketers to capture asymmetry, heavy-tails or multiple modes in density functions underlying the flexible distributions.

We organize this paper as follows. In Section 2, we present an overview of database marketing and describe the transaction database used in the empirical application. In Section 3, we state the database-marketing problem, formulate the isotonic single-index model, and develop the direct estimation approach. Section 4 discusses the findings from our empirical application, and Section 5 concludes by identifying research topics to further improve database-marketing practice.

2. Customer transaction databases

Here we present an overview of database marketing, and describe the high-dimensional database.

2.1. Overview of database marketing operations and extant literature

Database marketing operations consist of four distinct stages: data warehousing, modeling, optimization, and campaign execution. *Data warehousing* refers to storing "information inventory" on customer name, address, demographics, previous contacts, purchase history, payments, credit, product returns, phone and email contacts, and click-stream patterns.

The *modeling* stage involves statistical analyses of these data to gain insights into customer retention, customer segmentation, and customer propensity. To understand customer retention, marketers build statistical models to predict when customers might become inactive or switch suppliers (e.g., Schmittlein et al., 1987). To discover customer segments, marketers use finite mixture models and cluster analysis techniques (e.g., Wedel and Kamakura, 2000). To analyze customer propensity, marketers typically use logistic regression models (Hughes, 1996).

The third stage, *optimization*, combines the output from the modeling stage (e.g., estimated propensity scores) with information on costs and prices, imposes constraints on budget, timing and other resources, and recommends the optimal course of action (see, e.g., Bitran and Mondschein, 1996; Gonul and Shi, 1998). In the final stage, *campaign execution*, the marketer implements the recommended action, mails personalized offers to selected customers, collects feedback on their response (e.g., purchased or not, product returns, dissatisfaction) and updates the transaction databases with this event information.

In practice, a firm's performance depends on the optimization and campaign execution stages, which require inputs on those few variables that influence customer's purchase and those prospective customers who should receive the next direct mail. To this end, the flexible and scalable approach plays an important role in high-dimensional database marketing.

2.2. High-dimensional database

We consider the case of a catalog marketer who mailed catalogs with product information to its customers and maintained detailed records of purchase transactions for 12 years. These data are in pubic domain (see www.the-dma.org) and available for academic research from Direct Marketing Educational Foundation (data 03DMEF). Based on this transaction data, the marketer wants to determine which customers should receive a promotional catalog. We merge the credit data (98DMEF) and geo-demographic data (99DMEF) with this transaction data using postal ZIP code, and the resulting file contains 2424 customers with no missing values and 166 explanatory variables, which are described below.

Purchase transaction data: Six broad types of variables based on purchase transaction information are constructed. The first is *recency* of purchase, based on the time elapsed since a customer last bought the product. We create additional recency measures, such as the number of years since the last purchase, across all product classes. Secondly, marketers construct *frequency* variables based on how often a customer is contacted or buys a product. For example, one of the frequency measures is the lifetime number of contacts for each promotion. We augment the frequency measures by computing the average number of orders placed over the last 12 years. Thirdly, marketers measure the *monetary value* of purchase based on how much a customer spends. For example, they use the total sales per customer in the last year, 2 years ago, and so forth. We include further measures of monetary value by summing over a 4-year span for each promotion type.

In addition to the above three types of measures, the catalog marketer records *product category*, *transaction mode*, and *conversion purchase*. To plan and stock the product assortment for a season, the marketer tracks the product category from which a customer buys. They also record the transaction mode: whether the order was placed on phone, and whether it was paid for by credit card, house card, entertainment card, or cash. Finally, conversion purchase is the first purchase by a customer. The catalog marketer deems conversion purchase as special, and so maintains information on whether the first order was placed by telephone, whether it was on promotion, the product

category, and the dollar size. We obtained a total of 97 variables from the transaction data.

Credit data: Credit information is classified into five types of variables: *age*, *number of credit lines*, *balance amount*, *credit limit*, and *delinquency status*. The age variable provides information such as the age of the oldest or youngest credit card. The variables under "number of credit lines" indicate the number of bankcards, retail cards, and tradelines from financial institutions. The variables under "balance amount" show the average balance on open and active credit cards or the loan amount on automobiles. The credit limit variable gives information on the average credit limit for various financial instruments (e.g., bank cards, retail cards). Finally, the delinquency variable contains information such as the number of tradelines 30 days late, 60 days late, 90 days late, or seriously delinquent. Although financial information is available at the household level, a catalog marketer gets this financial data aggregated at the ZIP code level because of legal requirements for protecting individual privacy. A total of 38 variables are obtained from the credit data.

Geo-demographic data: This information is classified into five types of variables: *wealth*, *ethnic composition*, *family composition*, *neighborhood mix*, and *educational attainment*. The wealth variable consists of information such as wealth and income ratings, home value index, and current estimated median income. Ethnic composition indicates the percentage of households occupied by Whites, Blacks, or Hispanics. Family composition provides information on the number of persons per household, median age of the householder, and the percentage of households with children under 18, or the percentage of households who have families. The variables under "neighborhood mix" describe the percentage of houses occupied by owners or renters and the age of housing units. Finally, the variables under "educational attainment" measure aspects such as median years of school for people 25 years and older, or the percentage of households with a BA degree or more. The geo-demographic data furnishes an additional 31 variables.

Overall, the merged customer database consists of 166 explanatory variables. Many explanatory variables exhibit skewed and non-normal distributions, which are permissible in the proposed approach. Customer response—whether or not s/he buys the product—is the binary-dependent variable. We expect several explanatory variables to be irrelevant, whose presence not only diminish the precision of the estimated effects of relevant variables and the forecasts of response probability (Altham, 1984), but also detract managers from focusing their marketing effort on those few variables that actually drive customer response. Thus, to facilitate the analyses of high-dimensional databases, we next describe the isotonic single-index model and its estimation.

3. Isotonic single-index model

Here we describe the database marketing problem, formulate the isotonic single-index model, and propose a direct approach to estimate it.

3.1. Database marketing problem

Database marketers observe whether or not a customer responds to an offer, and record the transaction history of the customer's behavior. Let y_i denote the binary response of customer i to buy or not buy a product, let x_i be a p-dimensional vector of variables in the transaction database, and let β be the impact of these variables on customer response. As in Bult and Wansbeek (1995), the response of customer i is

$$y_i = \begin{cases} 1 & \text{if } x_i'\beta + \varepsilon_i > \alpha, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where the error term ε_i results from the unobservable characteristics of the ith customer, and α is the threshold for purchase action. The error term ε_i is a random variable with the distribution function $G_\varepsilon(\cdot)$, is independent of x_i, and is independently and identically distributed for $i = 1, \ldots, N$. The explanatory variables x can be non-normal, belonging to the broad class of elliptical distributions.

Eq. (1) indicates that customer i buys the product if $(x_i'\beta + \varepsilon_i)$ exceeds the threshold α. As in Bult and Wansbeek (1995), we denote $z_i = x_i'\beta$ as an index for the customer i. Thus, the likelihood that a customer buys the product, i.e., the *customer propensity*, is

$$P(y = 1) = P(\varepsilon > \alpha - z) = 1 - G_\varepsilon(\alpha - z). \tag{2}$$

We can compute customer propensity if the index z, the threshold α, and the distribution function $G_\varepsilon(\cdot)$ are known. However, as Cosslett (1983, p. 766) notes, we do not possess the knowledge to precisely specify $G_\varepsilon(\cdot)$ because ε is unobservable, and therefore $G_\varepsilon(\cdot)$ is assumed to be an *unknown* non-decreasing function with the range (0, 1). For identification, we assume that α is zero as in Bult and Wansbeek (1995, p. 393). As a result, Eq. (2) naturally becomes the *isotonic single-index model*:

$$E[y] = P(y = 1) = 1 - G_\varepsilon(-z) = F(z) = F(x'\beta), \tag{3}$$

where the unknown $F(\cdot) \in \Phi : \Re \to [0, 1]$, the space of distribution functions, and $\beta \in \Re^p$ with $\|\beta\| = 1$ for identification. The term *isotonic* means that F is "order-preserving" on the index set z (see Barlow et al., 1972, p. vi.), while *single-index* refers to the scalar z, which is a linear combination of x-variables (see, e.g., Naik and Tsai, 2000). Furthermore, because the link function belongs to the space of distribution functions $\Phi : \Re \to [0, 1]$, the underlying density function can exhibit asymmetry or heavy-tails or multiple modes. In other words, the proposed model does not rule out these possibilities as would be the case with logistic or probit regressions.

The marketing problem, then, is to determine (a) the parsimonious set of variables that affect customer response, and (b) the prospective set of customers for whom to initiate a marketing contact. Thus, we need to solve a dimension-reduction problem both column- and row-wise directions. In symbolic terms, for a given transaction database of dimension $N \times p$, we seek to find p^* significant variables ($p^* < p$) and M^* customer groups ($M^* < N$). Note that customers in a given group m ($m = 1, \ldots, M^*$) have the same predicted response probability. Next, we propose a direct approach for solving this dual dimension-reduction problem.

3.2. Direct approach

We estimate (F, β) in the space Γ: $\Phi \times \Re^p$ *without* iteration between Φ and $\Re^p$. To this end, we first estimate β without knowing F via sliced inverse regression (Li, 1991), and then estimate F via isotonic regression (Barlow et al., 1972).

Estimating β: Sliced inverse regression estimates the direction of β by the principal eigenvector γ_1 of the generalized eigenvalue decomposition

$$\Sigma_\eta \gamma_1 = \lambda_1 \Sigma_x \gamma_1, \tag{4}$$

where λ_1 is the largest eigenvalue, Σ_x denotes the covariance matrix, and $\Sigma_\eta = \mathrm{Cov}(E(x|y))$ is the covariance of the conditional means of x given y. To obtain $\hat{\beta}$, let Y denote the binary vector of customer response of dimension $N \times 1$, and let X be the matrix of explanatory variables of dimension $N \times p$. We estimate Σ_x by the usual covariance matrix

$$\hat{\Sigma}_x = \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})(X_i - \bar{X})', \tag{5}$$

where X_i' denotes a row for the ith customer, and $\bar{X}$ contains the means of variables across all customers. To estimate Σ_η, we partition X into two slices because the dependent variable is binary. Let X_0 denote a sub-matrix of X for $y_i = 0$, and X_1 be the sub-matrix of X for $y_i = 1$. We compute the means of variables in X_0 and X_1 over their respective customers, and denote them $\bar{X}_0$ and $\bar{X}_1$. The weighted average across the two slices yields the covariance matrix

$$\hat{\Sigma}_\eta = \sum_{h=0}^{1} \hat{p}_h (\bar{X}_h - \bar{X})(\bar{X}_h - \bar{X})', \tag{6}$$

where $\hat{p}_h$ is the proportion of customers in slice h, $h = 0$ and 1. Then we obtain $\hat{\beta} = \hat{\gamma}_1$ by substituting the estimates $\hat{\Sigma}_x$ and $\hat{\Sigma}_\eta$ into (4). We note that $\hat{\beta}$ is not overly sensitive to the assumption of elliptical distributions (Li, 1991, p. 338; Cook and Nachtsheim, 1994, p. 592). Furthermore, we compute the standard errors of $\hat{\beta}$ (see Chen and Li, 1998, p. 297; Duan and Li, 1991, Section 4), which are given by the squared root of the diagonal of the matrix:

$$\frac{1 - \hat{\lambda}_1}{\hat{\lambda}_1} N^{-1} \hat{\Sigma}_x^{-1}. \tag{7}$$

Using the parameter estimates and standard errors, we can compute the t-values to determine the p^* relevant variables at the 95% confidence level.

It is noteworthy that we can conduct parameter estimation and statistical inference without specifying G_ε a priori (as in parametric models), without estimating the distribution function F (e.g., Cosslett, 1983), and without estimating the derivatives of the joint density function of X (e.g., Powell et al., 1989).

Estimating F: The log-likelihood function for the binary response model (3) is given by

$$L(F) = \sum_{i=1}^{N} \{y_i \log[F(z_i)] + (1 - y_i) \log[1 - F(z_i)]\}. \tag{8}$$

For F to be a proper distribution function, we need to impose the constraints:

$$0 < F(z_j) \leqslant F(z_k) < 1 \tag{9}$$

for $z_j \leqslant z_k$. To solve the maximization problem posed by (8) and (9), we compute the index $z_i = X_i' \hat{\beta}$ for each customer i, and then sort them in ascending order so that

$$z_{(1)} \leqslant z_{(2)} \leqslant \cdots \leqslant z_{(N)},$$

where $z_{(i)}$ is the ith ordered statistic for the sorted customer i. The corresponding binary response is denoted by $y_{(i)}$. Next, we adopt the pooled adjacent violator algorithm from Ayer et al. (1955), which is briefly described below.

When $y_{(i)}=1$, the term $\log[F(z_{(i)})]$ in (8) is the largest possible if $F(z_{(i)})=F(z_{(i+1)})$, given the constraint in (9). By contrast, when $y_{(i)}=0$, the term $\log[1-F(z_{(i)})]$ is the largest possible if $F(z_{(i)})=F(z_{(i-1)})$. Consequently, the estimator $\hat{F}$ remains constant in any *decreasing run* of the ordered sequence $\{y_{(i)}\}$ with 1's followed by 0's. Therefore, we will construct customer groups that exhibit such decreasing runs. A decreasing run ends and the next one begins if, and only if, the sequence $\{01\}$ appears. We label the resulting customer groups as $m = 1, 2, \ldots, M$, and define the frequency

$$f_m = \frac{\#(1\text{s})}{\#(m)}, \tag{10}$$

where $\#(1\text{s})$ denotes the number of 1's, and $\#(m)$ is the total number of customers in group m. In case of ties, namely $z_{(i)} = z_{(i+1)}$, we keep these customers in the same group, arranging 1's before 0's. If $f_m < f_{m+1}$ holds for all $m = 1, \ldots, M-1$, then the sequence $\{f_m\}$ is the desired empirical distribution function, $\hat{F}$. If not, we combine—i.e., pool—the adjacent groups for which $f_m \geqslant f_{m+1}$, namely the groups that violate this non-decreasing order, and compute a new f_m value for the merged groups using Eq. (10), and re-label the resulting groups. We repeat this process until we obtain the non-decreasing sequence $\{f_m^*\}$, $m=1,\ldots,M^*$. Then the maximum-likelihood estimator of F is

$$\hat{F}(z_i) = f^*_{m(i)}, \tag{11}$$

where $m(i)$ denotes the group m that contains customer i, and the total number of groups is given by M^*. The subsequent remarks elaborate the properties of the direct approach, its relation to other approaches, and potential extensions.

Remark 1. Table 1 summarizes the algorithm to estimate isotonic single-index models. The theory of sliced inverse regression (e.g., Li, 1991; Hsing and Carroll, 1992) assures that $\hat{\beta}$ is a root-n consistent estimator of the direction of β even though the link function F is not known. It has been used previously to calibrate single-index models (Duan and Li, 1991; Naik and Tsai, 2000, 2001). As for the maximum-likelihood estimator

Table 1
Estimation algorithm for isotonic single-index models

1.	Slice the X matrix into X_0 and X_1. The sub-matrix X_0 contains customers with $y=0$, and X_1 contains those with $y=1$.
2.	Compute $\hat{\Sigma}_x$ and $\hat{\Sigma}_\eta$ by using (5) and (6), respectively.
3.	Solve the eigenvalue problem in (4). The principal eigenvector provides the parameter estimates $\hat{\beta}$.
4.	Compute standard errors of $\hat{\beta}$ by using (7) for statistical inference.
5.	Compute the single-index $z_i = X_i'\hat{\beta}$ for each customer i.
6.	Sort (y_i, z_i) in an ascending order of z values. If z values are tied, then keep the customers with $y_i = 1$ before those with $y_i = 0$.
7.	Group the customers in a decreasing run of the ordered sequence $\{y_j\}$. A group boundary exists between two customers $(i, i+1)$ if and only if $y_i = 0$ and $y_{i+1} = 1$.
8.	Compute f_m by using (10) for each customer group m.
9.	Merge customer groups m and $(m+1)$ if $f_m \geqslant f_{m+1}$. Compute f_m for this merged group, and re-label the remaining groups.
10.	Repeat step 9 until the sequence $\{f_m\}$ is monotonic in m.

of F, Kiefer and Wofowitz (1956) and Wald (1949) furnish the regularity conditions, and Cosslett (1983) verifies that they hold for model (3) to establish the consistency of $\hat{F}$ in (11).

Remark 2. We note the similarity and distinctions with other approaches. Specifically, Carroll et al. (1997) propose the generalized partially linear single-index model, and develop an iterative estimation approach. However, their approach is computationally intensive for high-dimensional data; and their link function is not order-preserving, which is a necessary property for probability models. In contrast, Gifi (1990, p. 370) develops a monotonic splines approach to fit binary variables; but it estimates a smoothed curve $\tilde{F}(\cdot)$, and so marketers cannot prioritize customers into M^* distinct groups for mailing catalogs.

Remark 3. We emphasize that the isotonic single-index model, akin to other probability models (e.g., logistic or probit regression), is "intimately related and yet conceptually quite different" (Cox and Snell, 1989, p. 132) from discriminant analysis. Conceptually, in discriminant analysis, $Y = 1$ or 0 represents two *distinct* populations (e.g., two different species of bacteria or two different kinds of plants). By contrast, in probability models, there is one population whose members respond differently depending on the level of covariates. See Cox and Snell (1989, p. 132) for further elaboration. Formally, the statistical properties of inverse regression differ from those for discriminant analysis. For example, inverse regression theory does not require the explanatory variables to follow a multivariate normal distribution, whose violation in discriminant analysis affects the significance tests and classification rates (Sharma, 1996, p. 332). See Li (2000, Chapter 14) for further differences and connections between inverse regression and discriminant analysis.

Remark 4. If managers expect latent segments in the market (Gonul et al., 2000), then they can extend the isotonic single-index model to incorporate unobserved heterogeneity. One approach is to conduct cluster analysis (say, via K-means algorithm), and then apply the isotonic single-index model in each segment. Alternatively, we consider its extension to *mixture isotonic single-index model*, $E[y_i] = \sum_{s=1}^{S} \pi_{is} F_s(x_i'\beta_s)$, where each customer i belongs to the latent segment $s = 1,\ldots,S$ with probability π_{is}, and $\sum_{s=1}^{S} \pi_{is} = 1$ for all i (see McLachlan and Peel, 2000; Wedel and Kamakura, 2000). The total number of segments S can be determined via selection criteria (e.g., Akaike information criterion, McLachlan and Peel, 2000).

Remark 5. We note that SIR extracts at most one index when response variable is binary. To study interactions between indexes, however, we need multiple indexes $z_l = x'\beta_l$, $l = 1,\ldots,L$, which can be retrieved via sliced average variance estimation or difference of covariance estimation (see Cook and Weisberg, 1991; Cook and Lee, 1999). Next, we apply the isotonic single-index model to analyze the high-dimensional database described in Section 2.

4. Empirical results

We first present parameter estimates to assess the impact of explanatory variables, and then provide probability estimates to characterize customer response.

4.1. Parameter estimates

Table 2 presents the SIR estimates with the absolute t-values greater than 1.96 (the 95% confidence level). They measure the relative importance of variables influencing the customer response. We find that customers' propensity to buy decreases as *recency* increases, a finding consistent with previous studies (e.g., Gonul and Shi, 1998). That is, customers are less likely to buy as time elapsed since the last purchase increases. The effect of *frequency* shows that potential customers who received more promotional contacts are less likely to buy. This finding indicates that a customer's decision to respond (or not) is based on the first few catalogs, and additional catalogs are less likely to change the initial decision. As expected, the *monetary value* positively influences customer propensity. In other words, the more money spent over the past 2 years, the more likely a customer buys from this product catalog. In addition, the *product category* effects suggest that products featured in the promotion 08 enhance customer propensity, but products in the promotion 85 or the class 5 do not. Such findings help catalog marketers to stock and promote appropriate products and devise marketing programs for cross-selling related merchandise.

Interestingly, *conversion purchase* on promotion decreases customer propensity. This finding is consistent with attribution theory (Dodson et al., 1978), which suggests that when consumers attribute their purchase to promotional incentives rather than to the merits of a product, the probability of subsequent purchases decreases. Hence, marketers

Table 2
Parameter estimates and t-values from sliced inverse regression

Variables	$\hat{\beta}$	t-values
Recency		
Years since the last order across all product classes	−0.7084	−3.06
Frequency		
Lifetime number of promotion 50 contacts	−0.2402	−5.80
Lifetime number of promotion 65 contacts	−0.1691	−2.53
Monetary value		
Dollar class of lifetime average order size	0.3508	3.04
Total sales last year	0.0142	3.18
Total sales 2 years ago	0.0175	3.90
Product category effects		
Lifetime orders placed in promotion 85	−0.5851	−2.16
Lifetime orders placed in promotion 08	1.3322	2.98
Sum of last 5 years sales in product class 5	−0.0169	−2.68
Conversion purchase		
First order's promotion	−0.0049	−2.56
Age		
Age of the youngest tradelines	−0.1232	−2.82
Average age of all active and paid tradelines	0.1324	2.83
Average age of all open and active tradelines	−0.0919	−2.50
Balance amount		
Average balance of all open and active credit cards	0.0043	2.00
Credit limit		
Avg. credit limit for all open and active credit cards	−0.0018	−2.30
Delinquency status		
Number of all tradelines once 30 days late	−3.0633	−2.20

should be cautious when assessing the value of customer acquisition via promotional incentives.

Another interesting result is the role of financial variables in customer behavior. From Table 2, we see that *age* of tradelines negatively affects customer propensity. This finding indicates that customers with younger tradelines are financially attractive. Moreover, the propensity to buy is positively related to *balance amount* and negatively related to *credit limit*. This means that customers with a large ratio of balance amount to credit limit are more likely to buy. As expected, poor *delinquency status* negatively impacts customer propensity.

Finally, Table 2 reveals that only 16 variables are significant at the 95% confidence level. The remaining variables, which comprise about 90% of all variables such as

Table 3
Probability estimates from isotonic regression

Group nos., m	Group boundary	Nos. of customers	Response probability, $\hat{F}$
0	$-\infty < z \leqslant 3.51$	0	0
1	$-3.51 < z \leqslant -1.94$	20	0.0000
2	$-1.94 < z \leqslant -0.26$	992	0.0272
3	$-0.26 < z \leqslant -0.02$	280	0.0357
4	$-0.02 < z \leqslant 0.21$	249	0.0361
5	$0.21 < z \leqslant 0.61$	351	0.0399
6	$0.61 < z \leqslant 0.65$	22	0.0455
7	$0.65 < z \leqslant 0.71$	50	0.0600
8	$0.71 < z \leqslant 0.82$	83	0.0602
9	$0.82 < z \leqslant 1.09$	136	0.0882
10	$1.09 < z \leqslant 1.83$	143	0.1468
11	$1.83 < z \leqslant 1.96$	6	0.1667
12	$1.96 < z \leqslant 2.61$	49	0.2041
13	$2.61 < z \leqslant 3.53$	21	0.2381
14	$3.53 < z \leqslant 4.42$	14	0.4286
$\hat{M}^* = 15$	$4.42 < z \leqslant 8.79$	8	0.7500
	$z > 8.79$	0	1

transaction mode, wealth, ethnic composition, family composition, neighborhood mix, and educational attainment, have no significant impact on customer propensity to buy. Thus, this finding resonates with Ehrenberg's (1969) insight: *Of many variables that could matter, only a few do*.

4.2. Probability estimates

Table 3 displays isotonic regression estimates of the response probability function $\hat{F}(\cdot)$, which reduces the row dimensionality from $N=2424$ customers to $\hat{M}^*=15$ groups. The largest group, $m=2$, contains 992 customers whose overall response probability is quite small, 2.7%. This finding suggests that most customer contacts within this group are likely to be unprofitable. We also observe that this approach identifies responsive customer groups, which is important for successful database marketing. The two most promising customer groups (namely, $m = 14$ and 15) consist of less than 1% of the sample, and exhibit a weighted average response probability of 54.5%.

To illustrate further the practical implications of this approach, we compare our results to those obtained by using the logistic distribution. Fig. 1 displays probability estimates from both the empirical and logistic distributions. It visually shows that the logistic distribution is different from the estimated $\hat{F}$. We formally assess whether $\hat{F}$ is close to the logistic distribution by applying the one-tailed Kolmogorov test statistic $D^- = \max[F_L - \hat{F}]$, where F_L denotes the logistic distribution. For the sake of comparison, we standardize the single-index scores by subtracting the mean and dividing by the standard deviation. The resulting value of $D^- = 0.7335$, which exceeds the critical value 0.295 at the 95% confidence level. Thus, the one-tailed test shows

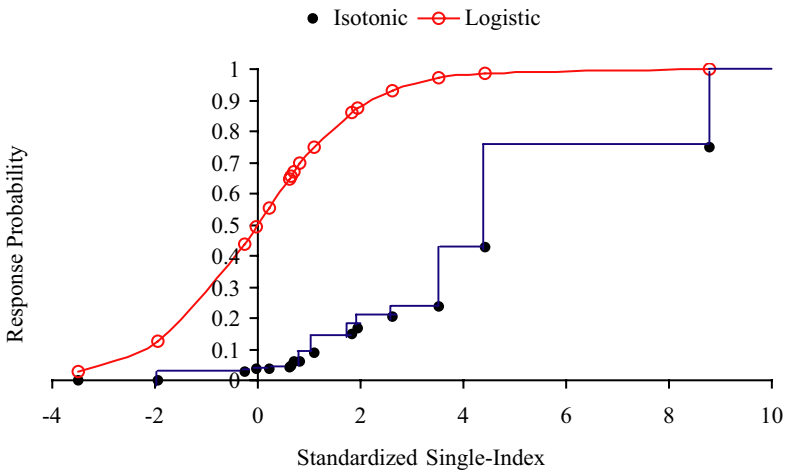

Fig. 1. Empirical (isotonic) and parametric (logistic) distribution functions.

F_L exceeds $\hat{F}$ significantly, which means that the *logistic distribution overestimates customer response probability substantially*.

Next, we illustrate how differences between the isotonic single-index and logistic models lead to different implications for customer selection. In isotonic single-index model, the probability that the most responsive customer group ($m = 15$) will respond to direct marketing contact is just 0.75, which is moderate compared to 0.99 based on the logistic distribution. Furthermore, the "tail" probability for $z \in (1.96, 8.79)$ drops dramatically from 0.75 to 0.2041, whereas it remains at about 0.90 level in the logistic model. Hence, the size and composition of customers selected for direct-marketing depends on the retained probability model. Because the shape of response probability function in the logistic model is predetermined *regardless of market data*, it potentially misleads marketers in mailing catalogs to non-responsive customers in some product-markets. In contrast, the proposed approach mitigates the problem of "junk mailing" by estimating the probability function using market data.

Finally, we note the relative computation times. Specifically, the estimation of the isotonic single-index model took 0.29 min to obtain the parameter estimates $\hat{\beta}$, the standard errors of $\hat{\beta}$, and the distribution function $\hat{F}$ using Gauss programming language on a 1.7 MHz PC. This timing compares favorably with 3.72 min to estimate the logistic regression (via iterative proportional fitting), thus indicating a speeding factor of about 12.

5. Discussion

Database marketers often use transaction data to identify prospective customers. In small data sets consisting of a few variables such as recency, frequency, and monetary value of past purchases, they could apply maximum-likelihood methods to calibrate

logistic regression models. In large databases, however, it seems impractical to implement maximum-likelihood methods in real-time (Balasubramanian et al., 1998). Hence, we propose the isotonic single-index model, develop a direct approach for its estimation, and investigate their applicability to high-dimensional database marketing.

The direct approach for estimating isotonic single-index models is both scalable and flexible. In the empirical example, we estimated the high-dimensional parameter vector for 166 variables in the transaction database. An advantage of this approach is that it possesses desirable properties even if (a) the error distribution in Eq. (1) is unknown, and (b) multivariate distribution of the explanatory variables is skewed/non-normal. Furthermore, in contrast to Cosslett (1983), it provides standard errors of the estimated parameters, which enable us to infer the significant variables influencing customer's purchase. Finally, it allows us to empirically *estimate*—rather than a priori specify—the response probability for a customer's purchase in a given product-market.

As for future research, other dimension-reduction approaches such as sliced average variance estimation (Cook and Weisberg, 1991) and difference of covariance estimation (Cook and Lee, 1999) need empirical investigation. Next, extend the information criterion of Naik and Tsai (2001) to encompass binary response variables. Thirdly, adopt Hall and Huang's (2001) "monotonizing general kernel-type estimation" to obtain the smoothed monotonic curve $\tilde{F}(\cdot)$ and then generate its bootstrapped estimates $\tilde{F}^*$ for assessing the impact of sampling variation on the estimated probability function. Such efforts would improve the practice of high-dimensional database marketing.

References

Altham, P.M.E., 1984. Improving the precision of estimation by fitting a model. J. Roy. Statist. Soc. Ser. B 46, 118–119.

Ayer, M., Brunk, H.D., Ewing, G.M., Reid, W.T., Silverman, E., 1955. An empirical distribution function for sampling with incomplete information. Ann. Math. Statist. 26, 641–647.

Balasubramanian, S., Gupta, S., Kamakura, W., Wedel, M., 1998. Modeling large data sets in marketing. Statist. Neerlandica 52, 303–323.

Barlow, R.E., Bartholomew, D.J., Bremner, J.M., Brunk, H.D., 1972. Statistical Inference under Order Restriction. Wiley, Chichester, UK.

Berry, M.J.A., Linoff, G., 1997. Data Mining Techniques for Marketing, Sales, and Customer Support. Wiley, New York, NY.

Brillinger, D.R., 1983. A generalized linear model with Gaussian regression variables. In: Bickel, P.J., Doksum, K.A., Hodges, J.L. (Eds.), A Festschrift for Erich L. Lehmann in Honor of His Sixty-Fifth Birthday. Wadsworth, California, pp. 97–114.

Bitran, G.R., Mondschein, S.V., 1996. Mailing decisions in the catalog sales industry. Manage. Sci. 42, 1364–1381.

Bult, J.R., Wansbeek, T., 1995. Optimal selection for direct mail. Marketing Sci. 14, 378–394.

Carroll, R.J., Fan, J., Gijbels, I., Wand, M.P., 1997. Generalized partially linear single-index models. J. Amer. Statist. Assoc. 92, 477–489.

Chen, C.H., Li, K.C., 1998. Can SIR be as popular as multiple linear regression? Statist. Sinica 8, 289–316.

Cook, R.D., Lee, H., 1999. Dimension reduction in binary response regression. J. Amer. Statist. Assoc. 94, 1187–1200.

Cook, R.D., Nachtsheim, C.J., 1994. Reweighting to achieve elliptically contoured covariates in regression. J. Amer. Statist. Assoc. 89, 592–599.

Cook, R.D., Weisberg, S., 1991. Discussion of sliced inverse regression. J. Amer. Statist. Assoc. 86, 328–332.

Cosslett, S.R., 1983. Distribution-free maximum likelihood estimator of the binary choice model. Econometrica 51, 765–782.

Cox, D.R., Snell, E.J., 1989. Analysis of Binary Data, 2nd Edition. Chapman & Hall, New York, NY.

Dodson, J.A., Tybout, A.M., Sternthal, B., 1978. Impact of deals and deal retraction on brand switching. J. Marketing Res. 15, 72–81.

Duan, N., Li, K.C., 1991. Slicing regression: a link-free regression method. Ann. Statist. 19, 505–530.

Ehrenberg, A.S.C., 1969. Laws in marketing. In: Bogart, L. (Ed.), Current Controversy in Marketing Research. Markham, Chicago, IL, pp. 141–152.

Gifi, A., 1990. Nonlinear Multivariate Analysis. Wiley, New York, NY.

Gonul, F., Shi, M.Z., 1998. Optimal mailing of catalogs: new methodology using estimable structural dynamic programming models. Manage. Sci. 44, 1249–1262.

Gonul, F., Kim, B.D., Shi, M.Z., 2000. Mailing smarter to catalog customers. J. Interactive Marketing 14, 2–16.

Hall, P., Huang, L.S., 2001. Nonparametric kernel regression subject to monotonicity constraints. Ann. Statist. 29, 624–647.

Horowitz, J.L., 1998. Semiparametric Methods in Econometrics. Springer, New York, NY.

Hsing, T., Carroll, R.J., 1992. An asymptotic theory for sliced inverse regression. Ann. Statist. 20, 1040–1061.

Hughes, A.M., 1996. The Complete Database Marketer. McGraw-Hill, New York, NY.

Ichimura, H., 1993. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. J. Econom. 58, 71–120.

Kiefer, J., Wofowitz, J., 1956. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Ann. Math. Statist. 27, 887–906.

Leeflang, P., Wittink, D.R., Wedel, M., Naert, P.A., 2000. Building Models for Marketing Decisions. Kluwer Academic Publishers, Boston, MA.

Li, K.C., 1991. Sliced inverse regression for dimension reduction. J. Amer. Statist. Assoc. 86, 316–342 (with discussions).

Li, K.C., 2000. High Dimensional Data Analysis Via the SIR/PHD Approach. Unpublished manuscript dated April 6, 2000 obtained at the Internet site www.stat.ucla.edu∼kcli/sir-PHD.pdf.

McLachlan, G., Peel, D., 2000. Finite Mixture Models. Wiley, New York, NY.

Naik, P.A., Raman, K., 2003. Understanding the impact of synergy in multimedia communications. J. Marketing Res. 40 (4), 375–388.

Naik, P.A., Tsai, C.L., 2000. Partial least squares estimator for single-index models. J. Roy. Statist. Soc. Ser. B 62, 763–771.

Naik, P.A., Tsai, C.L., 2001. Single-index model selections. Biometrika 88 (3), 821–832.

Powell, J.L., Stock, J.H., Stoker, T.M., 1989. Semiparametric estimation of index coefficients. Econometrica 57, 1403–1430.

Schmittlein, D.C., Morrison, D.G., Colombo, R., 1987. Counting your customers: who are they and what will they do next?. Manage. Sci. 33, 1–24.

Sharma, S., 1996. Applied Multivariate Techniques. Wiley, New York.

Silverman, B.W., 1986. Density Estimation. Chapman & Hall, London.

Simonoff, J.S., 1996. Smoothing Methods in Statistics. Springer, New York, NY.

Statistical Fact Book, 2000. The Direct Marketing Association, 22nd Edition. New York, NY.

Stoker, T.M., 1986. Consistent estimation of scaled coefficients. Econometrika 54, 1461–1481.

Wald, A., 1949. Note on the consistency of the maximum likelihood estimate. Ann. Math. Statist. 20, 595–601.

Wedel, M., Kamakura, W., 2000. Market Segmentation: Conceptual and Methodological Foundations, 2nd Edition. Kluwer Academic, Boston, MA.