

Multi-Factor Binary Response Analysis of Large Sparse Databases for CRM¹

By

Prasad A. Naik
Associate Professor of Marketing
University of California, Davis, CA.

Michel Wedel
Dwight F. Benton Professor of Marketing
University of Michigan, Ann Arbor, MI.

Final Report: August 25, 2004

Contact Information

Prof. Prasad A. Naik
One Shields Avenue
AOB IV, Room 151
University of California Davis
Davis, CA 95616.
Tel: (530) 754 9834
Fax: (530) 752 2924
Email: panaik@ucdavis.edu.

Copyright © 2004 Prasad A. Naik and Michel Wedel.

¹ The authors appreciate both the financial support of the Teradata Center and the valuable suggestions of Professors Rick Staelin and Wagner Kamakura on this research.

Multi-Factor Binary Response Analysis of Large Sparse Databases for CRM

SUMMARY

Customer Relationship Management (CRM) places individual customers at the heart of the company and its strategies, and seeks to establish and strengthen relationships with them by developing an understanding for the motives for their individual choices. Typically, parametric choice models such as the logit and probit models have been used for that purpose. But parametric binary models are not sufficiently flexible to accommodate sparse data, whereas the available nonparametric models are not scalable to large databases. To overcome these drawbacks, we propose the Multi-Factor Binary Response (MBR) model and develop a non-iterative two-step approach to estimate it. Specifically, in the projection step, we estimate the factor structure underlying the predictor space via sliced average variance estimation; in the calibration step, we estimate the unknown link function that relates the estimated factors to the expected response via multivariate kernel regression. We investigate finite sample performance using Monte Carlo studies and illustrate MBR in an empirical application. We glean four insights from the empirical results. First, we observe that a vast landscape in the uncovered dimensions is nearly flat, indicating those customers who will not buy. Second, the results reveal that marketing plans for acquiring new customers in “top deciles” may be too aggressive. Third, responsive customers are spread out in the edges of the high-dimensional space and outliers are the norm rather than exception. Finally, we find evidence that responsive customers belong to *disconnected* sets, i.e., pockets in the data. We discuss the implications of these findings and the potential of the MBR model for CRM.

1. INTRODUCTION

Customer Relationship Management (CRM) places individual customers at the heart of the company and its strategies, and seeks to establish and strengthen relationships with them by developing an understanding for the motives for their individual choices (Day 2000; Winer 2001). The customer transaction database, compiled by many companies, is crucial in achieving that goal. The CRM focus on individual customers invokes the need for explaining choice of individuals, which induces CRM-analysts to look at such databases at finer levels of granularity. But, as companies' offerings continue to expand, transactions for single offers (e.g., single books on Amazon.com, CD's on CDnow.com, fmcg-SKU's on Peapod.com, or DVD's on Netflix.com) are a relatively rare phenomenon and may only be sporadically encountered in the database. Usually transaction databases contain only transactions with the focal company, and the fear of having insufficient information to explain sparse choices prompts marketers to search for additional relevant data sources with richer explanatory variables. For example, data sources such as car and voter registrations provide information on age, name, address, and telephone; county records reveal personal information such as home value; census surveys yield geo-demographic data; purchase transactions at online and off-line stores describe buying habits; credit card companies reveal spending patterns; and airline companies know travel patterns. By combining such rich information with customer transaction data and analyzing it jointly in efforts to better explain individual choices, marketers have attempted to develop more effective CRM programs. However, this wealth of information also presents challenges. Driven by the need for comprehensive information on individual customers, the size of the databases, not only the number of customers, but especially the number of variables, has exploded.

In CRM and database marketing, parametric models such as logistic regression have become popular tools to predict whether a customer would buy or not buy a firm's offerings. But logistic regression—similar to probit regression and other parametric probability models—pre-specifies the shape of an underlying probability function of choice. Because that function is mostly symmetric, it cannot accurately predict rare events with low probability of occurrence (Cosslett 1983). One class of models that overcomes this drawback are nonparametric binary models. These models estimate a flexible probability function without specifying a particular shape (see Bult 1993). However, as the number of predictor variables increases, nonparametric models suffer from the “curse of dimensionality” due to the empty space phenomenon, which we will explain later (see, e.g., Simonoff 1996, p. 101). Consequently, these nonparametric approaches break down for large databases encountered in CRM applications; their applicability is restricted to small datasets with ten or fewer variables. In sum, parametric binary models are not sufficiently flexible to accommodate sparse CRM data, whereas nonparametric binary models are not scalable to large CRM databases.

Recently, in the tradition of approaches such as Sliced Inverse Regression (SIR, Li 1991) and Sliced Average Variance Estimation (SAVE, Cook and Weisberg 1991), Naik and Tsai (2004) proposed an isotonic single-factor model that is both flexible and scalable and developed a non-iterative approach to estimate it. This approach yields dimension reduction both column-wise and row-wise, which makes it scalable in these two dimensions of transaction databases; and, marketers can not only discover significant variables in large databases, but also prioritize customers into a few distinct groups based on estimated response probability (to enable direct mailing of catalogs).

In single-factor discrete choice models (e.g., Pagan and Ullah 1999, Ch. 7), a customer's response probability is given by $P(y = 1 | \mathbf{x}) = g(\mathbf{x}'\boldsymbol{\beta})$, where \mathbf{x} is a high-dimensional vector of predictor variables, and $g(\cdot)$ denotes the flexible distribution function. Consequently, the customer's purchase behavior depends on *only one* linear combination of explanatory variables or the factor $z = \mathbf{x}'\boldsymbol{\beta}$. This formulation with only one factor underlying customers' behavior seems to be too restrictive, especially without the ability to test for the absence of other factors.

We therefore intend to move beyond single-factor formulations in trying to answer the questions, "Is there more than a single factor, $z_k = \mathbf{x}'\boldsymbol{\beta}_k$, $k = 1, 2, \dots, K$, that influences customer's purchase behavior?" If so, how many factors are there? How do we extract these multiple factors? Can we estimate multiple factors without knowing the exact link function $g(\cdot)$ a priori? These questions are particularly relevant in CRM applications because (a) misspecification of the link function or the number of factors is likely in high-dimensional contexts with hundreds of variables; and (b) ignoring such issues would lead to biased prediction of choice and incorrect selection of prospective customers.

To address these questions and augment the class of existing probability models, we propose Multi-Factor Binary Response (MBR) model and develop a non-iterative consistent two-step method to estimate it. We call these two steps projection and calibration. Specifically, in the projection step, we combine information available in high-dimensional predictor space and project it to a low-dimensional factor space. To achieve this dimension reduction, we estimate the factor structure (i.e., the number of factors to retain and their composition in terms of the predictors, \mathbf{b}_k for $k = 1, \dots, K$)

without specifying the link function via SAVE (Cook and Weisberg 1991). In the calibration step, we estimate the unknown link function via kernel regression.

The intuition behind the two-step estimation of MBR models is as follows. The projection step retrieves the principal components of a certain covariance matrix derived from the sub-samples of customers, namely, those who choose the company's product and those who don't. This covariance matrix incorporates, non-parametrically, the information contained in customers' choice behaviors. A unique property of this projection step is its ability to extract multiple factors even if the response variable is binary; other approaches assume the existence of one factor without testing for the lack of multiple factors. Furthermore, the computational effort of this step reduces dramatically because the estimation of factor structure does not depend on the estimation of the link function $g(\cdot)$, thereby rendering this approach scalable to large CRM databases. The calibration step explicitly estimates the unknown link function to enable the prediction of the binary dependent variable.

The resulting estimated multi-factor model allows marketers to identify prospective customers in a more refined manner and forecast their responses from the identified factor structure. In contrast to the extant single-factor probability models (e.g., Bult 1993; Cosslett 1983) in which customers lie on a *line* (i.e., $z \in \mathfrak{R}^1$), suppose we find empirical evidence for the presence of two factors (i.e., $K = 2$); then customers will live in a two-dimensional *plane*, $z = (z_1, z_2) \in \mathfrak{R}^2$, some of them more responsive than others. Thus, by allowing for multiple factors, the MBR model facilitates a more refined understanding and targeting of customer groups, as we will illustrate using real data from a catalog company.

We organize this report as follows. Section 2 briefly reviews the extant probability models used in marketing and identifies their limitations in high dimensional data analyses. Section 3 presents the MBR model and its estimation approach. Section 4 investigates the finite sample performance via Monte Carlo studies, and section 5 illustrates an empirical application. Section 6 concludes by discussing the contributions and suggesting avenues to further extend this work.

2. LITERATURE REVIEW

Here we briefly review three kinds of probability models—parametric, nonparametric, and semi-parametric models—to understand their strengths and drawbacks in high dimensional data analysis. We summarize these approaches and mention one or two main references; for a comprehensive review, see Hastie, Tibshirani, and Friedman (2001).

2.1 Parametric Models

In CRM applications, the customer choice probability is estimated as a function of predictor variables in vector x (dimension $p \times 1$). Using a logistic regression model, this *response probability* is

$$P(y = 1 | x) = e^{x'b} / (1 + e^{x'b}), \quad (1)$$

where b_j denotes an effect of variable x_j on customer's likelihood of purchase ($j = 1, \dots, p$), and y indicates whether a customer bought ($y = 1$) or did not buy ($y = 0$) the product. We can estimate b in parametric models via maximum-likelihood estimation (MLE). The link function relates the variables x with the expected response $P(y = 1|x)$. In the logistic model (1), the link function is $g(z) = e^z / (1 + e^z)$; in the probit model $g(z) = \hat{O}(z)$, the cumulative distribution function of the Normal density.

Two main drawbacks of these and other parametric models are as follows. First, managers need to know the shape of the link function $g(\cdot)$, which characterizes the likelihood of customer's purchase. The logistic (or normal) link is symmetric *regardless of customer behavior or marketing situation facing the company*. Consequently, parametric models are inflexible for modelling rare events because the ratio of the right and the left tail-probability areas are invariant to data characteristics, and so parametric models yield very small estimates for $P(y = 1 | x)$ when applied to rare events. Hence, we need to relax parametric assumptions on $g(\cdot)$ to be able to predict such rare events.

Second, even if managers knew the exact shape of the link function for their particular market, the estimation of parametric models proceeds *iteratively*. The iterations are time-consuming and slow in the presence of many predictors (large p), which restricts the applicability of MLE in large CRM data sets. Alternatively, stepwise model selection and principal component regression are two possible ways to handle many predictors. But stepwise model selection increases the computational time enormously, because as many as $(2^p - 1)$ models are estimated, which is an astronomical number when $p \approx 100$ variables. On the other hand, principal components regression, while appealing in its ability to reduce dimensionality of the predictor-space, runs the risk of eliminating predictive factors that do not explain large variation among predictors. This risk is especially high when the predictor set is large and a few components need to be retained.

2.2 Nonparametric Models

Nonparametric probability models are expressed as

$$P(y = 1 | x) = g(x_1, x_2, \dots, x_p), \quad (2)$$

where the link function $g(\cdot)$ itself is an object of estimation. To estimate $g(\cdot)$, we can apply kernel regression (e.g., Fan et al. 1995, Simonoff 1996) to characterize the empirical shape of the link function.

The main disadvantage of non-parametric models is the curse of dimensionality, which induces an “empty space” phenomenon. To understand this phenomenon, consider the standard normal density and observe that 68.3% of its total mass lies within ± 1 standard deviation from the origin. For a bivariate normal, the mass within a unit square centered at the origin is about $(0.683)^2$, which is less than 50% of its total mass. In p -variate normal density, the mass within a unit hypercube is about $(0.683)^p$, which tends to zero rapidly as p increases. For example, when $p = 10$, only 2% of the finite sample is near the center of the density, i.e. in the unit hypercube. Consequently, as Silverman (1986, p. 92) cautions, “large regions of high [dimensional] density may be completely devoid of observations in a sample of moderate size.” In other words, local neighborhoods in high dimensions are almost surely empty, and those that are not empty are almost surely not local (Simonoff 1996, p. 101). Due to this empty space phenomenon, model (2) cannot be estimated accurately for p greater than ten or more variables and much fewer in practical applications with limited data. Therefore we need to formulate semi-parametric models to overcome the curse of dimensionality.

2.3 Semi-parametric Models

A semi-parametric probability model is given by

$$P(y = 1 | x) = g(x' \mathbf{b}), \quad (3)$$

where both the link function $g(\cdot)$ itself and the vector \mathbf{b} are the objects of estimation. In model (3), we overcome the curse of dimensionality because the index $z = x' \mathbf{b}$, together with the link $g(\cdot)$, serves as the first projective approximation to a general

nonparametric function $f(x_1, x_2, \dots, x_p)$. Semi-parametric models do not assume a particular shape for the link function, thereby introducing the necessary flexibility in determining customers' response probability based on market data.

The main advantage of not pre-specifying the shape of $g(\cdot)$ is as follows. If the pre-specified parametric function (as in equation 1) was identical to the true link that generated the data, then both the parametric and semi-parametric models would work satisfactorily. Should the pre-specified form differ from the true link, then the parametric model will definitely fail to estimate response probabilities accurately, while the semi-parametric model mitigates this risk of misspecification. The reason is that the parametric link functions are mostly symmetric and understate the choice probabilities for rare events (and conversely overstate the probabilities for common events).

One of the two main drawbacks of most semi-parametric models is that these models ignore the possibility of multiple factors. In other words, the response probability $P(y = 1|x)$ depends on the single factor $z_1 = x'\beta_1$ but not on any other factors $z_k = x'\beta_k$, $k = 2, \dots, K$. The multi-factor representation acquires its appeal from principal components regression, which provides a low-dimensional mapping of the high-dimensional predictor space. However, in that method the extraction of components and the prediction of a dependent variable are dissociated and a linear link function is assumed.

The second drawback of semi-parametric models is that the estimation approach is iterative and slow. To substantiate, consider the direct-marketing application in Bult and Wansbeek (1995), who apply Cosslett's (1983) estimator to calibrate model (3). This estimator is computationally intensive because the estimation of β depends on estimating $g(\cdot)$, and the estimated $\hat{g}(\cdot)$ is a discontinuous step function, which rules out

all gradient-based methods for efficiently estimating β . Other semi-parametric estimators of discrete choice models as reviewed in Pagan and Ullah (1999, Ch.7) also are computationally “very costly” (Pagan and Ullah 1999, p. 283). Consequently, marketers are unable to apply these approaches in real-time to large CRM data sets with hundreds of variables.

To summarize, parametric models are not flexible to characterize customers’ response probability in diverse product-markets; nonparametric models suffer from the curse of dimensionality due to the empty space phenomenon; and most semi-parametric models ignore multiple factors and their estimation approaches are iterative and slow. Next, we propose a probability model with multiple factors and a flexible link function; it is suitable for high-dimensional CRM applications with hundreds of variables because it can be estimated non-iteratively in two steps.

3. MULTI-FACTOR BINARY RESPONSE MODEL

We first formulate the model and then describe an approach to estimate it.

3.1 The MBR Model

Let the matrix X denote a large CRM data set with dimensions $N \times p$, where N is the number of respondents, and p is the number of variables. Let the dependent variable be a binary vector Y of dimension $N \times 1$. Then we propose the multi-factor binary response (MBR) model that relates Y and X as follows:

$$P(Y = 1|X) = g(X\beta_1, X\beta_2, \dots, X\beta_K), \quad (4)$$

where $Z_k = X\beta_k$ is a *factor*, i.e., a linear combination of the original variables in X , K denotes the total number of factors to be retained ($k = 1, \dots, K$), and the link function g :

$\mathfrak{R}^K \rightarrow (0, 1)$ is unknown. The subsequent three remarks distinguish the MBR model (4) from existing approaches.

Remark 1. Both parametric and semi-parametric models are special cases of the MBR model. Specifically, in parametric model (1), the shape of $g(\cdot)$ is restricted to known link functions with at most one factor (i.e., $K = 1$). For example, $g(z) = e^z/(1+e^z)$ in the logistic regression, and $g(z) = \hat{O}(z)$, the cumulative Normal density, in the probit model. In the semi-parametric model (3), the link function has a flexible shape, but the total number of factors is restricted to $K = 1$.

Remark 2. The MBR model is a nonparametric model similar to (2), but it lives in a low dimensional factor space $z = (z_1, z_2, \dots, z_K)$ rather than in the original variables space $x = (x_1, x_2, \dots, x_p)$, where $K \ll p$. If we achieve dimension reduction from $p \approx 100$ to $K < 10$ (say), then the nonparametric link $g(\cdot)$ becomes estimable using reasonable sample sizes, thus overcoming the curse of dimensionality.

Remark 3. Although Li's (1991) formulation appears similar to the model (4), sliced inverse regression (SIR) can estimate as many factors as $\min(p, H-1)$, where H is the number of slices. When response variable is binary, $H = 2$ and so SIR can find only one effective dimension. In other words, SIR cannot estimate multiple factors (i.e., $K > 1$) in the MBR model.

Next, we present an approach to extract multiple factors and estimate the nonparametric link in the MBR model.

3.2 Estimation Approach

The estimation problem is to discover the latent factor structure and its relation with the binary response variable. The latent factor structure consists of the number of factors K and the factor composition $B = \{\beta_1, \dots, \beta_K\}$. In addition, we need to

determine the link function $g(z_1, z_2, \dots, z_K)$, where $z_k = \mathbf{x}'\boldsymbol{\beta}_k$ are factor scores. The objects (g, \mathbf{B}) belong to the space $\Gamma : \Omega \times \mathfrak{R}^{p \times K}$, and we estimate them *without* iterating between the function space $\dot{U} : \mathfrak{R}^K \rightarrow (0,1)$ and the parameter space $\mathfrak{R}^{p \times K}$. To construct a non-iterative approach, we first obtain $\hat{\mathbf{B}}$ similar to SAVE (Cook and Weisberg 1991) and then find \hat{g} via kernel regression (e.g., Simonoff 1996).

3.2.1 Estimating the factor structure

Let $\tilde{\mathbf{x}} = \boldsymbol{\Sigma}_x^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ be the standardized \mathbf{x} , where $\boldsymbol{\mu} = E(\mathbf{x})$ and $\boldsymbol{\Sigma}_x^{1/2}$ denotes the Cholesky factor of $\boldsymbol{\Sigma}_x = \text{Cov}(\mathbf{x})$. We first transform the \mathbf{X} matrix to $\tilde{\mathbf{X}} = \hat{\boldsymbol{\Sigma}}_x^{-1/2}(\mathbf{X} - \bar{\mathbf{X}}')$, where $\hat{\boldsymbol{\Sigma}}_x = \frac{1}{N} \sum_{i=1}^N (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'$, \mathbf{X}_i' denotes the i -th row, $\bar{\mathbf{X}}$ contains the sample means, and N is the sample size and the resulting $\tilde{\mathbf{X}}$ has uncorrelated columns with zero means and unit variances.

We then estimate the factor composition \mathbf{B} by finding the linear combinations of $\tilde{\mathbf{x}}$, namely $(\tilde{\mathbf{x}}'\boldsymbol{\eta}_1, \dots, \tilde{\mathbf{x}}'\boldsymbol{\eta}_K)$, and we recover the parameter estimates in original \mathbf{x} -scale via $\boldsymbol{\beta} = \boldsymbol{\Sigma}_x^{-1/2}\boldsymbol{\eta}$. Specifically, SAVE estimates each direction $\boldsymbol{\eta}_k$ by the eigenvector $\boldsymbol{\gamma}_k$ obtained from the eigenvalue decomposition

$$\mathbf{M}\boldsymbol{\gamma}_k = \lambda_k \boldsymbol{\gamma}_k, \quad (5)$$

where λ_k is the k -th eigenvalue, $\mathbf{M} = E(\mathbf{I} - \text{Cov}(\tilde{\mathbf{x}} | \mathbf{y}))^2$, and $\text{Cov}(\tilde{\mathbf{x}} | \mathbf{y})$ is the conditional covariance of $\tilde{\mathbf{x}}$ given \mathbf{y} .

To obtain $\hat{\boldsymbol{\eta}}_k$, we next estimate \mathbf{M} by partitioning $\tilde{\mathbf{X}}$ into two slices $\tilde{\mathbf{X}}_0$ and $\tilde{\mathbf{X}}_1$, where $\tilde{\mathbf{X}}_0$ is a sub-matrix of $\tilde{\mathbf{X}}$ with all $Y_i = 0$, and $\tilde{\mathbf{X}}_1$ is the sub-matrix of $\tilde{\mathbf{X}}$ for $Y_i = 1$. In each slice $s = \{0, 1\}$, the conditional covariance matrix is

$$\hat{V}_s = \frac{1}{N_s} \sum_{i=1}^{N_s} (\tilde{X}_{si} - \bar{\tilde{X}}_s)(\tilde{X}_{si} - \bar{\tilde{X}}_s)', \quad (6)$$

where \tilde{X}'_{si} denotes the i -th row in slice s , $\bar{\tilde{X}}_s$ contains the sample means in slice s , and N_s is the sample size in slice s . The weighted average across the two slices yields the kernel matrix

$$\hat{M} = \sum_{s=0}^1 \hat{p}_s (\mathbf{I} - \hat{V}_s)^2, \quad (7)$$

where \hat{p}_s is the proportion of customers in slice s . Finally, we obtain $\hat{\eta}_k = \hat{\gamma}_k$ by substituting \hat{M} into (5) and solving the resulting eigenvalue problem. Since \hat{M} is of full rank, $K = p$ eigenvalues can be potentially retained.

Although statistical tests are available to infer the number of factors to be retained (Cook and Lee 1999, p. 1192; Cook and Yin 2001, p. 155), Cook (2000) recognizes that “these tests can be helpful but, as with many asymptotic tests, accuracy issues can arise in applications. Another option is to consider the relative magnitude of $\hat{\lambda}_j$ along with visual assessment, with relatively large eigenvalues indicating a probable significant contribution to the regression.” As in factor analysis, the “scree plot” facilitates the visual assessment (Catell 1966).

After we infer the number of factors and their composition, we compute the standardized factor scores $\tilde{Z} = (\tilde{z}_1, \dots, \tilde{z}_K) = (\tilde{X}\hat{\eta}_1, \dots, \tilde{X}\hat{\eta}_K)$. We emphasize that this determination of the latent factor structure did not require the knowledge of the link function $g(\cdot)$, whose estimation we discuss next.

3.2.2 Estimating the link function

If we achieve dimension reduction from p variables to K factors such that $K \ll p$, then we can estimate the unknown link function by applying nonparametric methods.

Fan et al. (1995) studied the theoretical aspects of binary nonparametric regression and, based on their results, we apply the Nadaraya-Watson estimator to sample observations (Y_i, \tilde{Z}_i) , $i = 1, \dots, N$, to estimate the response probability $P(Y = 1 | \tilde{Z})$ by

$$\begin{aligned} \hat{E}[Y = 1 | \tilde{Z}] &= \hat{g}(\tilde{Z}_1, \dots, \tilde{Z}_K) \\ &= \frac{\sum_{i=1}^N w_i Y_i}{\sum_{i=1}^N w_i}, \end{aligned} \quad (8)$$

where $w_i(t) = h^{-1} \exp(-(\tilde{Z}_i - t)'(\tilde{Z}_i - t)/h)$, is the multivariate normal kernel and h denotes the bandwidth.

For a prospective customer located at the point $t = (t_1, t_2, \dots, t_K)'$ in K -dimensional factor space, the function $w_i(t)$ assigns positive weights to every observation i . The closer an observed customer \tilde{z}_i to the prospective customer t , the greater the kernel-weight $w_i(t)$. It follows from (8) that the prospective customer's response probability $\hat{g}(t_1, t_2, \dots, t_K)$ is a weighted average of all the observed responses $\{Y_i\}$, and that this computation does not involve any iteration. In addition, the bandwidth h is a scalar because the factor scores are orthogonal with unit variances (i.e., $\text{Cov}(\tilde{Z}) = I_K$). Because the optimal bandwidth is $h^* = O(N^{-1/(K+4)})$ for the K -variate normal kernel (e.g., see Simonoff 1996, p. 105), we suggest using bandwidth $h = \sigma N^{-1/(K+4)}$, where σ measures the average dispersion of the estimated \tilde{z} . We close this section with three further remarks on the properties of this approach.

Remark 4. The idea underlying dimension reduction is to replace the $p \times 1$ predictor vector x by the $K \times 1$ predictor vector $B'x$ without loss of regression information. Inverse regression methods seek to estimate the central subspace $\mathcal{S}(B)$,

which identifies the space spanned by the columns of B . When y is binary and $K = 1$, SIR yields a consistent estimator of $\beta \subseteq \mathcal{S}(B) \subseteq \mathfrak{R}^p$ even if $g(\cdot)$ in (4) is not known (Li 1991, Hsing and Carroll 1992). For binary y , $K > 1$, and unknown $g(\cdot)$, SAVE provides the consistent estimate of $\mathcal{S}(B)$ via $\mathcal{S}(\hat{\eta}_1, \dots, \hat{\eta}_K)$, where $\hat{\eta}_k$, $k = 1, \dots, K$, are the first K eigenvectors of \hat{M} (Cook and Lee 1999). Finally, Fan, Heckman and Wand (1995) establish the conditions for consistent estimation of the multivariate link function $g(\cdot)$ when y is a binary response variable.

Remark 5. Table I summarizes the non-iterative algorithm for estimating MBR models. We acknowledge that further refinements in \hat{g} are possible at the cost of introducing iterative approaches. For example, one can replace the weights w_i by w_i/v_i , where $v_i = g_i(1-g_i)$, or use different bandwidth selection methods (e.g., Naik and Tsai 2001). Depending on the extent of dimension-reduction achieved in a particular CRM application, the users can appropriately trade-off the extra computational costs with the potential gains in estimation accuracy. While such extensions follow from standard theory (Fan et al. 1995), we advocate the simple approach because of the importance we attach to the scalability of the resulting procedure for it to be used in large CRM applications.

[INSERT TABLE I ABOUT HERE]

Remark 6. In semi-parametric regression models, including the MBR model, the regression coefficients are not of primary interest (unlike parametric models). One reason for this is that the number of estimated parameters is too large to be interpretable in most empirical applications. Instead of regression coefficients, the interest focuses on the estimated factor scores $\tilde{Z} = (\tilde{z}_1, \dots, \tilde{z}_K) = (\tilde{X}\hat{\eta}_1, \dots, \tilde{X}\hat{\eta}_K)$ and the estimated link

function $\hat{g}(\tilde{z}_1, \dots, \tilde{z}_K)$. Regression graphics (see Cook 1998) are then employed to facilitate interpretation by displaying shape of the link as a function of standardized factor scores. This graphical output of MBR models reveals the expected choice probabilities $\hat{g}(t_1, t_2, \dots, t_K)$ of prospective customers located at $t = (t_1, t_2, \dots, t_K)'$ in the factor-space.

4. MONTE CARLO RESULTS

The goal of this simulation study is to investigate performance of the projection step in extracting multiple factors in MBR models. To this end, we generated data using the model

$$y = \begin{cases} 1 & \text{if } \mu > \bar{\mu} \\ 0 & \text{otherwise,} \end{cases} \quad (9)$$

$$\mu = \exp(x'\beta_1) + 1/x'\beta_2, \quad (10)$$

where (9) creates the binary response variable, and (10) provides the link function. We consider four predictor variables in x , and each variable x_j is drawn from a normal density with mean zero and variance = 25. That is, $x_j \sim N(0, \sigma_j^2)$, $j = 1, 2, 3$ and 4, and $\sigma_j = 5$. Using this information, we construct the first factor $z_1 = x_1 + x_2$ and the second factor $z_2 = x_3 + 2x_4$. For each observation i , $i = 1, \dots, N = 1000$, we set $y_i = 1$, if $\mu_i = \exp(x_i'\beta_1) + 1/x_i'\beta_2$ exceeds the average $\bar{\mu} = \sum_{i=1}^N \mu_i / N$; else $y_i = 0$. The resulting 1000×5 matrix contains the variables $(y_i, x_{1i}, x_{2i}, x_{3i}, x_{4i})$. Applying the steps 1 through 4 of the estimation algorithm (see Table I), we estimated the factor composition $\hat{B} = (\hat{\beta}_1, \hat{\beta}_2)$ and the factor scores $\tilde{Z} = (\tilde{X}\hat{h}_1, \dots, \tilde{X}\hat{h}_K)$. We replicated this data generation and model estimation 100 times. Below we present the averaged results.

Figure 1 displays the scree plot of eigenvalues as a function of number of factors. These eigenvalues contain valuable information on how many factors to retain in a model. Specifically, a large value reveals an important factor, and a small value suggests that the factor lacks predictive capability. Figure 1 clearly reveals an “elbow” indicating a sharp change in slope at two factors (Catell 1966), where the first two eigenvalues are large compared to the last two. Hence, the projection step recovers the number of factors correctly even though it was neither supplied with or estimated the nonlinear link function (10) that generated this data. Given that the existing semi-parametric discrete choice models cannot estimate more than one factor (Pagan and Ullah 1999, Ch. 7), we consider this finding encouraging.

[INSERT FIGURE 1 ABOUT HERE]

We estimate the factor composition $B = (\beta_1, \beta_2)$ using the eigenvectors in (5). Because two factors were retained based on the scree-plot, we extract the first two eigenvectors, average them over 100 replications, and report the averaged values in Table II. For the sake of comparison, we standardize the first and the third element of $\hat{\beta}_1$ and $\hat{\beta}_2$ to unity, respectively. We observe that the first two elements of $\hat{\beta}_1$ are large and comparable in magnitude, while its last two elements are both small, reflecting the assumed pattern for β_1 . Similarly, the first two elements of $\hat{\beta}_2$ are small, and the last two elements are large, corresponding well to the true pattern of β_2 . To assess consistency of the projection step, we increase the sample size from $N = 1000$ and $N = 5000$. Table II shows that the estimated parameters tend to their true values as the sample size increases. In particular, note that the last two elements of $\hat{\beta}_2$ approach the true 1:2 ratio.

[INSERT TABLE II ABOUT HERE]

In Figure 2, we display the true and estimated factor scores. Panels A and B, respectively, show that the estimated scores for the factors 1 and 2 (plotted on y-axes) are proportional to the original factors (plotted on x-axes). In sum, we conclude that the projection step extracts the multiple factors from binary response data even if the link function is not known. We next illustrate the estimation of the unknown link function using a large CRM database.

[INSERT FIGURE 2 ABOUT HERE]

5. EMPIRICAL RESULTS

We consider the case of a database marketer who mailed catalogs to customers and maintained detailed records of purchase transactions for twelve years. These data are available from Direct Marketing Educational Foundation (data 03DMEF) for academic research (see www.the-dma.org). We merged the credit data (98DMEF) and geo-demographic data (99DMEF) with purchase transaction data using postal ZIP codes, and the resulting data matrix contains 2424 customers with no missing values and 166 predictor variables. While the sample size is small, the number of variables is representative of those encountered in typical CRM applications for us to illustrate the method on publicly available data (which offers other advantages).

Applying the algorithm in Table I to this customer database, we estimated the multi-factor binary response model. Figure 3 presents the scree plot of estimated eigenvalues as a function of the number of factors. Although 166 factors were extracted, we display the first fifty eigenvalues for clarity. An “elbow” in the scree plot suggests the number of factors to retain in the model. More specifically, the first six eigenvalues exceed unity; they exhibit a sharp decrease compared to the rest; and they

account for 62.9% of the total variation. So we use the first six factors in creating multiple factors in the MBR model. In other words, we reduced the dimensionality from $p = 166$ original variables to $K = 6$ factors without specifying any particular link function.

[INSERT FIGURE 3 ABOUT HERE]

To estimate the link function, we apply multivariate kernel regression in the six-dimensional factor space. It is important to recognize that such estimation would be impossible with 166 original variables and $N = 2424$ customers. Using (8), we estimate the link function $P(Y = 1 | \tilde{Z}) = \hat{g}(\tilde{z}_1, \tilde{z}_2, \tilde{z}_3, \tilde{z}_4, \tilde{z}_5, \tilde{z}_6)$, which has six input factor scores and one scalar output for the response probability. The estimated link function fits the data well. Specifically, the adjusted R-squared is 53.5% for the MBR model and 29% for the logistic regression; and the root mean squared error is 0.1534 for the MBR model compared to 0.1833 for the logistic regression.

More importantly, we learn about the response probability of *prospective* customers outside of the observed sample. Consider a prospective customer—not necessarily a customer listed in the database—located at a point $t = (t_1, t_2, t_3, t_4, t_5, t_6)'$ in the six-dimensional factor space. Substituting these values in (8), we evaluate the chance that this prospective customer would buy the company's financial product: $\hat{g}(t_1, t_2, t_3, t_4, t_5, t_6)$. This resulting probability estimate does not assume any specific functional form for $g(\cdot)$, and it utilizes information contained in the database on customers' behavior. By systematically evaluating this probability estimate for different values of $t \in \mathfrak{R}^6$, we can identify where high-potential prospects live.

To explore regions of high-potential prospects, we construct the probability surface $\hat{g}(\cdot)$ in six-dimensional space. Specifically, we present 3D plots with $\hat{g}(t_k, t_m; \bar{z}_{-k, -m})$ on the vertical axis and the prospective customers at points (t_k, t_m) in the horizontal plane of the two factors (k, m) , holding the other factors fixed at their sample means (i.e., $\bar{z}_{-k, -m}$). Figures 4 through 8 depict the probability surfaces and the corresponding contour projections for the factor combinations $k = 1$ and $m = 2, 3, \dots, 6$. For the sake of brevity, we do not present the other ten plots for the various combinations of pairs (k, m) . To aid customer identification, we use color graphics to visualize the quartiles of response probabilities: non-responsive customers (≤ 0.25), low probability (0.25 to 0.50), medium probability (0.50 to 0.75), and high probability (> 0.75). Such a color visualization of the results may be useful in practical applications as well.

[INSERT FIGURES 4 THROUGH 8 ABOUT HERE]

We glean four insights from these empirical results. First, we observe that a vast landscape in six dimensions is nearly flat. This finding suggests that many prospective customers have response probability close to zero; most prospective customers are non-responsive and transactions are relatively rare. The good news is that we know *whom* to not mail the catalogs, thus reducing printing and mailing costs and minimizing “junk mails.” The framework of Bult and Wansbeek (1995), based on profit maximization, can then be applied to optimally select customers using the MBR model (instead of their single-factor model).

Second, marketing plans for acquiring new customers in “top deciles” seem to be overstated. Realistic estimates for prospective customers’ chance of buying lie within the range $\frac{1}{4}$ to $\frac{3}{4}$. This finding also corroborates with a recent study by Naik

and Tsai (2004). Therefore, we recommend that an unknown link function should be estimated using a nonparametric method (rather than pre-specified) because it not only mitigates misspecification errors, but also promotes a more conservative approach for database marketing.

Third, because probability mass is not concentrated near the center of the density due to the empty space phenomenon, customers are spread out in the edges of the high-dimensional space. Consequently, the support (or domain) of the estimated link function exceeds the typical range of ± 3 standard deviations (see Figures 4-8). In other words, outliers are the norm rather than exception.

Finally, we find evidence that responsive customers belong to *disconnected* sets. Specifically, Figure 4 (see the lower panel) reveals three sets of customers: one set of non-responsive customers and two sets of potential customers. Note that potential customers are not all contiguous to one another. The probability models in the extant literature, regardless of whether the link function is parametric or nonparametric, *cannot* identify such disconnected customer sets because the response probability is a monotonic function of one single factor. To be able to identify pockets of responsive customers, a topic of great interest for CRM, marketers need multi-factor models so that they can conduct a more refined targeting and profiling of customer groups.

6. CONCLUSIONS

We believe the Multi-Index Binary Response (MBR) model holds promise for application to analytical CRM. One of the most important features of the approach is its scalability to large datasets. It consists of projection and calibration steps, both of

which involve simple non-iterative computations. This makes the approach feasible for the analysis of customers' discrete choices in transaction databases that are large both in the number of customers and the number of variables. The second important feature of the approach is its flexibility. It is more flexible than extant parametric choice models, since it allows for a non-parametric link function, which we think is useful especially in the analysis of sparse data, i.e., data with a small number of transactions observed among a large number of customers. The MBR approach in addition offers greater flexibility of calibrating the response function in a *multidimensional* space. Whereas non-parametric and parametric discrete choice models to date all assume a single factor to underlie customer behavior (e.g., see Pagan and Ullah 1999, Ch. 7), the MBR model allows and tests for the presence of multiple factors. Those multiple factors when present create new possibilities for targeting prospective customers.

Our analysis of empirical data revealed important insights: the response surface is flat over much of the factor space, with much of the transaction-activity occurring at the boundaries, in disconnected pockets of the data. These observations have important implications for CRM. CRM programs may need to be, and based on MBR analyses will be, more conservative in the number of customers to target for specific transactions, but may need to be more aggressive in the intensity with which these customers are targeted. While much progress has been made with the targeted CRM programs that recognize the potential of individual customers, still much marketing effort is wasted on potentially non-responsive customers. The past decades have seen a proliferation of the size and use of customer databases, leading to application of direct marketing techniques based on inappropriate and simplified (parametric) models for reasons of scalability. This has resulted in markets being canvassed with ineffective

marketing efforts, increasing consumer resistance to “junk mail,” and reducing the profitability of marketing activity. We think that techniques such as MBR may assist in developing more effective targeting of customers, greater responsiveness to direct marketing efforts, reduction of customer resistance, and greater profitability for the firm.

In closing, we suggest two avenues for extending this work. First, MBR models predict a single transaction, and their extension to multiple, possibly correlated, binary variables seems of interest. This effort would enable, for instance, the development of cross selling efforts (Kamakura et al. 2004). Second, further work can focus on either incorporating prior information on factor composition (Naik and Tsai 2005) or estimating nonparametric link functions under constraints (e.g., Hall and Huang 2001, Naik and Tsai 2004) when such constraints can be motivated from marketing considerations. We hope that this study presents a useful starting point for such further developments that may have an impact on the development of CRM programs and their effectiveness.

REFERENCES

- Bult, Jan-Roelf (1993), "Semiparametric versus Parametric Classification Models: An Application to Direct Marketing," *Journal of Marketing Research*, 30 (August), 380-390.
- Bult, Jan-Roelf and Tom Wansbeek (1995), "Optimal Selection for Direct Mail," *Marketing Science*, 14 (4), 378-394.
- Cattell, R. B. (1966), "The Scree Test for the Number of Factors," *Multivariate Behavioral Research*, 1, 245-276.
- Cook, R. Dennis (1998), *Regression Graphics: Ideas for Studying Regressions through Graphics*, New York: John Wiley & Sons.
- Cook, R. Dennis (2000), "SAVE: A Method for Dimension Reduction and Graphics in Regression," *Communications in Statistical Theory and Methods*, 29, 2109-2121.
- Cook, R. Dennis and Hakbae Lee (1999), "Dimension Reduction in Binary Response Models," *Journal of the American Statistical Association*, 94 (448), 1187-1200.
- Cook, R. Dennis and S. Weisberg (1991), "Discussion of Li (1991)," *Journal of the American Statistical Association*, 86, 328-332.
- Cook, R. Dennis and Xiangrong Yin (2001), "Dimension Reduction and Visualization in Discriminant Analysis (with Discussions)," *Aust. N. Z. J. Stat.*, 43 (2), 147-199.
- Cosslett, S. R. (1983), "Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model," *Econometrica*, 51, 765-782.
- Day, G. S. (2000), "Capabilities for Forging Customer Relations," MSI Working Paper 00-118, Cambridge: Massachusetts: MSI.
- Fan, J. Q., Heckman, N. E. and Wand, M. P. (1995), "Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions," *Journal of the American Statistical Association*, 90, 141-150.
- Hall, Peter and Huang, Li-Shan (2001), "Nonparametric Kernel Regression Subject to Monotonicity Constraints," *The Annals of Statistics*, 29 (3), 624-647.
- Hastie, Treor, Tibhirani, Robert, and Friedman, Jerome (2001), *The Elements of Statistical Learning*, New York: Springer.
- Hsing, Tailen and Raymond J. Carroll (1992), "An Asymptotic Theory for Sliced Inverse Regression," *The Annals of Statistics*, 20 (2), 1040-1061.
- Kamakura, Wagner A. de Rosa, Fernando, Wedel, Michel, Mazzon, Jose A. (2003), "Cross-selling Financial Services with Database Marketing," *International Journal of Research in Marketing*, 2003, 20(1), 45-65.
- Li, Ker-Chau (1991), "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*, 86, 316-342.
- Naik, Prasad A., and Tsai, Chih-Ling (2001), "Single-index Model Selections," *Biometrika*, 88 (3), 821-832.
- Naik, Prasad A., and Tsai, Chih-Ling (2004), "Isotonic Single-Index Model for Database Marketing," *Computational Statistics and Data Analysis*, forthcoming.

- Naik, Prasad A., and Tsai, Chih-Ling (2005), "Constrained Inverse Regression for Incorporating Prior Information," *Journal of the American Statistical Association*, forthcoming.
- Pagan, Adrian and Ullah, Aman (1999), *Nonparametric Econometrics*, New York: Cambridge University Press.
- Silverman, B. W. (1986), *Density Estimation*, London: Chapman & Hall.
- Simonoff, Jeff S. (1996), *Smoothing Methods in Statistics*, New York: Springer.
- Winer, Russell S. (2001), "A Framework for Customer Relationship Management," *California Management Review*, 43(4), 89-105.

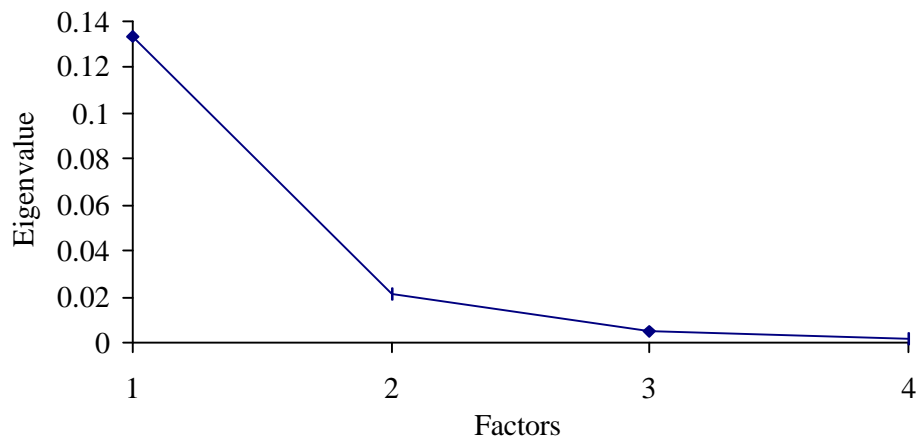
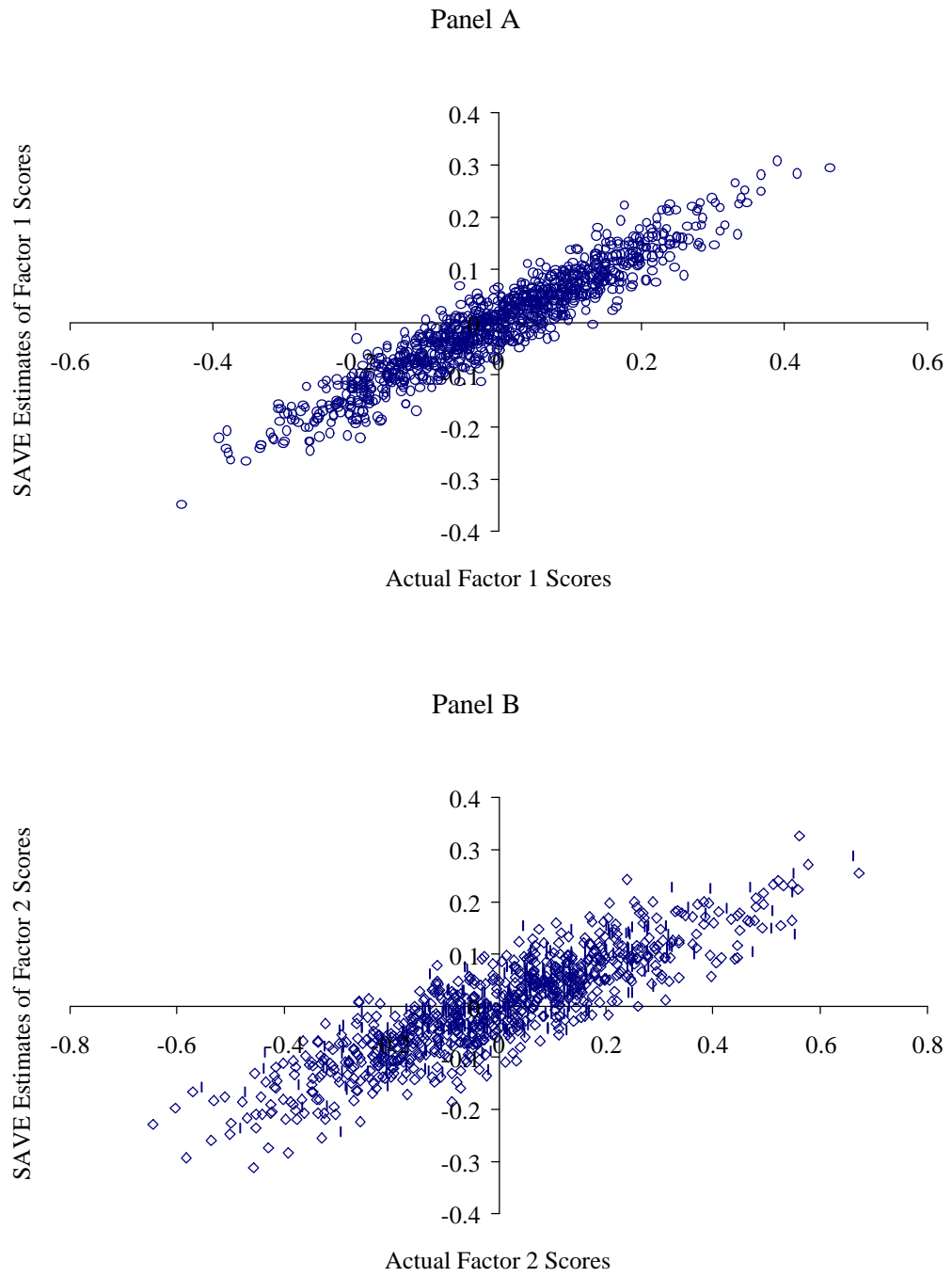


Figure 1. Scree plot for the simulation example

Figure 2. Estimated and actual factor scores in the simulation example



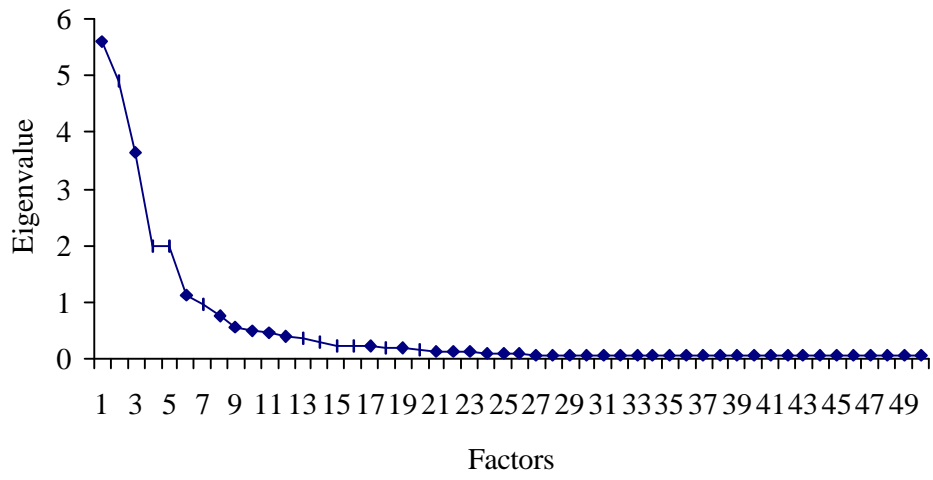


Figure 3. Scree plot for the empirical example (first 50 factors)

Figure 4. Response probability $\hat{g}(t_1, t_2; \bar{z}_{-1, -2})$ and its contour plot (lower panel)

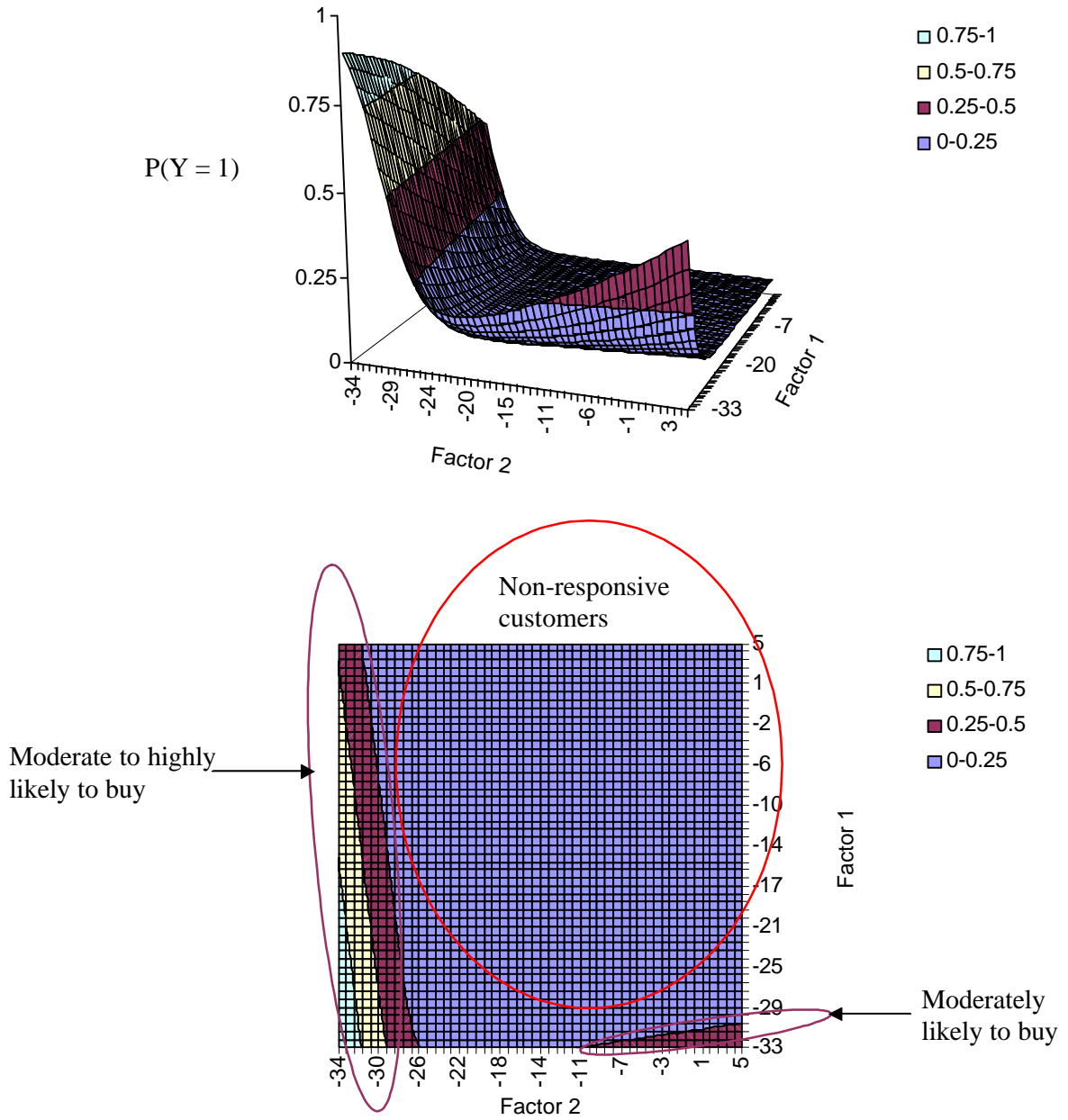


Figure 5. Response probability $\hat{g}(t_1, t_3; \bar{z}_{-1, -3})$ and its contour plot (lower panel)

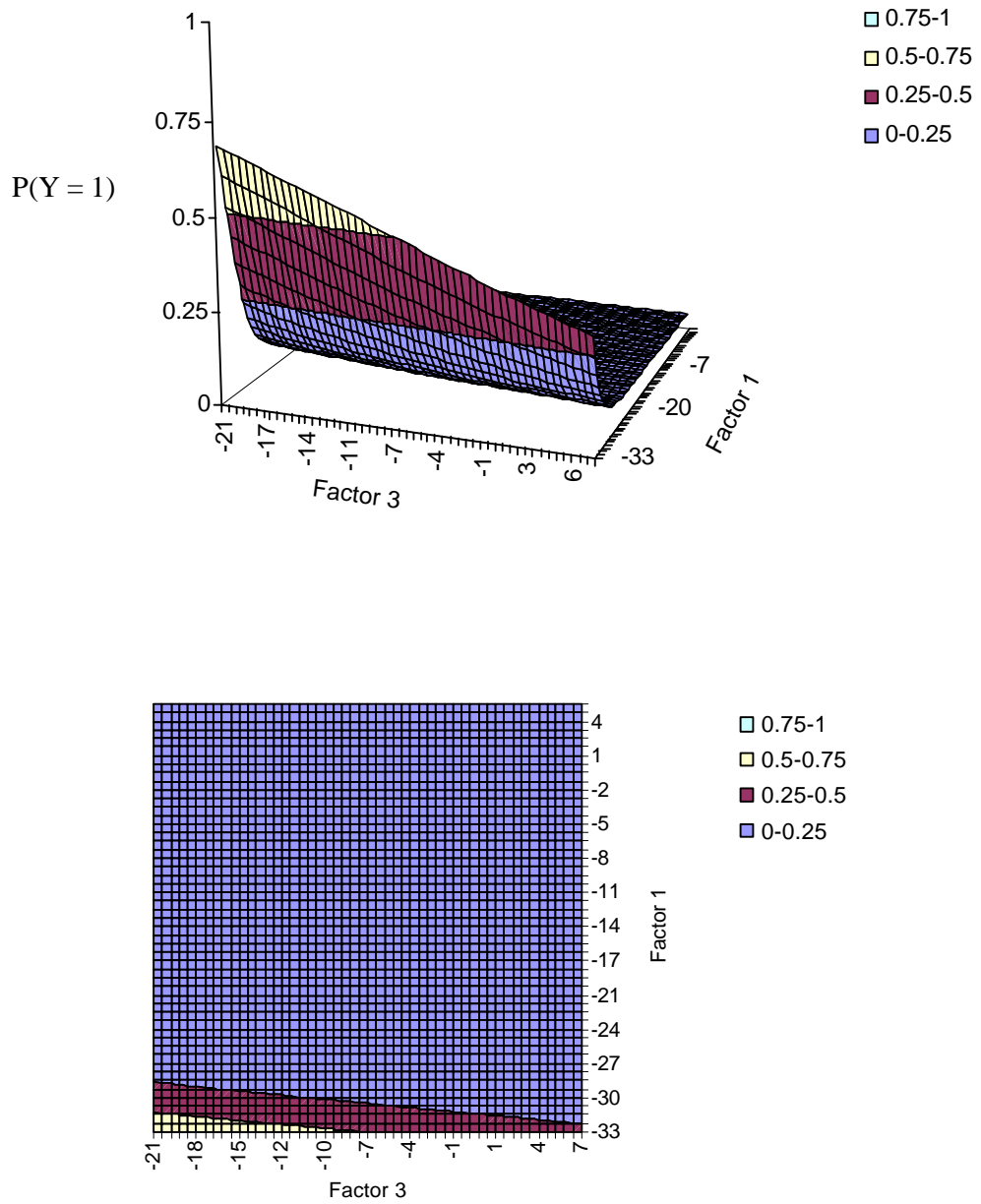


Figure 6. Response probability $\hat{g}(t_1, t_4; \bar{z}_{-1, -4})$ and its contour plot (lower panel)

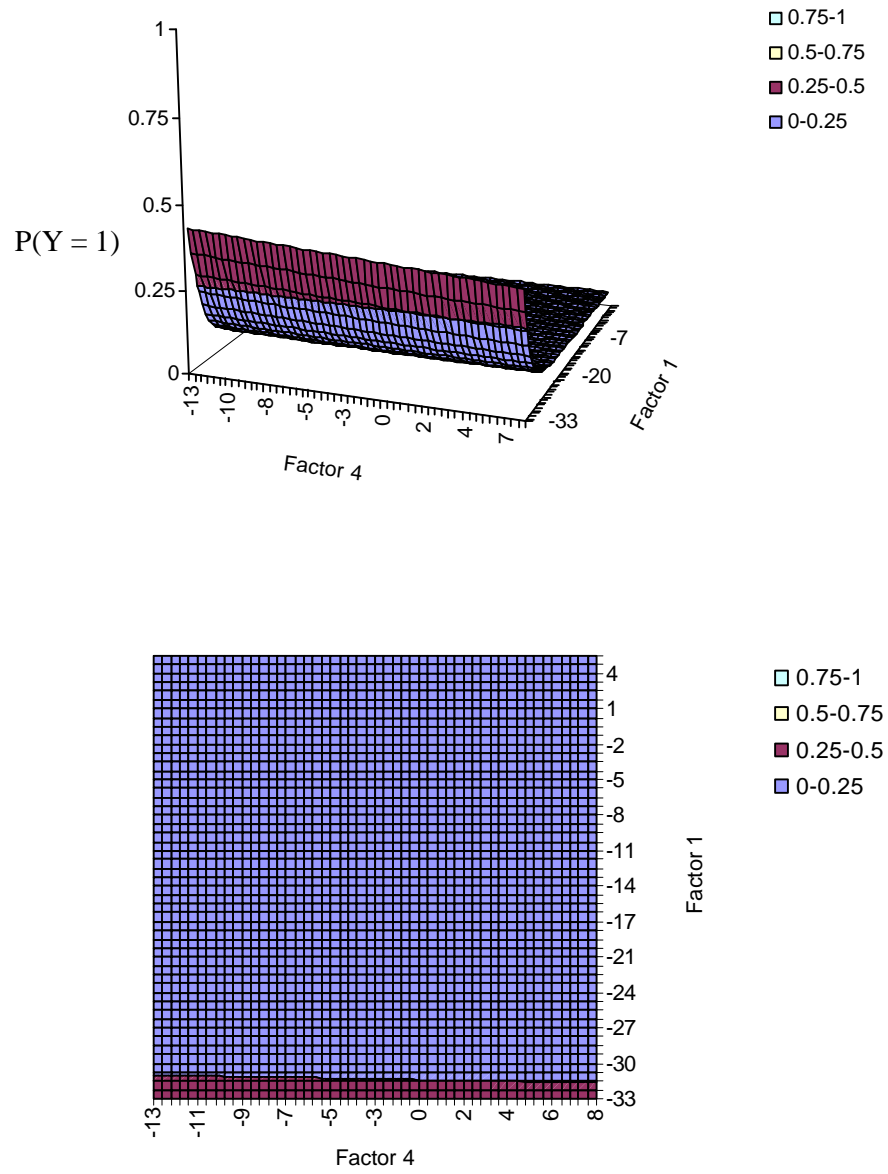


Figure 7. Response probability $\hat{g}(t_1, t_5; \bar{z}_{-1, -5})$ and its contour plot (lower panel)

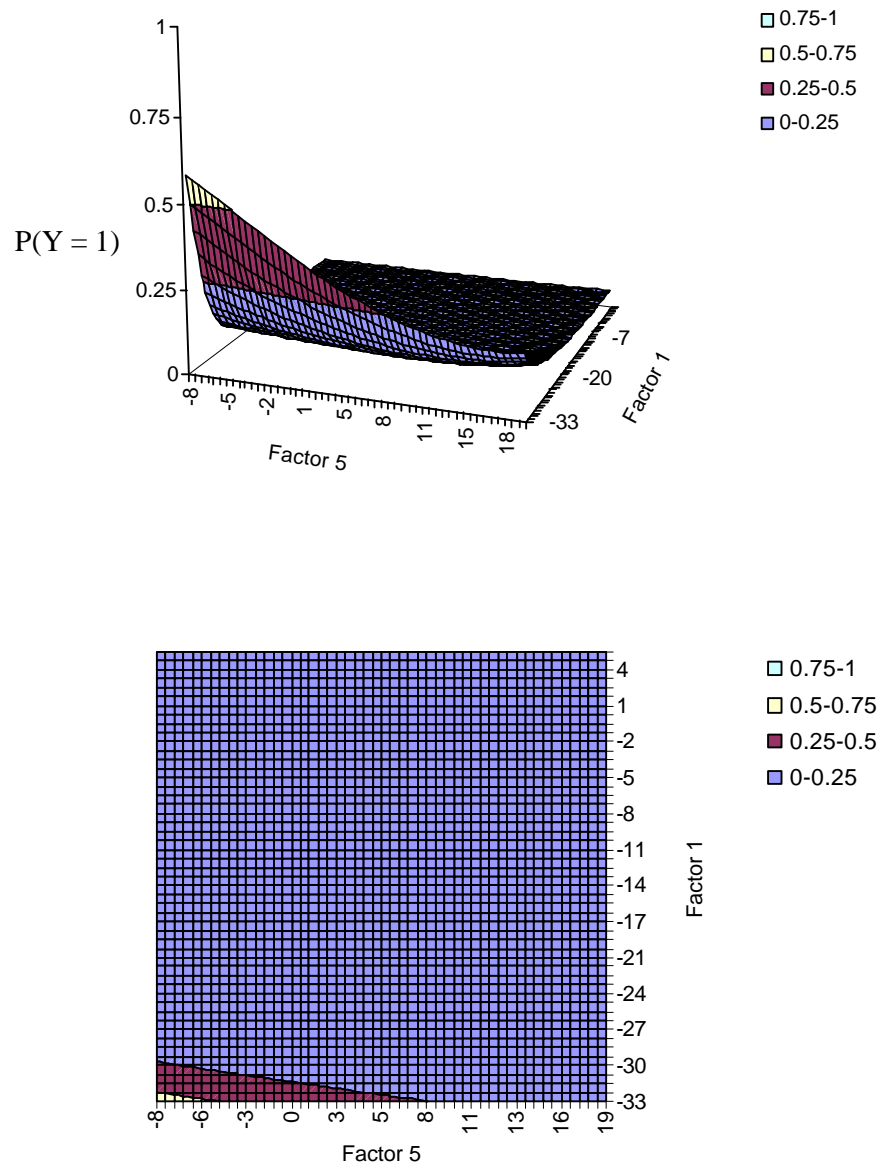


Figure 8. Response probability $\hat{g}(t_1, t_6; \bar{z}_{-1, -6})$ and its contour plot (lower panel)

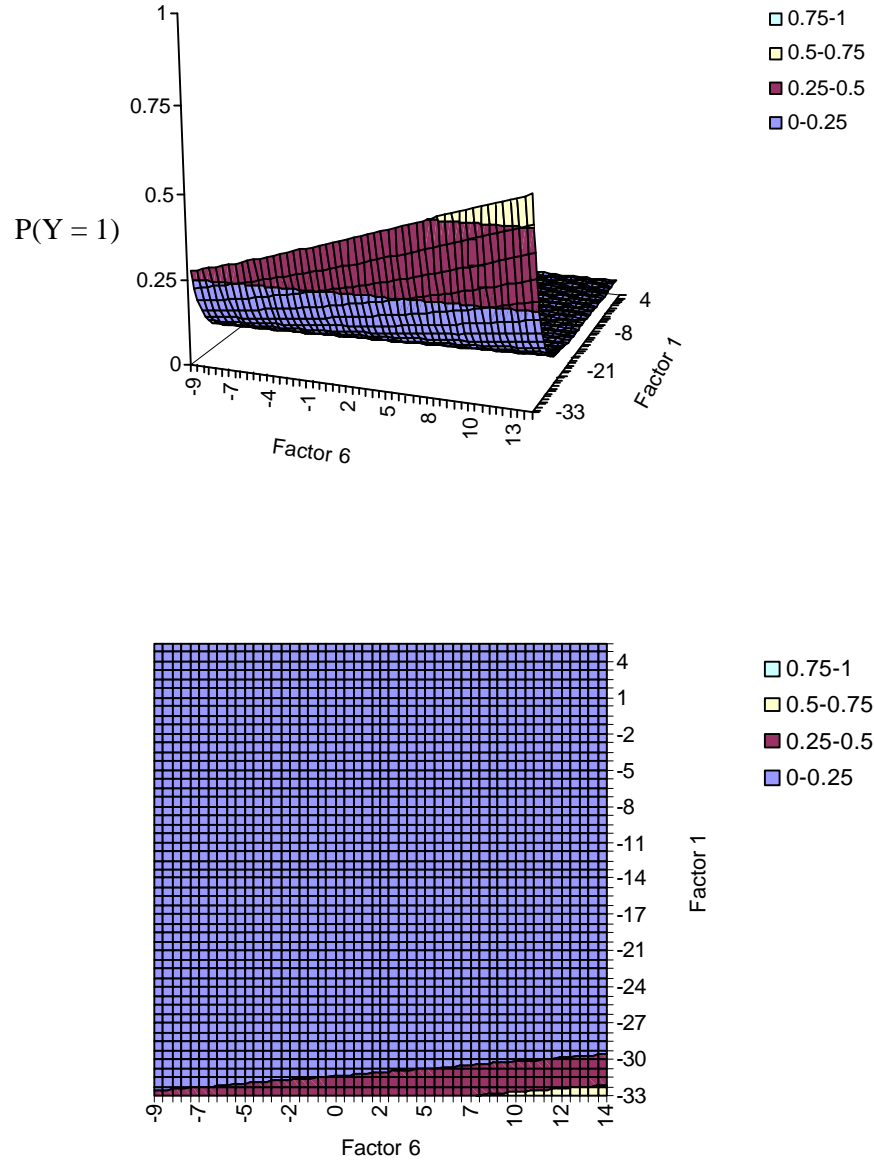


Table I. Estimation algorithm for MBR models

-
1. Transform the X matrix into $\tilde{X} = \hat{\Sigma}_x^{-1/2}(X - \bar{X}')$, where \bar{X} and $\hat{\Sigma}_x$ represent the sample analogs of the mean vector and covariance matrix, respectively.
 2. Partition \tilde{X} into \tilde{X}_0 and \tilde{X}_1 . The sub-matrix \tilde{X}_0 contains customers with $y = 0$, and \tilde{X}_1 contains those with $y = 1$.
 3. Compute \hat{M} by using (7) and solve the eigenvalue problem in (5). Scree plot of eigenvalues suggests the number of factors to be retained, K . The eigenvectors provide the basis vectors $\hat{\eta}_k$ ($k = 1, \dots, K$) for the span of B .
 4. Compute the standardized factor scores $\tilde{z}_k = \tilde{X}\hat{\eta}_k$ for each factor k .
 5. Evaluate the kernel weights $w_i(t) = h^{-1} \exp(-(\tilde{z}_i - t)'(\tilde{z}_i - t)/h)$, where $\tilde{z}_i = (\tilde{z}_{i1}, \dots, \tilde{z}_{iK})'$ represents the location of the customer i , and $t = (t_1, t_2, \dots, t_K)'$ denotes a prospective customer.
 6. A prospective customer's response probability $\hat{g}(t_1, t_2, \dots, t_K)$ is computed using equation (8).
-

Table II. Consistency of the projection step as the sample size increases

Variables	True		N = 1000		N = 5000	
	β_1	β_2	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
x_1	1	0	1.0000	0.5552	1.000	0.3258
x_2	1	0	0.9971	0.5163	0.9973	0.3139
x_3	0	1	0.1770	1.0000	0.1473	1.000
x_4	0	2	0.3033	1.6959	0.2641	1.9293