Multi-Index Binary Response Analysis of Large Data Sets

Prasad A. NAIK
Graduate School of Management, University of California Davis, Davis, CA 95616 (*panaik@ucdavis.edu*)

Michel WEDEL
Department of Marketing, Robert H. Smith School of Business, University of Maryland

Wagner KAMAKURA
Fuqua School of Business, Duke University

We propose a multi-index binary response model for analyzing large databases (i.e., with many regressors). We combine many regressors into factors (or indexes) and then estimate the link function via parametric or nonparametric methods. Neither the estimation of factors nor the determination of the number of factors requires ex ante knowledge of the link between the response and regressors. Furthermore, applying perturbation theory, we furnish a new asymptotic result to facilitate significance tests of factor loadings. We illustrate this approach with an empirical application in which we reduced dimensionality from 124 regressors to 4 factors.

KEY WORDS: Customer relationship marketing; Discrete choice; Factor model; Inverse regression; Semiparametric estimation; Sliced average variance estimation.

1. INTRODUCTION

In business and economics, parametric models such as logistic regression have become popular tools for predicting whether a customer would buy a firm's products or defect to a competitor's services. But parametric models have the drawback of prespecifying the shape of an underlying probability function. When the prespecified function departs from the true unknown link, it cannot accurately predict rare events with low probability of occurrence (Cosslett 1983). Nonparametric binary models overcome this drawback. These models estimate a flexible probability function that does not depend on a particular functional form. But as the number of regressors increases, nonparametric models suffer from the curse of dimensionality, which induces the empty-space phenomenon (see, e.g., Simonoff 1996, p. 101). Consequently, nonparametric approaches break down when data sets contain a large number of regressors (i.e., predictor variables), which often occurs in many business applications, and so their applicability is restricted to small data sets with 10 or fewer variables. Thus parametric models are not sufficiently flexible, whereas nonparametric models are not scalable to large databases encountered by businesses.

To overcome these problems, semiparametric binary choice models are proposed that relate an index—a linear combination of regressors—to the expected response via a nonparametric link function. Although this formulation improves the computational properties, it assumes that a single index underlies consumers' response, and the index often consists of a limited number of variables for computational reasons. In this study we propose a model that addresses these limitations by allowing for multiple indexes, estimating these multiple indexes, and determining the number of indexes to retain.

Specifically, we propose a multi-index binary response (MBR) model and develop a noniterative two-step method to estimate it. We call these two steps *projection* and *calibration*. In the projection step, we combine information available in high-dimensional predictor space and project it into a low-dimensional index space. We achieve this dimension reduction by estimating an index structure (i.e., their composition in terms of original regressors and the number of indexes to retain) without specifying the link function (Cook and Weisberg 1991; Li and Zhu 2006). In the calibration step, we estimate the unknown link function via local polynomial regression (or a parametric model).

The intuition behind these two steps is as follows. The projection step, via sliced average variance estimation (SAVE), retrieves the principal components of a certain covariance matrix constructed from two subgroups of consumers, that is, those for which the dependent variable is 0 and those for which the dependent variable is 1. This covariance matrix incorporates, nonparametrically, the information contained in the dichotomous dependent variable, for example, consumers' choice behavior. SAVE can extract multiple indexes even when the response variable is binary; other approaches (e.g., logistic regression) assume a single index or cannot extract multiple indexes at all (e.g., sliced inverse regression). Because the index structure estimated by SAVE does not depend on the link function, the computational effort is dramatically reduced, making the MBR model applicable to databases with high-dimensional covariates. The calibration step, via multivariate local polynomial regression, estimates a flexible shape of the underlying response probability as a function of the few multiple indexes.

One of the main advantages of the MBR model is that, by allowing for multiple indexes, it facilitates a more refined understanding of consumer behavior. For example, consumers can lie in a two-dimensional plane, $z = (z_1, z_2) \in \Re^2$, if we find empirical evidence for the presence of two indexes. In contrast,

Journal of Business & Economic Statistics
???? 0, Vol. 0, No. 00
DOI: 10.1198/jbes.????.????

consumers must lie on a line (i.e., $z \in \Re^1$) in the existing single-index models. Thus the MBR model augments the scope of possibilities for profiling and targeting consumers.

The article is organized as follows. Section 2 reviews the existing probability models and identifies their limitations in high-dimensional data analyses. Section 3 presents the MBR model and its estimation approach. Section 4 illustrates an empirical application to a customer churn database. Section 5 concludes by suggesting avenues for further research.

2. LITERATURE REVIEW

Here we briefly review parametric, nonparametric, and semiparametric binary choice models to gain insight into their strengths and drawbacks in high-dimensional data analysis and thus motivate our approach.

2.1 Parametric Models

In applications of parametric choice models, the choice probability is estimated as a function of the regressors in the vector x of dimension $p \times 1$. This response probability is $P(y = 1|x) = F(x'\beta)$, where β is a conformable parameter vector and y is a binary random variable. The link function $F(\cdot)$ relates the expected response $P(y = 1|x)$ to the linear predictor $x'\beta$, according to some known parametric function [often a cumulative distribution function (cdf)]. For example, F is the logistic cdf in the Logit model; it is the normal cdf in the Probit model. Such parametric choice models are estimated by iterative quasi-Newton-type methods, which can be time-consuming in the presence of many predictors (large p).

To reduce the number of predictors, stepwise model selection or principal component regression could be applied. But stepwise model selection increases computational time substantially, because an astronomical number of models need to be estimated when there are more than 100 variables ($p \approx 100$). In contrast, principal components regression, although appealing in its ability to reduce the dimensionality of the predictor space, runs the risk of eliminating components that do not explain large variation among predictors but are important in predicting the response. This risk is especially high when the predictor set is large and a few components need to be retained.

2.2 Nonparametric Models

Nonparametric binary models are expressed as $P(y = 1|x) = g(x_1, x_2, \ldots, x_p)$, where the link function $g(\cdot)$ itself is an object of estimation; for example, local polynomial regression may be applied to characterize the shape of g empirically (e.g., Fan, Heckman, and Wand 1995; Simonoff 1996). But these approaches suffer from 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 <50% of its total mass. In a p-variate normal density, the mass within a unit hypercube is about $(0.683)^p$, which tends to 0 rapidly as p increases. For example, when $p = 10$, only 2% of the finite sample is near the center of the density, that is, in the unit hypercube. Consequently, as Silverman (1986, p. 92) cautioned, "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). Because of this empty-space phenomenon, fully nonparametric models cannot be estimated accurately for more than 10 regressors (and far fewer than that in applications with limited sample size). In addition, the computational power required for these methods even today is prohibitive for applications with massive databases (i.e., large sample sizes and many predictors). This situation led to the development of semiparametric single-index models.

2.3 Semiparametric Single Index Models

A semiparametric binary choice model is given by $P(y = 1|x) = h(x'\beta)$, where both the link function $h(\cdot)$ itself and the vector β are objects of estimation. In this model we overcome the curse of dimensionality because the index $z = x'\beta$, together with the link $h(\cdot)$, serves as the first projective approximation to a general nonparametric function $g(x_1, x_2, \ldots, x_p)$. Semiparametric choice models assume neither a particular shape for the link function, such as F, in parametric models, nor a general link, such as $g(\cdot)$, in nonparametric models, yet introduce sufficient flexibility in estimating consumers' response probability. The benefit of not prespecifying the shape on $h(\cdot)$ is to mitigate the risk of misspecification inherent in parametric models; for example, if the prespecified form in logistic regression differs from the true link, then it fails to estimate response probabilities accurately. Semiparametric approaches, developed by Gallant and Nychka (1987), investigated by Hall (1990), and applied to binary choice by Gabler, Laisney, and Lechner (1993), for example, fall within this class of models (also see Manski 1975, 1985; Cosslett 1983, 1987; Horowitz 1992, 1993, 1998; Gabler, Laisney, and Lechner 1993; Klein and Spady 1993; Lewbel 2000).

Another semiparametric approach is the class of generalized additive models (GAMs) and related models (Hastie and Tibshirani 1986, 1987), which Abe (1999) extended to multichotomous choice. For binary choice models, the GAM specifies $P(y = 1|x) = h_0(\sum_{j=1}^{p} \beta_j h_j(x_j))$, where nonparametric functions h_j $(j = 0, \ldots, p)$ are estimated using iterative algorithms (Abe 1999) and the computations become intensive when p is large (say hundreds of regressors).

Another binary choice model is the isotonic single-index model $P(y = 1|x) = h(x'\beta)$, where $\beta \in \Re^p$ and $h(\cdot)$ belongs to $\Phi : \Re \to [0, 1]$, the space of proper distribution functions. To estimate this model, we need to estimate $h(\cdot)$ and β in the space $\Gamma : \Phi \times \Re^p$. Cosslett (1983) originally proposed an iterative maximization of the likelihood function $L(h, \beta)$. But, this approach requires substantial computational effort to extract just a single index. Moreover, Bult and Wansbeek (1995, p. 388) noted an additional drawback: "The asymptotic distribution is unknown. Therefore, we are not able to give standard errors."

To simplify the estimation and inference of isotonic single-index models, Naik and Tsai (2004) proposed a two-step

approach, which first determines the orientation of a high-dimensional parameter vector via sliced inverse regression (SIR) and then obtains the nondecreasing link function using isotonic regression. The simplicity arises because the estimated $\hat{\beta}_{SIR}$ does not require knowledge of the link function $h(\cdot)$, thus eliminating the iterative nature of Cosslett's (1983) estimation procedure. Standard errors of $\hat{\beta}_{SIR}$ have been provided by Chen and Li (1998, p. 297). A limitation of Naik and Tsai's (2004) two-step estimator is that it can extract only one linear combination of regressors (i.e., $z = x'\beta_{SIR}$). Consequently, the response probability, $P(y = 1|x)$, depends on the index $z_1 = x'\beta_1$ only, and the possibility of multiple indexes, $z_k = x'\beta_k$, $k = 2, \ldots, K$, is ruled out.

In sum, parametric models are not sufficiently flexible to characterize customers' response probability in diverse product markets; nonparametric models suffer from the curse of dimensionality due to the empty-space phenomenon, whereas semiparametric binary choice models ignore multiple indexes. Thus we propose a binary choice model with multiple indexes and a flexible link function, and we present a noniterative estimation approach to estimate it that is suitable for high-dimensional applications based on recent developments in dimension-reduction theory, which we describe next.

3. MULTI–INDEX BINARY RESPONSE MODEL

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

3.1 The MBR Model

Let X denote a large 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 MBR model, which relates Y and X as follows:

$$P(Y = 1|X) = g(X\beta_1, \ldots, X\beta_k, \ldots, X\beta_K), \tag{1}$$

where $Z_k = X\beta_k$ is an index (i.e., a linear combination of the original variables in X), $k = 1, \ldots, K$; K denotes the total number of indexes; and $g : \Re^K \to (0, 1)$ is the link function. The parameter space $B = (\beta_1, \ldots, \beta_K)$, the number of indexes K, and the link function $g(\cdot)$ are unknown (to be estimated). The following remarks further elaborate the MBR model (1) to distinguish it from or relate it to the other models.

Remark 1. Parametric and semiparametric choice models are special cases of the MBR model. Specifically, in parametric models, the shape of $g(\cdot)$ is restricted to the known link functions (e.g., logistic, probit). Moreover, the total number of indexes is restricted to $K = 1$ in both parametric and semiparametric single-index models [e.g., binary single-index (Cosslett 1983) or binary GAM (Abe 1999)].

Remark 2. The nonparametric link of the MBR model is based on the low-dimensional index space $z = (z_1, z_2, \ldots, z_K)$ rather than on the original variables space $x = (x_1, x_2, \ldots, x_p)$, where $K < p$. If we achieve dimension reduction from $p \approx 100$ or more to $K < 10$ or less (say), then nonparametric estimation of $g(\cdot)$ becomes feasible with moderate sample sizes, mitigating the curse of dimensionality.

Remark 3. The MBR model assumes that the regressors X are continuous variables. But some regressors are either discrete-valued (e.g., binary, ordered) or categorical (e.g., nominal or string variables, such as city names). To incorporate discrete-valued regressors into the MBR model, as in factor analysis (see Basilevsky 1994, chap. 8), we conduct a spectral decomposition of their covariance matrix, extract the eigenvectors, and construct factor scores. We then retain all of the factor scores in the MBR model, so we do not lose any covariance information in the predictor set. Furthermore, by construction, the resulting factor scores are continuous, because they are convex combinations of the original regressors (with the weights given by the elements of an eigenvector). For multinomial categorical variables, we recommend using the approach developed by Chiaromonte, Cook, and Li (2002). Finally, we caution that an index model with only discrete variables is not identified (see example 2.3 in Horowitz 1998). For the parameters to be identified, besides the conditions in Remark 8 (also see sec. 2.4.2 in Horowitz 1998), the support of the index must not separate into disjoint sets (see example 2.4 in Horowitz 1998) when we vary the discrete variables in an index comprising both continuous and discrete variables and the link function must not be periodic (see Ichimura 1993 for details).

Remark 4. Approximate factor models (e.g., Bai and Ng 2002), principal component regression with many predictors (e.g., Stock and Watson 2006), and linear or nonlinear principal component models (e.g., Gifi 1990) are seemingly related to the MBR model in (1). The similarity derives from the fact that all involve linear combinations of the predictors; however, the main difference from MBR models is that in the estimation, the MBR models explicitly incorporate the information in a dependent variable in the process of constructing factors or components. In contrast to the estimation procedure for the MBR model (described in the next section), factors are extracted by decomposing a matrix such as covariance, correlation, or product moment, which is based on predictors only, thereby ignoring their possible relationship to the response variable and resulting in a potential loss of predictive accuracy.

In nonlinear principal components, we find linear combinations of the nonlinearly transformed covariates (including binary, ordered, or categorical variables) by maximizing some convex function defined on the space of correlation matrixes (e.g., the sum of the correlation coefficients or the sum of the largest eigenvalues of the correlation matrix). (See de Leeuw 2005 for details and related methods, e.g., multiple correspondence analyses; see PRINQUAL in SAS for its software.) Conceptually, the benefit of nonlinear transformation is that it concentrates more information in the first few principal components (de Leeuw 2005, p. 8). We document empirical evidence for a similar result in the context of inverse regression (see Sec. 4.2.1).

Remark 5. We focus on binary response in the MBR model; however, the model generalizes to multinomial responses, where Y takes L discrete values $\{0, 1, \ldots, L - 1\}$. Specifically, we construct the kernel covariance matrix [see Equation (4)] by summing over all L levels, instead of two, and the remainder of the analyses proceeds as in the binary case. Li and Zhu (2006) established consistency of SAVE for this general setting.

Remark 6. Although Li's (1991) formulation appears similar to the MBR model (1), SIR can estimate at most $\min(p, H-1)$ indexes, where H is the number of slices. When $H = 2$ for binary response, SIR finds only one effective dimension. Thus SIR cannot extract multiple indexes (i.e., $K > 1$), whereas the MBR model can.

Remark 7. Naik and Tsai (2004) developed a two-step approach for the estimation and inference of isotonic single-index models, thereby extending Cosslett's (1983) distribution-free estimation to high-dimensional covariates. But Naik and Tsai's (2004) estimator is based on SIR, so it cannot estimate multiple indexes, such as those in the MBR model (see Remark 6). Table 1 further differentiates the study of Naik and Tsai (2004) from our study.

Remark 8. Identification of index models has been studied in the econometric and statistics literature. Specifically, the single-index model $E[y|x] = h(x'\beta)$ has a unique representation when $h(z)$ is differentiable and nonconstant on the support of the index z, the variables in x are continuously distributed random variables with nondegenerate joint density function and no exact linear dependence among them, and the parameter vector β is normalized to unit length with a positive first element. Alternatively, the first element could be set to $\beta_1 = 1$ with no restriction on the length of the parameter vector. (For further details, see Manski 1988; Ichimura 1993; Horowitz 1998, theorem 2.1.) Recently, Lin and Kulasekara (2007, theorem 1) proved identification of single-index models under weaker conditions of continuity rather than differentiability of $h(z)$. For the partially linear single-index model $E[y|x] = x_A'\beta_A + h(x_B'\beta_B)$, where $x = (x_A', x_B')'$, we also need exclusion restrictions such that variables in x_A are not linear combinations of those in x_B, which implies that, for example, x_A and x_B cannot have common variables (Horowitz 1998, p. 12). Alternatively, we can specify a model $E[y|x] = x'\beta_1 + h(x'\beta_2)$ as done by Lin and Kulasekara (2007, p. 2), thus allowing common variables and replacing the usual exclusion restrictions with the orthogonality condition $\beta_1 \perp \beta_2$, where $\perp$ denotes perpendicular and the β_k's are conformable parameter vectors (see Lin and Kulasekara 2007, theorem 2, for details). Donkers and Schafgans (2005, assumption 3) showed how orthonormality of the space $B = (\beta_1, \beta_2)$ (i.e., $B'B = I_{2\times 2}$) substitutes for the usual normalization and exclusion restrictions.

For multi-index models such as $E[y|x] = \tilde{g}(x_1\beta_1, \ldots, x_K\beta_K)$, Ichimura and Lee (1991) furnished the identification conditions, which consist of scale normalization (i.e., $\|\beta_k\| = 1$ with positive first element for each k), exclusion restrictions on the variables across various indexes, differentiability of $\tilde{g}(z_1, \ldots, z_K)$, and no linear dependence among the set of partial derivatives $\dot{g}_k = \partial\tilde{g}/\partial z_k$ (for details, see lemmas 2 and 3 in Ichimura and Lee 1991). More recently, allowing for common variables $x = (x_1', \ldots, x_K')'$, Li (1991) specified the multi-index model $y = \tilde{g}(x'\beta_1, \ldots, x'\beta_K, \varepsilon)$ and proposed the concept of the effective dimension-reduction (EDR) space, $S(B)$, which is spanned by the columns of the $p \times K$ matrix $B = (\beta_1, \ldots, \beta_K)$, where $\beta_k \perp \beta_{k'}$ for every (k, k'), $k \neq k'$, $\|\beta_k\| = 1$, so that $B'B = I_{K\times K}$, and the error term ε is independent of x. We note that β_k is not individually identified; rather, the EDR space $S(B)$ is identifiable (Li 1991). Specifically, when x has a positive density over a convex support and $\tilde{g}$ is a nonconstant continuous function, this EDR space is uniquely identified up to scale normalization (see Li 1991, p. 318; Cook 1998, p. 113; Xia et al. 2002, p. 364; Li, Zha, and Chiaromonte 2005, p. 1581). Cook (1998, chap. 6) has provided further details on the notion of dimension-reduction spaces and the central subspace formed by their intersection. Thus the index space $S(B)$ is uniquely defined, and "as soon as the operator B which maps $\Re^p$ onto $\Re^K$ is fixed, the link function $\tilde{g}$ can be estimated in a nonparametric way" (see Hristache et al. 2001, p. 1538; notation ours). Finally, to identify K as the minimal number of indexes, Donkers and Schafgans (2005) required both $\text{rank}(B) = K$ (to rule out multicollinearity of the indexes) and $\text{rank}(D) = K$ (so that each index provides unique information on the shape of g), where the $K \times K$ matrix $D = \{d_{k,k'}\}$, $d_{k,k'} = E[\frac{\partial\tilde{g}}{\partial z_k}\frac{\partial\tilde{g}}{\partial z_{k'}}]$, for all $(k, k') \in \{1, \ldots, K\}$.

Next we present our approach to estimating the MBR model.

Table 1. Main differences between the study of Naik and Tsai (2004) and the present study

Features	Naik and Tsai (2004)	Present study
Link function	Monotonic	General
Number of factors in the model	One	Multiple
Variable types in the analysis	Continuous	Continuous and discrete
Estimation method	SIR and isotonic regression	SAVE and local polynomial
Can extract multiple factors?	No	Yes
Bandwidth selection	NA	Equation (13)
Determination of number of factors	NA	Equations (6) and (7)
Empirical application	Catalog mailing decision	Customer defection (churn) detection
Models comparison using out-of-sample data	No	Yes
Forecasts combination	No	Yes
Novel substantive finding and implication	Logistic distribution overstates the top decile probability; consider using nonparametric distribution function	Prospective customers exist in disconnected sets (see Figure 4); consider using multifactor models

3.2 Estimation Approach

The estimation problem is to discover the index structure and its relationship to the binary response variable. The index structure consists of the number of indexes K and the EDR space $B = (\beta_1, \ldots, \beta_K)$. In addition, we need to estimate the link function $g(z_1, z_2, \ldots, z_K)$, where $z_k = x'\beta_k$ are the index scores. The objects (g, B) belong to the space $\Gamma : \Omega \times \Re^{p\times K}$, and we can estimate them *without* iterating between the function space $\Omega : \Re^K \to [0, 1]$ and the parameter space $\Re^{p\times K}$. The noniterative approach entails obtaining $\hat{B}$ in a high-dimensional regressor space using SAVE (Cook and Weisberg 1991) without knowing g and then finding $\hat{g}$ via local polynomial regression in a low-dimensional projected subspace (e.g., Fan and Gijbels 1996; Simonoff 1996). The resulting estimates are consistent via the theorems of inverse regression (e.g., Li 1991; Cook and Lee 1999; Li and Zhu 2006) and nonparametric statistics (Fan, Heckman, and Wand 1995; Pagan and Ullah 1999).

3.2.1 Estimating the Indexes. Let $\tilde{x} = \Sigma_x^{-1/2}(x - \mu)$ be the standardized x, where $\mu = E(x)$, and $\Sigma_x^{1/2}$ denotes the Cholesky factor of $\Sigma_x = \text{cov}(x)$. We transform the data matrix X to $\tilde{X} = \hat{\Sigma}_x^{-1/2}(X - \bar{X}')$, where $\hat{\Sigma}_x = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(X_i - \bar{X})'$, X_i' denotes the ith row, $\bar{X}$ contains the sample means, and N is the sample size. The resulting $\tilde{X}$ has zero means, unit variances, and uncorrelated variables.

The parameter estimates in (1) are given by $\beta_k = \Sigma_x^{-1/2}\eta_k$, where the direction vectors η_k are obtained from the eigenvectors of the decomposition

$$M\gamma_k = \lambda_k\gamma_k, \tag{2}$$

for $k = 1, \ldots, p$. In (2), λ_k is the kth eigenvalue and γ_k is the associated eigenvector of the kernel matrix $M = E(I - \text{cov}(\tilde{x}|y))^2$, where $\text{cov}(\tilde{x}|y)$ is the conditional covariance of $\tilde{x}$ given y.

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

$$\hat{V}_s = \frac{1}{N_s}\sum_{i=1}^{N_s}(\tilde{X}_{si} - \bar{\tilde{X}}_s)(\tilde{X}_{si} - \bar{\tilde{X}}_s)', \tag{3}$$

where $\tilde{X}_{si}'$ denotes the ith 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 over the two slices yields the kernel covariance matrix

$$\hat{M} = \sum_{s=0}^{1}\hat{p}_s(I - \hat{V}_s)(I - \hat{V}_s)', \tag{4}$$

where $\hat{p}_s$ is the proportion of observations in slice s. Finally, we obtain $\hat{\eta}_k = \hat{\gamma}_k$ by substituting $\hat{M}$ into (2) and solving the resulting eigenvalue problem.

Li and Zhu (2006) investigated the asymptotics of SAVE and proved its consistency. Specifically, they proved that

$$\sqrt{N}(\hat{M} - M) \Rightarrow Q, \tag{5}$$

where "$\Rightarrow$" denotes convergence in distribution and Q is the limiting random matrix such that $\text{vec}(Q) \sim \text{N}(0, \Sigma_Q)$, where $\Sigma_Q = \text{cov}(\text{vec}(V(Y, z)))$, which was defined by Li and Zhu (2006, theorem 2.3). Li and Zhu (2006, theorem 3.3) also established consistency of the eigenvalues and eigenvectors of $\hat{M}$. They also conducted Monte Carlo studies to support their theoretical results in small samples.

To gain intuition for SAVE, we observe that the index $z_k = x'\beta_k$ is a new variable formed by combining the regressors, where the weights come from the elements of the eigenvector $\hat{\gamma}_k$. The new variables z_k and $z_{k'}$ are uncorrelated with each other $(k \neq k')$, as in principal components analysis, where the eigendecomposition in (2) is based on the covariance matrix $M = \Sigma_x$. Thus we refer to an index z_k as a "factor" and the elements of an eigenvector $\hat{\gamma}_k$ as "factor loadings," which indicate the composition of factors in terms of the original regressors. The main difference, however, is that the kernel matrix M for SAVE depends on both X and Y through (V_s, p_s) in Equations (3) and (4), whereas M in principal components analysis depends solely on X and is unrelated to the binary response variable (see Remark 4). Thus, intuitively, SAVE extracts predictive factors that are functions of the response variable.

Remark 9. Our proposed approach has three main advantages. First, we reduce the number of regressors in x from a large p to small K *without* using a model-fitting process (either parametric or nonparametric). Second, we achieve this dimension reduction without the knowledge of the link function g in Equation (1). Several existing estimators (e.g., average derivative estimation, outer product of gradients) use information in the gradient ∇g to estimate nonparametrically the parameter space B (see, e.g., Hardle and Stoker 1989; Samarov 1993; Hristache et al. 2001; Xia et al. 2002; Donkers and Schafgans 2005). In contrast, we estimate B based on the eigenvectors of $\hat{M}$, which is constructed without estimating g or its derivatives (either parametrically or nonparametrically), thus imposing weaker assumptions on the conditional distribution of $y|x$. Third, we determine the number of indexes K to be retained using the plug-in formula in Equation (6), which uses the eigenvalues already obtained while estimating B, thereby avoiding the numerically intensive approaches based on gradient information $\nabla\tilde{g}$, cross-validation, or resampling, all of which are computationally expensive when dimensionality is high ($p \approx 100$).

Remark 10. The factors extracted by SAVE can be, but need not be, interpretable. For a clean interpretation of the factors, users should have ex ante knowledge of the composition of latent constructs (i.e., the factor structure) in terms of the original variables; for example, the first factor consists of the variables x_1, x_3, and x_5, whereas the second factor consists of the variables x_2, x_4, and x_6. When prior knowledge of factor structure is available, constrained inverse regression (CIR; Naik and Tsai 2005) is the optimum procedure.

In CIR, the information about the factor structure is incorporated as constraints on model parameters. For example, if the first factor excludes x_2, x_4, and x_6, then we can express this knowledge as $A_1'\beta = 0$, where the constraint matrix

$$A_1' = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and β is the 6×1 parameter vector in this example. Similarly, we can express the second factor's structure as $A_2'\beta = 0$, where

$$A_2' = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix},$$

so that the first, third, and fifth variables are excluded from the second factor by construction. Then, to extract factors subject to such linear homogenous constraints, we solve the eigenvalue problem in (2) by replacing M by $(I - P_c)M$, where $P_c = A_c(A_c'A_c)^{-1}A_c'$ is the projection matrix for the constraint matrix A_c $(c = 1, 2)$. Naik and Tsai (2005) developed the theory of constrained inverse regression, provided simulation-based evidence, and illustrated its applicability for a brand logo design study. Thus factors can be extracted that satisfy the prespecified factor structure such that the extracted factors are interpretable and meaningful to users. Future research could investigate the extension of CIR to binary response models.

3.2.2 Determining the Number of Indexes. Because $\hat{M}$ is full rank, all of its eigenvalues exist, so we can potentially extract p factors (or indexes). We can determine the number of factors to retain, K^*, via permutation tests or information criteria. Cook and Yin (2001) developed permutation tests, which rely on resampling the data; one of these tests estimates the upper bound $\tau > K^*$ (i.e., the maximum number of factors likely to be present in the population). Alternatively, Zhu, Miao, and Peng (2006) developed an information criterion and established its consistency. Specifically, their information criterion is given by

$$G(k) = \frac{N}{2} \sum_{i=\min(\tau,k)+1}^{p} (\mathrm{Ln}(\hat{\theta}_i) + 1 - \hat{\theta}_i) - \frac{C_N k(2p - k + 1)}{2}, \tag{6}$$

where $\hat{\theta}_i = \hat{\lambda}_i + 1$, $\hat{\lambda}_i$ is the ith eigenvalue of $\hat{M}$, and C_n depends on sample size [e.g., $\mathrm{Ln}(N)$ as in the Bayes information criterion (BIC)]. Then the number of factors to be retained is given by

$$K^* = \arg\max G(k), \tag{7}$$

over $k \in \{0, 1, \ldots, p - 1\}$.

Zhu, Miao, and Peng (2006, theorem 2) established the consistency of K^* by proving weak and strong convergence (see Bai and Ng 2002 for approximate factor models). They also conducted Monte Carlo studies and reported satisfactory performance in recovering the correct number of factors in small samples.

The intuition underlying (6) is as follows. We retain the factors whose eigenvalues, $\hat{\lambda}_i$, are significant; equivalently, we keep the factors such that $\hat{\theta}_i = \hat{\lambda}_i + 1$ exceeds unity. In other words, we test whether $\theta_1 > \cdots > \theta_k > 1$ and $\theta_{k+1} = \cdots = \theta_p = 1$. The first term on the right side of (6) measures the goodness of fit, whereas the second term penalizes model complexity (i.e., number of free parameters in M). Specifically, when k is fixed, the $p \times p$ matrix M has rank k, so the last $(p - k)$ columns are linear combinations of the first k columns. Of the $p \times k$ parameters in M, as many as $k(k - 1)/2$ are duplicated due to symmetry. Thus the number of free parameters equals $pk - k(k - 1)/2 = k(2p - k + 1)/2$. Thus, Equation (6) balances the fit and parsimony, while Equation (7) attains the best trade-off.

Before deriving the asymptotic distribution of factor loadings, we emphasize the novel property of SAVE: Neither the estimation of factor composition via (2) nor the determination of the number of factors in (7) requires a priori knowledge of the nonlinear link function g in the model Equation (1).

3.2.3 Asymptotic Distribution of Factor Loadings. To derive the asymptotic distribution of the loadings, we apply perturbation theory (Tyler 1981; Kato 1995), which shows how small perturbations of a matrix affect its eigenvalues and eigenvectors. Specifically, if we perturb an element a_{ij} of the matrix A, by a small value denoted by $\frac{\partial A}{\partial a_{ij}}$, then the corresponding change in its kth eigenvector e_k is given by

$$\frac{\partial e_k}{\partial a_{ij}} = \sum_{\substack{l=1 \\ l \neq k}}^{p} \frac{e_l e_l'}{\lambda_k - \lambda_l} \times \frac{\partial A}{\partial a_{ij}} \times e_k.$$

We know that the perturbation $(\hat{M} - M)$ is small for large N [see Equation (5)]; thus the associated change in an eigenvector is given by

$$\sqrt{N}(e_k(\hat{M}) - e_k(M)) = \sum_{\substack{l=1 \\ l \neq k}}^{p} \frac{e_l e_l'}{\lambda_k - \lambda_l} \times \sqrt{N}(\hat{M} - M) \times e_k$$

$$\Rightarrow \tilde{C}_k Q e_k, \quad \text{where } \tilde{C}_k = \sum_{\substack{l=1 \\ l \neq k}}^{p} \frac{e_l e_l'}{\lambda_k - \lambda_l}$$

is a $p \times p$ matrix, Q is the random matrix in (5), and $e_k(M)$ is the kth eigenvector of M. Because $e_k(\hat{M})$ is the estimated eigenvector $\hat{\gamma}_k$ obtained using (2), we get

$$\sqrt{N}(\hat{\gamma}_k - \gamma_k) \Rightarrow C_k Q \gamma_k, \quad \text{where } C_k = \sum_{\substack{l=1 \\ l \neq k}}^{p} \frac{\gamma_l \gamma_l'}{\lambda_k - \lambda_l}.$$

Next, to find the variance of $\hat{\gamma}_k$, we note that $\mathrm{vec}(C_k Q \gamma_k) = (\gamma_k' \otimes C_k)\,\mathrm{vec}(Q)$. Given $\mathrm{vec}(Q) \sim \mathrm{N}(0, \Sigma_Q)$ from (5), we get the standard errors of the factor loadings by taking the square-root of the diagonal elements of the $p \times p$ matrix,

$$N^{-1}(\gamma_k' \otimes C_k)\Sigma_H(\gamma_k \otimes C_k'), \qquad k = 1, \ldots, K^*, \tag{8}$$

where Σ_Q is a $p^2 \times p^2$ matrix (see Li and Zhu 2006, theorem 2.3). Equation (8) provides the asymptotic distribution of factor loadings of SAVE, which furnishes a new result to facilitate inference. Specifically, using Equations (2) and (4), we can estimate $(\hat{\lambda}_k, \hat{\gamma}_k)$ for all k and thus compute $\hat{C}_k$. If a consistent estimator of $\hat{\Sigma}_Q$ were available (which we hope can be derived in future studies), we can apply (8) to test for the significance of the factor loadings, $\hat{\gamma}_k$. Because we transformed the original X data matrix [where $X \sim \mathrm{N}(\mu, \Sigma_x)$] to $\tilde{X}$ [which follows $\mathrm{N}(0, I)$], $\hat{\gamma}_k$ serves as the estimator for β, which can be recovered in original units via $\beta_k = \Sigma_x^{-1/2}\gamma_k$. In other words, we have one-to-one correspondence between them, and thus the significance of the factor loadings in $\hat{\beta}_k$ follows directly from those of the corresponding elements of $\hat{\gamma}_k$ (i.e., the t-ratios of $\hat{\gamma}_k$ and $\hat{\beta}_k$ are identical).

In summary, for statistical inference, we first solve the eigenvalue problem in (2) using (4) to obtain $\hat{\gamma}_k$ (which can be transformed back if necessary to obtain the point estimates of $\hat{\beta}_k = \hat{\Sigma}_x^{-1/2}\hat{\gamma}_k$). We then extract the diagonal matrix in (8), take the square root of its elements to obtain the standard errors, and compute the t-ratios by dividing the estimates of $\hat{\gamma}_k$ with its corresponding standard errors. The resulting t-ratios facilitate testing for the significance of $\hat{\beta}_k$.

3.2.4 Estimating the Link Function. After determining the number of factors and their composition, we compute the factor scores $\tilde{Z} = (\tilde{Z}_1, \ldots, \tilde{Z}_{K^*}) = (\tilde{X}\hat{\eta}_1, \ldots, \tilde{X}\hat{\eta}_{K^*})$. If K^* is small, then we can apply nonparametric methods to estimate the unknown link function in (1). But many nonparametric methods suffer from boundary problems; the response becomes flat when $\tilde{Z}$ is close to the ends of their range. Local polynomial regression not only mitigates such boundary effects (Fan 1992; Fan and Gijbels 1996, p. 60), but also is less sensitive to bandwidth choice compared with other nonparametric methods (e.g., ordinary kernel regression). Thus we recommend using it to calibrate $g(\cdot)$ in the MBR model Equation (1).

We predict the response probability $\hat{g}(t_1, \ldots, t_{K^*})$ for any customer located at the point $t = (t_1, \ldots, t_{K^*})'$ in the K^*-dimensional factor space by applying local linear regression (i.e., local polynomial with degree 1) to the sample of observations, $(Y_i, \tilde{Z}_i')$, $i = 1, \ldots, N$. Specifically, we estimate the probability $P(Y = 1|\tilde{Z})$ by

$$\hat{E}[Y = 1|t] = \hat{g}(t_1, \ldots, t_{K^*}). \tag{9}$$

Toward this end, we first create the design matrix of dimension $N \times (K^* + 1)$,

$$X_D = \begin{bmatrix} 1 & \tilde{Z}_{11} - t_1 & \cdots & \tilde{Z}_{1K^*} - t_{K^*} \\ \vdots & \vdots & & \vdots \\ 1 & \tilde{Z}_{N1} - t_1 & \cdots & \tilde{Z}_{NK^*} - t_{K^*} \end{bmatrix}, \tag{10}$$

and the $N \times N$ diagonal weight matrix,

$$W = \text{diag}\{w_i(t)\}, \tag{11}$$

using the multivariate kernel $w_i(t) = b^{-1}\exp(-(\tilde{Z}_i - t)'(\tilde{Z}_i - t)/b)$, where b denotes the bandwidth. For a prospective customer located at the point $t = (t_1, \ldots, t_K^*)'$, the function $w_i(t)$ assigns positive weights to every observation i. The closer the observation $\tilde{Z}_i$ is to a customer t, the greater the weight $w_i(t)$. Then the least squares estimates are given by (e.g., Fan and Gijbels 1996, p. 298)

$$\hat{m} = \begin{bmatrix} \hat{m}_0 \\ \hat{m}_1 \\ \vdots \\ \hat{m}_{K^*} \end{bmatrix} = (X_D'WX_D)^{-1}X_D'Wy, \tag{12}$$

where $y = (Y_1, \ldots, Y_N)'$. Finally, $\hat{m}_0$, the first element of $\hat{m}$ in Equation (12), furnishes the desired response probability $\hat{g}(t_1, \ldots, t_{K^*})$.

Because $\hat{m}$ is a closed-form estimator, the computation of $\hat{g}(t)$ does not involve iterations. Furthermore, the bandwidth b is scalar, because the factor scores are orthogonal with unit variances [i.e., $\text{Cov}(\tilde{Z}) = I_{K^*}$]. However, as usual, the choice of the bandwidth affects both the in-sample and out-of-sample performance. Specifically, when the bandwidth b is small, the model approximates the in-sample data accurately (low bias), but the estimate has large variance because a few observations fall within the small window. This high variability could adversely affect out-of-sample forecasts. When b is large, the bias increases but the estimate is more stable, which could improve the out-of-sample performance.

For bandwidth selection, theoretical approaches balance this trade-off between bias and variance by minimizing the mean squared error, but the asymptotic results depend on unknown population quantities. The optimal bandwidth has order $b^* = O(N^{-1/(d+4)})$, so, as recommended by Fan and Gijbels (1996), we can set $b = 0.79\sigma N^{-1/(d+4)}$ as a reasonable choice, where σ is the average range of estimated $\tilde{Z}$ and d denotes the number of dimensions for smoothing. Alternatively, approaches based on cross-validation or information criteria can be used. For example, by evaluating the expected Kullback–Leibler divergence between the true and candidate models, Naik and Tsai (2001) derived an information criterion that we can apply to determine the bandwidth of the link function in multi-index models,

$$AIC_C(b) = -2\,\text{Ln}(L_b) + \frac{N(N + p_b)}{N - p_b - 2}, \tag{13}$$

where $\text{Ln}(L_b) = \sum_{i=1}^{N}[Y_i\,\text{Ln}(\hat{g}_{i,b}) + (1 - Y_i)\,\text{Ln}(1 - \hat{g}_{i,b})]$ for MBR models and p_b denotes the "effective number of parameters" measured by $\text{tr}(Q_b)$, where Q_b is the *hat* matrix in the projection $\hat{y} = Q_b y$. We use (12) to construct Q_b, whose ith row is $e'(X_{D,i}'WX_{D,i})^{-1}X_{D,i}'W$, where $e = (1, 0, \ldots, 0)'$ is a conformable unit vector and $X_{D,i}$ evaluates (10) at $t = \tilde{Z}_i$. Formally, we select the bandwidth that minimizes the Akaike information criterion $AIC_C(b)$. We also modify (13) to obtain a BIC-type criterion by replacing the penalty term by $p_b\,\text{Ln}(N)$. Intuitively, as the bandwidth decreases, the first term on the right side of (13) decreases, thereby decreasing the deviance $[\propto -2\,\text{Ln}(L_b)]$, whereas the second term increases to penalize the resulting increased model complexity $[\propto \text{tr}(Q_b)]$ relative to the sample size.

If AIC_C and BIC suggest similar bandwidths, then a typical bandwidth can be used (e.g., average); however, if they suggest different bandwidths, then the forecasts resulting from several MBR models estimated using multiple bandwidths (Bates and Granger 1969) can be combined. Specifically, we generate the forecasts $P_j(Y = 1) = \hat{g}_{b_j}(\cdot)$, where $\hat{g}_{b_j}(\cdot)$ is the MBR model with bandwidth b_j for $\{b_j\}, j = 1, \ldots, J$. Then we average $\hat{g}_{b_j}$ across j to obtain the combined forecasts (see "Forecasts Combination" in Sec. 4.3 for details).

Table 2 summarizes the noniterative algorithm for estimating MBR models, which we next apply to a real data set from customer relationship marketing (CRM).

4. CUSTOMER RELATIONSHIP MARKETING APPLICATION

CRM places individual customers at the heart of a company's strategies, hoping to establish and strengthen relationships with them by better understanding the motives for their individual choices (Day 2000; Winer 2001). The database of customers' transactions, which many companies compile, is central to achieving such goals, because it allows marketers to explore customer behavior at finer levels of granularity. Although such

Table 2. 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.
2. Partition $\tilde{X}$ into $\tilde{X}_0$ and $\tilde{X}_1$. The submatrix $\tilde{X}_0$ contains customers with $y = 0$, and $\tilde{X}_1$ contains those with $y = 1$. The conditional covariance matrix is $\hat{V}_s = \frac{1}{N_s}\sum_{i=1}^{N_s}(\tilde{X}_{si} - \bar{\tilde{X}}_s)(\tilde{X}_{si} - \bar{\tilde{X}}_s)'$, where $\tilde{X}'_{si}$ is the ith row, $\bar{\tilde{X}}_s$ is the sample mean, $\hat{p}_s$ is the proportion, and N_s is the sample size in slice, $s = \{0, 1\}$.
3. The parameter estimates are $\hat{\beta}_k = \hat{\Sigma}_x^{-1/2}\hat{\eta}_k$. Compute $\hat{M} = \sum_{s=0}^{1}\hat{p}_s(I - \hat{V}_s)(I - \hat{V}_s)'$ as in Equation (4) and solve the eigenvalue problem $M\gamma_k = \lambda_k\gamma_k$ as in Equation (2). The resulting eigenvectors $\hat{\gamma}_k$, $k = 1, \ldots, K$, are equal to $\hat{\eta}_k$, which can be expressed in the original units via $\hat{\beta}_k = \hat{\Sigma}_x^{-1/2}\hat{\eta}_k$.
4. Determine K^*, the number of factors to retain, using (6) and (7). Specifically, for each k, compute $G(k) = \frac{N}{2}\sum_{i=\min(\tau,k)+1}^{p}(\mathrm{Ln}(\hat{\theta}_i) + 1 - \hat{\theta}_i) - \frac{C_N k(2p-k+1)}{2}$, where $\hat{\theta}_i = \hat{\lambda}_i + 1$, $\hat{\lambda}_i$ is the ith eigenvalue of $\hat{M}$, and $C_n = \mathrm{Ln}(N)$. Then choose K^* that maximizes $G(k)$.
5. Compute the factor scores $\tilde{Z}_k = \tilde{X}\hat{\eta}_k$ for each index k, $k = 1, \ldots, K^*$.
6. Apply local linear regression to estimate $\hat{m} = (X_D'WX_D)^{-1}X_D'Wy$ as in Equation (12), with the design matrix X_D given in (10), and the kernel weight matrix $W = \mathrm{diag}\{w_i(t)\}$ given in (11), where $w_i(t) = b^{-1}\exp(-(\tilde{Z}_i - t)'(\tilde{Z}_i - t)/b)$ and b is the bandwith. Save the first element of $\hat{m}$, $\hat{m}_0$, which furnishes the response probability $\hat{g}(t_1, \ldots, t_{K^*})$ for an out-of-sample customer at $t = (t_1, \ldots, t_{K^*})'$ in K^*-dimensional factor space.

databases contain transactions with the focal company, the lack of complete information to explain choices prompts marketers to include additional explanatory variables from other sources. For example, car and voter registrations provide information on age, name, address, and telephone; county records reveal information on home value; census surveys yield geodemographic data; credit card companies reveal spending patterns; and airline companies know travel patterns. By analyzing the augmented databases, marketers attempt to develop more effective CRM programs. But this wealth of information brings new challenges; the size of the databases—not only the number of customers, but especially the number of variables to be used as predictors—explodes from a few to hundreds. Next we analyze a CRM database from a telecom company, present the estimation results from the MBR model, and describe its out-of-sample performance.

4.1 Churn Data Description

In this empirical application, we study the case of a telecom company that seeks to predict customer defection (i.e., churn) using many variables from customers' past transactions and usage behavior, available from the firm's customer database. The Teradata Center for CRM at the Duke University provided the calibration and validation data sets (see Lemmens and Croux 2006; Neslin et al. 2006 for details). The calibration data set contains 100,000 customers and 171 variables; this sample is balanced to represent 50% churners (i.e., customers that defected from the company) and 50% nonchurners. To handle missing values, we delete variables with >50% missing values or string variables (e.g., city names) and then impute the few missing values using averages of the corresponding variables. For model estimation, we randomly select 10,000 customers across 104 continuous variables and 20 discrete variables (e.g., binary, ordered). Following Basilevsky (1994, chap. 8), we transform 20 discrete variables into continuous variables by conducting factor analysis of their covariance matrix and retain all 20 factors so that we do not lose any covariance information (which follows from the spectral decomposition theorem). Thus the resulting data comprise 124 regressors and a single binary response.

These regressors provide three kinds of information: customer behavior (e.g., average monthly minutes of use over the previous 3 months, the total revenue of a customer account, trends in usage), company interaction (e.g., calls to customer center), and customer demographics (e.g., age, home ownership, lifestyle variables, such as ownership of boat, RV, or motorcycle). Given the long list of regressors, we refer the readers to table 1 of Lemmens and Croux (2006, p. 278) for a description of some variables. The dependent variable, called "churn," is binary and measured by $Y_i = 1$ if the customer i defects and by $Y_i = 0$ otherwise. Next we report the estimation results using the foregoing calibration data and model comparisons based on validation data (described in Sec. 4.3).

4.2 Estimation Results

4.2.1 Variable Transformation and Eigenplots. The histograms of 104 continuous regressors reveal different degrees of asymmetry and peakedness. Following Cook and Lee's (1999) suggestion, to achieve normality, we apply the Box–Cox transform $x_j^{(\lambda_j)} = (x_j^{\lambda_j} - 1)/\lambda_j$ (taking logarithms would imply letting $\lambda_j \to 0$ for all j). Specifically, we translate the domain of each x_j such that its minimum value equals 1 (i.e., $x_j \leftarrow x_j - \underline{x}_j + 1$), and then estimate $\hat{\lambda}_j$ by maximizing the likelihood,

$$L(\lambda_j) = -(N/2)\,\mathrm{Ln}\left(\sum_{i=1}^{N}\left(x_{ij}^{(\lambda_j)} - \mu_{\lambda_j}\right)\right) + (\lambda_j - 1)\sum_{i=1}^{N}\mathrm{Ln}(x_{ij}),$$

where $\mu_{\lambda_j} = (1/N)\sum_{i=1}^{N} x_{ij}^{(\lambda_j)}$. After rounding off the estimates, we present the frequency of estimated power transforms in Figure 1. We find that in this application, the logarithmic transform is valid for about half of the variables, whereas the square root ($\hat{\lambda} = 0.5$) and linear ($\hat{\lambda} = 1$) transforms are not the best choices for the other half. Thus the shapes of the distributions differ widely across variables, requiring different power transforms for various regressors.

Does the transformation matter? To assess it, we apply eigendecomposition in (2) to the data set with original and transformed regressors. Figure 2 displays the eigenplots for the first 50 out of 124 factors (for the sake of clarity). Comparing the two curves demonstrates a relatively sharper decline for

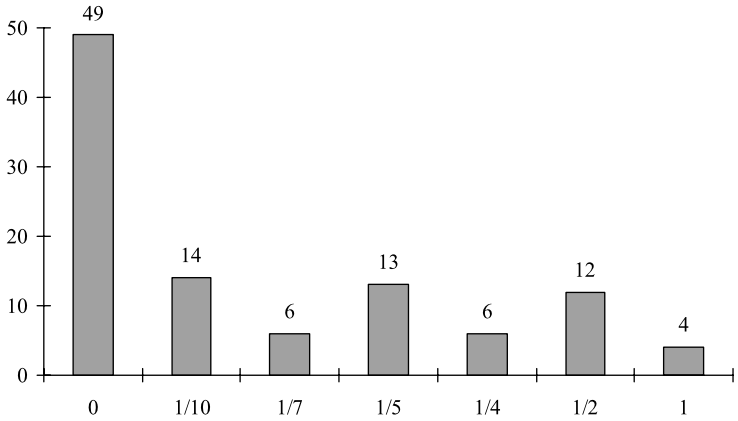

Figure 1. Frequency of estimated power transforms.

the transformed regressors. For example, for the transformed regressors, just 10 out of 104 eigenvalues exceed $\lambda = 0.5$, whereas it takes 25 factors to attain that cutoff (indicated by the dashed line) based on original regressors. Furthermore, a similar result holds for every cutoff level (see Figure 2). Thus using nonlinear transformations of the regressors compress greater information into fewer factors (also see de Leeuw 2005, p. 8).

4.2.2 Number of Factors. Figure 2 further reveals that we cannot determine the number of factors to retain by eyeballing the eigenplot of transformed regressors, because its gradient changes smoothly (unlike the "scree plot" in principal component analysis). Thus we apply the permutation test and information criterion. Cook and Yin (2001, p. 155) developed the permutation test based on resampling the data (see their proposition 2i for its theoretical basis). For this application, this indicates that we retain fewer than $\hat{\tau} = 75$ factors (i.e., the upper bound on the number of factors in the population). Next, based on the information criterion of Zhu, Miao, and Peng (2006), we compute $G(k)$ in (6) and present it in Figure 3. It shows that $G(k)$ attains the maximum value at $K^* = 4$, and thus we retain the first four factors, which strike the optimal balance between fit and parsimony.

This result raises three important points. First, we achieved a marked reduction in dimensionality from $p = 124$ regressors to $K^* = 4$ indexes (or factors) without prespecifying a particular link function. Second, our data provide empirical evidence to support the need for more than a single index (i.e., $K = 1$) to describe binary response variable, justifying the proposed multi-index binary response model. Finally, does substantial dimension reduction lead to severe information loss relative to retaining all of the regressors, ceteris paribus (e.g., the link function)? We address this question in Section 4.3 using out-of-sample forecasts.

4.2.3 Fitting Multivariate Link and Locating Prospective Customers. Nonparametric estimation of $g(\cdot)$ would be practically impossible using the original 124 regressors. However, having projected all of the regressors on a four-dimensional subspace, we can calibrate the churn probability $P(Y = 1) = \hat{g}_b(\tilde{Z}_1, \tilde{Z}_2, \tilde{Z}_3, \tilde{Z}_4)$ nonparametrically. Toward this end, we apply multivariate local polynomial regression of order 1 (i.e., local linear), because even ordered fits (e.g., local constant or quadratic) reduce efficiency and suffer from boundary effects (see Fan and Gijbels 1996, p. 79). Specifically, we apply Equation (12) to the four factor scores and binary response.

To select the bandwidth b, we evaluate the AIC-type information criterion in (13), modified to obtain a BIC-type criterion by replacing the penalty term by $p_b \operatorname{Ln}(N)$. Table 3 presents the

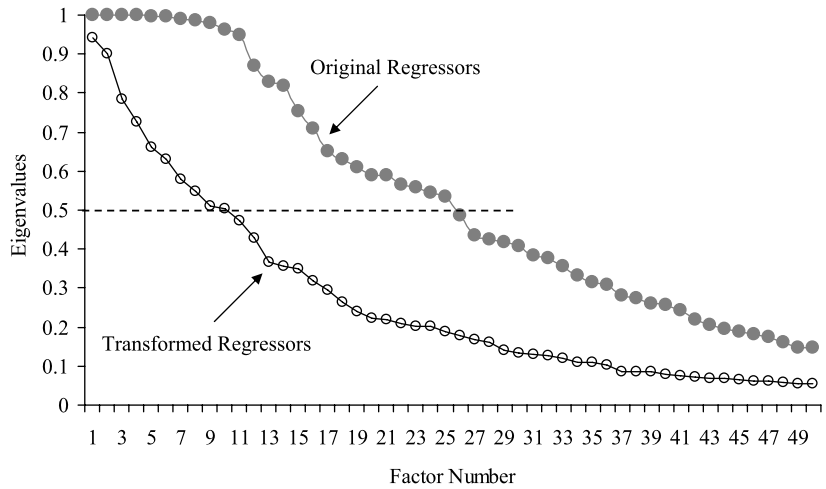

Figure 2. Eigenplot for the first 50 factors.

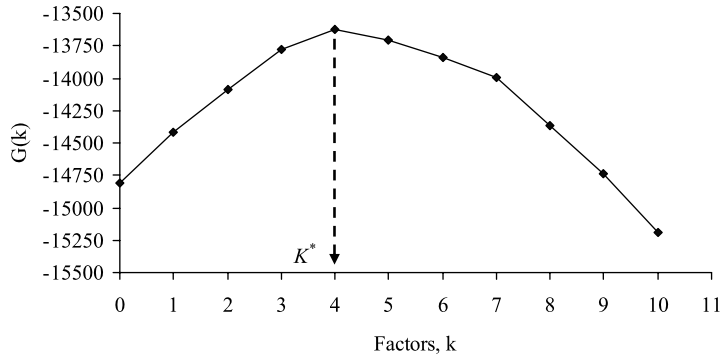

Figure 3. Factor determination using the information criterion.

AIC and BIC and the trace of the hat matrix for various bandwidths. Note that the effective number of parameters (via the trace) decreases as bandwidth increases; however, its impact on AIC_C and BIC is different. AIC_C increases with bandwidth because $\text{tr}(Q_b)$ is small relative to the sample size $N = 10{,}000$, recommending that we select the smallest bandwidth; whereas BIC decreases with bandwidth because $\text{Ln}(N)$ imposes a strong penalty compared with the improvements in the model fit, indicating that we select a larger bandwidth. Given these opposite recommendations, we compute $b = 0.79\sigma N^{-1/5}$, where σ is the average range over the four factor scores $(\tilde{Z}_1, \tilde{Z}_2, \tilde{Z}_3, \tilde{Z}_4)$, yielding the empirical bandwidth $\hat{b} = 7.2$ for this sample.

Without assuming any functional form for $g(\cdot)$, we calibrate the link function using the closed-form estimator (12), the empirical bandwidth in (11), the four estimated factor scores, and the binary response. The resulting link function helps us predict the response probability for *prospective* customers (not just the existing ones). Specifically, consider a prospective customer—not listed in the CRM database—located at an arbitrary point $t = (t_1, t_2, t_3, t_4)'$ in a four-dimensional index space. The chance that this prospective customer churns is given by $\hat{g}(t_1, t_2, t_3, t_4)$. By evaluating the churn probability for customers located in the region $t \in \Re^4$, we identify where high-risk prospects live.

To locate high-risk prospects, we construct contour plots of $\hat{g}(\cdot)$ in two-dimensional spaces. Specifically, for a prospective customers at any point (t_k, t_m) in two-factor space, we represent similar values of $\hat{g}(t_k, t_m; \bar{z}_{-k,-m})$ by a certain color, holding other factors fixed at their sample means (i.e., $\bar{z}_{-k,-m}$). Figure 4 depicts six contour plots resulting from the index combinations $(k, m) = 1, \ldots, 4$ and $k \neq m$. To aid customer identification, we use color graphics to visualize the quintiles of churn probabilities: no-risk (<0.2), low-probability (0.2–0.4), medium-probability (0.4–0.6), high-probability (0.6–0.8), and high-risk (>0.8) customers. Such a visualization is useful in practical applications as well.

We glean four insights from these results. First, we observe that a vast landscape in four dimensions is nearly flat. This suggests that many prospective customers have churn probability close to 0; they exhibit inertia perhaps due to contractual constraints or product satisfaction. The good news is that we know who not to mail the retention promotions to proactively, and know that this constitutes a large proportion of customers, thus leading to substantial savings and avoiding "overtouching" the customer with unnecessary mailings. The framework proposed by Bult and Wansbeek (1995) can be adapted to optimally select customers at risk by maximizing profit based on this MBR model.

Second, our findings corroborate the results of Naik and Tsai (2004), which show that the logistic distribution overestimates customers' response probability. Estimating the unknown link function by nonparametric methods (rather than prespecified) not only mitigates misspecification errors, but also promotes a more conservative approach to database marketing.

Third, 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 Figure 4). In other words, outliers are the norm rather than exception.

Finally, we document evidence that high-risk customers belong to disconnected sets in the four-dimensional latent space. Specifically, Figure 4 shows the sets of customers with no churn risk, moderate churn risk, and high churn risk. Note that the set of moderately risky customers is not contiguous. Existing probability models, regardless whether they fit parametric or nonparametric link functions, cannot identify such disconnected

Table 3. Information criteria and $\text{tr}(Q)$ across bandwidths

Bandwidth, b	AIC_C	BIC	Trace(Q)
0.05	**23,475**	15,201	241
0.1	23,570	14,593	143
0.5	23,760	14,080	45
1	23,804	14,006	28
2	23,816	13,955	20
5	23,826	13,917	13
10	23,832	**13,901**	10

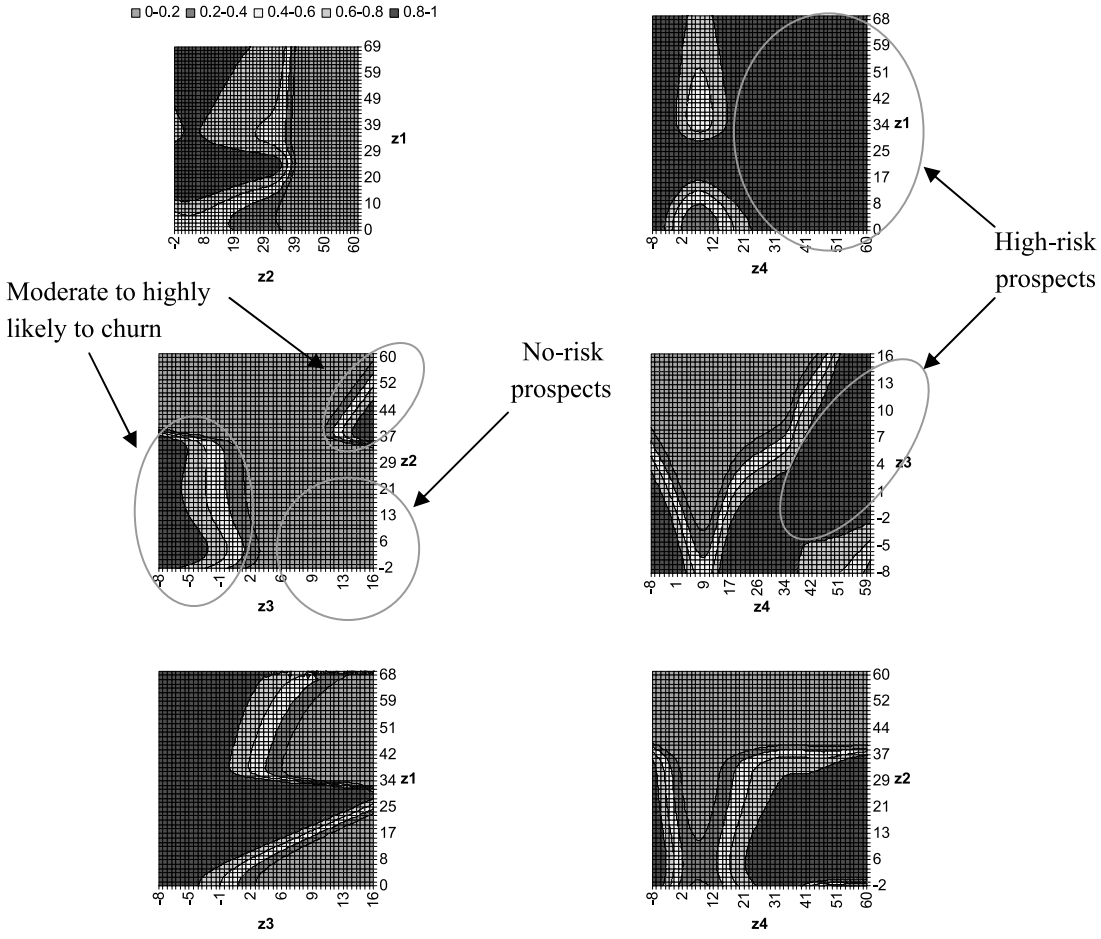

Figure 4. Contour plots $\hat{g}(t_k, t_m; \bar{z}_{-k,-m})$ for $(k, m) = 1, \ldots, 4, k \neq m$.

customer sets, because the response probability is a monotonic function of only one index. To be able to identify different pockets of the same customer type, a topic of great interest in CRM, marketers need to apply multiple-index models, such as the MBR model, to enable more granular profiling and targeting of customer groups.

4.3 Models Comparison

Does dimension reduction induce information loss? How does the MBR model compare with other specifications (e.g., neural network, single-index model)? Can the combination of forecasts resolve bandwidth selection uncertainty? Before we address these three issues, we briefly describe the validation data and the comparison criteria.

The validation data set contains the same variables as the calibration data set, but contains 51,306 different customers. As in Section 4.2, here we use the binary churn indicator, 104 continuous variables, and 20 discrete variables (i.e., their factor scores). We eliminate cases with missing data and then draw $\tilde{N} = 10{,}000$ customers randomly to create the out-of-sample data. The data are sparse, because we observe only 163 churn events from 10,000 customers (i.e., 1.63% defection probability). To compare the out-of-sample forecasts generated by alternative models, we evaluate three kinds of criteria: statistical, intuitive, and managerial. The statistical criterion, proportional to the sample deviance, is twice the negative log-likelihood,

$$-2LogLikelihood = -2\sum_{i=1}^{\tilde{N}}[Y_i \operatorname{Ln}(\hat{Y}_i) + (1 - Y_i)\operatorname{Ln}(1 - \hat{Y}_i)], \quad (14)$$

where Y_i is churn indicator and $\hat{Y}_i$ is the predicted probability for an out-of-sample customer i. The smaller the value of $-2LogLikelihood$, the better the model predicts. Our intuitive criterion is the percentage correct classification, which is given by

$$\%CorrectClassified = 100 \times (1/\tilde{N}) \sum_{i=1}^{\tilde{N}} C_i, \quad (15)$$

where the term C_i refers to a "correct" match. Specifically, there is such a match (i.e., $C_i = 1$) if the observed outcome $Y_i = \tilde{Y}_i$, where $\tilde{Y}$ denotes the discretized probability. We discretize the fractional probability estimates as either 1 or 0 when the estimated probability $\hat{Y}_i$ exceeds a threshold, which equals the mean probability, that is, $(1/\tilde{N}) \sum_{i=1}^{\tilde{N}} \hat{Y}_i$. Thus $C_i = 1$ if $Y_i = \tilde{Y}_i$ and $C_i = 0$ otherwise, where the discretized predicted probability $\tilde{Y}_i = 1$ if $\hat{Y}_i$ exceeds the threshold and $\tilde{Y}_i = 0$ otherwise. The larger the correctly classified percentage, the better the model.

The managerial criterion is the top decile lift (see Lemmens and Croux 2006, p. 280), which is given by the proportion of churners in the top 10% of the sample ($=\pi_{10\%}$) divided by the proportion of churners in the entire validation sample ($=\pi_{100\%}$),

$$TopDecile = \frac{\pi_{10\%}}{\pi_{100\%}}. \tag{16}$$

The greater the lift, the better the model predicts. Top decile lift focuses on "prime" customers because companies target their marketing efforts at this subgroup, thus affecting their profits (see Neslin et al. 2006). When it equals unity, the model has no predictive power, because the targeted customers are as likely to churn as the rest of the sample. But although this criterion is used frequently in practice, we attribute less weight to it, because, unlike the other two criteria, it ignores the accuracy of a classifier across the remaining 90% of the sample and is misleading for models that overestimate the top decile probability (as noted in Naik and Tsai 2004). Next we address the three issues that we raised earlier.

Information Loss. We compare the accuracy of out-of-sample forecasts when the forecasting model uses information contained in the 4 SAVE factors versus all 124 regressors ceteris paribus, i.e., keeping the link function the same. The SAVE factors are obtained by combining the values of the 124 regressors in the validation data with the factor loadings obtained from the estimation sample. We apply logistic regression to the calibration data and estimate the model parameters, which we combine with the regressors (or four SAVE factors) from the validation data to predict the out-of-sample probability, $\hat{Y} = P(Y = 1)$. In other words, we do not use the actual churn indicator Y in the validation sample to estimate the model parameters, and thus our model comparisons are based on true predictions.

Table 4 summarizes the performance with and without dimension reduction. Comparing the rows A and B in Table 4, we find that predictive log-likelihood improves substantially when we use the four SAVE factors relative to all of the regressors. Because the top-decile lifts exceed unity, both the regressors and SAVE factors contain predictive information according to this managerial criterion. We notice a relative drop in the top-decile lift for SAVE, however; this outcome may be due to the local nature of this performance metric, which relies on 10% of the sample and ignores the other nine deciles. Indeed, SAVE factors outperform all 124 regressors, as indicated by the percentage classified correctly, a global measure that accounts for the entire sample. Thus, ceteris paribus, dimension reduction via SAVE engenders no loss of information.

Other Models. By treating the 4 SAVE factors as new regressors, we open the possibilities to include nonparametric models, which otherwise are impractical (if not impossible) to estimate in the presence of 124 regressors. Given this benefit of dimension reduction, we fit four models: neural network with one hidden layer, single-index model, and two local polynomial regressions of order 1 with four SAVE factors using a small bandwidth and a large bandwidth. Note that all of these models are based on the SAVE factors, so that our comparison of predictive validity assesses the use of the local polynomial regression smoother with different bandwidths for the link function versus the link functions of the neural network and single-index models.

The neural network model with a single hidden layer is specified by

$$\begin{aligned} P(Y=1) = \exp\big(&\alpha_0 + \alpha_1 \operatorname{Tanh}(X\beta_1) + \alpha_2 \operatorname{Tanh}(X\beta_2) \\ &+ \alpha_3 \operatorname{Tanh}(X\beta_3) + \alpha_4 \operatorname{Tanh}(X\beta_4)\big) \\ &/\big(1 + \exp\big(\alpha_0 + \alpha_1 \operatorname{Tanh}(X\beta_1) + \alpha_2 \operatorname{Tanh}(X\beta_2) \\ &+ \alpha_3 \operatorname{Tanh}(X\beta_3) + \alpha_4 \operatorname{Tanh}(X\beta_4)\big)\big), \end{aligned} \tag{17}$$

where $\operatorname{Tanh}(t)$ is a sigmoid function. In (17), we extract the factors $X\hat{\beta}_k, k = 1, \ldots, 4$, by applying SAVE to the calibration data; we obtain parameter estimates $\{\hat{\alpha}_l\}, l = 0, \ldots, 4$, via logistic regression of Y in the calibration data on the transformed SAVE factors, $\operatorname{Tanh}(X\hat{\beta}_k)$.

The single-index model specification is

$$P(Y = 1) = h(X\beta), \tag{18}$$

where $\hat{h}(\cdot)$ is estimated by local linear regression and $\hat{\beta}_{SIR}$ estimated via SIR. As noted in Remark 6, SIR can only estimate the single-index specification, because its kernel covariance matrix has rank 1 when the response variable is binary.

The two local polynomial regressions are MBR models fitted via (12) with bandwidth $b = 0.79\sigma N^{-1/5}$, where σ is the average range and the average standard deviation of estimated $\tilde{Z}$. As before, we estimate the parameters using calibration data and then produce forecasts based on the estimated parameters and validation data.

Table 4 reports the results (see the rows marked C, D, E, and F). Specifically, the SIR-based single-index model yields the smallest predictive (negative) likelihood, the MBR model with bandwidth $b = 0.1$ attains the highest percentage classified correctly, and the MBR model with bandwidth $b = 7.2$ achieves

Table 4. Forecasting performance of various models

Models	Predictive $-2Log\ likelihood$	Percent correctly classified	Top decile lift
Information loss			
(A) Logistic with 124 original variables	17,026.5	50.5	1.35
(B) Logistic with four SAVE factors	14,116.2	53.2	1.23
Other models			
(C) Neural network with four SAVE factors	14,536.2	51.9	1.35
(D) Single-index model with SIR factor	13,913.0	33.6	1.29
(E) Local linear with four SAVE factors, $b = 7.2$	14,665.0	46.8	1.41
(F) Local linear with four SAVE factors, $b = 0.1$	24,357.8	64.5	0.80
(G) Forecasts combination across bandwidths	14,671.4	56.2	1.10
(H) Forecasts combination across models	12,561.0	58.4	0.92

the biggest top-decile lift. Note that the neural network does not dominate on any of these measures. Thus, inverse regression theory enables marketers to apply nonparametric models effectively, which otherwise would remain impractical to use in the presence of many predictors. Which inverse regression model is favored depends on the choice of performance metric, but either SIR or SAVE with a small bandwith does quite well on these criteria. Note the impressive performance of the latter method on the hold-out hit rate, with >64% of choices predicted correctly. Once again, it is important to note that all of these methods are based on the SAVE factors, and thus the comparison pertains only to a limited aspect (the link function) of the models.

Forecasts Combination. Table 4 indicates that a bandwidth of 0.1 or 7.2 yields different results (cf. rows E and F), which raises the question of which bandwidth to use in practice. To resolve this uncertainty, we could combine predicted probabilities generated by MBR models with different bandwidths. Bates and Granger (1969) first suggested this idea of forecasts combination. We investigate its performance in this context. We operationalize it by generating the forecasts $P_j(Y=1) = \hat{g}_{b_j}(t_1,\ldots,t_4)$, where $\hat{g}_{b_j}(\cdot)$ is the MBR model with bandwidth b_j. For each bandwidth $\{b_j\}, j=1,\ldots,J$, we forecast the probabilities $P_j(Y=1)$. Then we combine these J forecasts using the simple average $P(Y=1) = (1/J)\sum_{j=1}^{J}\hat{g}_{b_j}(t_1,\ldots,t_4)$. The simple average avoids the uncertainties inherent in a weighted average due to estimating the weights and the variability in the resulting estimates (see Timmermann 2006); it is also simple to apply in practice. Indeed, the real value emerges from (a) combining the accuracy of small bandwidths and the stability of large bandwidths, (b) avoiding optimistic or pessimistic (i.e., extreme) forecasts, and (c) not relying on one specific bandwidth choice.

Specifically, we predict the $\tilde{N}\times 1$ vector of probabilities $P_j(Y=1)$ for $J=10$ different bandwidths $\{b_1=0.1,\ b_2=0.5,\ b_3=1,\ b_4=1.5,\ b_5=2,\ b_6=3,\ b_7=4,\ b_8=5,\ b_9=6,\ b_{10}=7\}$, which spans a realistic range from small to large bandwidths (based on AIC_C, BIC, and b^*). We then assess the performance of the "combined" forecast. Table 4 shows that the combined forecast is indeed competitive (cf. row G with row E, F, or A). In addition, we compute the combined forecast by averaging the forecasts from all of the models and report the results in Table 4 (see row H). As expected, this combined models forecast outperforms several models and performs no worse than any individual forecast. We conclude that multi-index binary response models not only perform satisfactorily in out-of-sample predictions, but also, and more importantly, augment the applicability of nonparametric models to high-dimensional covariates.

5. CONCLUSION

We believe that the MBR model holds promise for application to CRM databases with high-dimensional covariates. The estimation approach comprises a projection step and a calibration step, both involving noniterative computations. Consequently, the approach is scalable to databases with many customers and predictors, as illustrated in our application. Another important feature of this approach is its flexibility. It is more flexible than existing parametric choice models, because it allows for a nonparametric link function, which is useful for analyzing sparse choices, that is, a small number of "success" outcomes observed among a large number of customers. In addition, the MBR model formulates and tests for the presence of multiple factors, whereas the existing nonparametric and parametric choice models all assume a single index to underlie customer behavior. Those multiple factors, when present, create new possibilities for targeting and profiling prospective customers.

Our empirical analysis has provided several insights. First, distributional asymmetry and peakedness differ across variables, thereby requiring heterogeneous power transforms (instead of uniformly applying log transform), allowing us to compress greater information into fewer factors. Second, inverse regression not only reduces dimensionality substantially (from 124 regressors to 4 factors), but also engenders virtually no loss of predictive information. Third, the response surface may be flat over much of the index space, with response-variable activity occurring at the boundaries, in disconnected pockets of the data. Finally, in our application the MBR model facilitated the targeting of prospective customers and the visualization of high-dimensional databases.

In our illustrative application, the estimation sample was balanced with respect to the categories of the dependent variable. For unbalanced samples, flexible approaches, such as the one we propose, may benefit even more from their versatile forms at the tails of the distribution; thus our application may be a conservative test of the benefits of the proposed approach. On the other hand, a limitation of the MBR model is that its predictive fit is sensitive to the metrics used for comparisons, the number of factors retained, and the choice of bandwidth; thus caution should be exercised when making these choices in practical applications.

Given that we use an eigenvalue analysis to compress information from 124 predictors into 4 dimensions and we refer to these dimensions as "factors" with "loadings" on the original predictors, one would naturally think of interpreting these factors. We have not attempted to interpret the resulting factor loadings and associate meanings or labels to them for the two reasons. First, such an interpretation would require us—and the reader—to look at a 124×4 loading matrix, the mere presentation of which is rather unwieldy. Second, to be able to interpret the factors, the analyst or manager should have ex ante knowledge of the composition of latent constructs (i.e., the factor structure) in terms of the original variables. The presence of such prior information is likely in a factor analysis of scales designed to measure specific psychological constructs, but less likely in an analysis of data on consumer choice behavior, where the assumption that latent constructs are uniquely responsible for explaining the observed variation in the data is tenuous. When such prior knowledge is available, CIR is the appropriate procedure. In Remark 10, we explain how to use this approach to incorporate meanings and extract only interpretable factors. When such prior knowledge on the factor structure is not available, as in the present application, the eigenvectors are used merely as a convenient way to represent the solution in a lower-dimensional space. This limitation is common to all

factor-analytic methods that lack ex ante knowledge of the latent factor structures, as is generally the case in applications to revealed preference data.

In summary, this study not only broadens the scope of binary choice models to embrace nonlinear, nonparametric, multifactor models in the presence of many predictors, but also provides a practical and computationally feasible approach for estimating factor scores, determining the number of factors, calibrating the nonlinear link function, selecting the bandwidth, and combining out-of-sample forecasts. We conclude by suggesting three avenues for further research:

- *Alternative transformations*. Figure 2 shows that we can concentrate more information into fewer factors when we attain normality, which we achieve through univariate power transformation of each regressor. Alternatively, we could attain multivariate normality through the approach of Velilla (1993). Similarly, through the approach of Cook and Nachtsheim (1994), we could induce more general elliptically contoured distributions for the regressors rather than restricting them to be normal. Future research may investigate the costs and benefits of these alternative transformations.
- *Asymptotic distribution of factor loadings*. We derived the asymptotic distribution of factor loadings of SAVE to obtain a new result, which can be used to conduct statistical inference on factor loadings in applied research. However, to apply this result, future studies need to construct a consistent estimator of Σ_Q.
- *Multivariate response*. The MBR model predicts a single transaction, for example, the churn indicator, which is a single Y variable. Its extension to multiple correlated response variables is of substantial interest, because this would enable companies to improve their cross-selling efforts across multiple products and services (Kamakura et al. 2003). Toward this end, future research could apply the ideas in "Alternating SIR," developed by Li et al. (2003), which reduces the dimensionality of both $Y \in \Re^m$ and $X \in \Re^p$ simultaneously without specifying the vector-valued link functions. An analysis similar to alternating SIR, but through the kernel matrix from SAVE, would facilitate the inclusion of multiple response variables in MBR models.

We hope that this study marks a useful starting point for further development and that it stimulates researchers to join in the effort to advance the theory and practice of high-dimensional data analyses.

ACKNOWLEDGMENTS

The first two authors appreciate the financial support of the Teradata Center at Duke University for this research. The authors acknowledge that the Editor, Area Editor, and the reviewers provided constructive comments, which improved the manuscript.

[Received July 2007. Revised January 2008.]

REFERENCES

Abe, M. (1999), "A Generalized Additive Model for Discrete Choice Data," *Journal of Business and Economic Statistics*, 17 (3), 271–284.

Bai, J., and Ng, S. (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70 (1), 191–221.

Basilevsky, A. (1994), *Statistical Factor Analysis and Related Methods: Theory and Applications*, New York: Wiley.

Bates, J. M., and Granger, C. W. J. (1969), "The Combination of Forecasts," *Operations Research Quarterly*, 20, 451–468.

Bult, J. R., and Wansbeek, T. (1995), "Optimal Selection for Direct Mail," *Marketing Science*, 14 (4), 378–394.

Chen, C.-H., and Li, K.-C. (1998), "Can SIR Be as Popular as Multiple Linear Regression?" *Statistica Sinica*, 8, 289–316.

Chiaromonte, F., Cook, R. D., and Li, B. (2002), "Sufficient Dimension Reduction in Regressions With Categorical Predictors," *The Annals of Statistics*, 30 (2), 475–497.

Cook, R. D. (1998), *Regression Graphics: Ideas for Studying Regressions Through Graphics*, New York: Wiley.

Cook, R. D., and Lee, H. (1999), "Dimension Reduction in Binary Response Models," *Journal of the American Statistical Association*, 94 (448), 1187–1200.

Cook, R. D., and Nachtsheim, C. J. (1994), "Reweighting to Achieve Elliptically Contoured Covariates in Regression," *Journal of the American Statistical Association*, 89 (426), 592–599.

Cook, R. D., and Weisberg, S. (1991), Discussion on "Sliced Inverse Regression for Dimension Reduction," by K.-C. Li, *Journal of the American Statistical Association*, 86, 328–332.

Cook, R. D., and Yin, X. (2001), "Dimension Reduction and Visualization in Discriminant Analysis" (with discussion), *Australian & New Zealand Journal of Statistics*, 43 (2), 147–199.

Cosslett, S. R. (1983), "Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model," *Econometrica*, 51, 765–782.

——— (1987), "Efficiency Bounds for Distribution-Free Maximum Likelihood Estimator of the Binary Choice and Censored Regression Models," *Econometrica*, 55, 559–585.

Day, G. S. (2000), "Capabilities for Forging Customer Relations," Working Paper 00-118, MSI, Cambridge, MA.

De Leeuw, J. (2005), "Nonlinear Principal Components Analysis and Related Techniques," eScholarship Repository, University of California, available at *http://repositories.cdlib.org/uclastat/papers/20005070801*.

Donkers, B., and Schafgans, M. (2005), "A Method of Moments Estimator for Semiparametric Index Models," Working Paper EM/05/493.

Fan, J. (1992), "Design-Adaptive Nonparametric Regression," *Journal of the American Statistical Association*, 87 (420), 998–1004.

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modeling and Its Applications*, Suffolk: Chapman & Hall.

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.

Gabler, S., Laisney, F., and Lechner, M. (1993), "Semi-Nonparametric Estimation of Binary Choice Models With an Application to Labor-Force Participation," *Journal of Business and Economic Statistics*, 11 (1), 61–80.

Gallant, A. R., and Nychka, D. N. (1987), "Semi-Nonparametric Maximum Likelihood Estimation," *Econometrica*, 57, 1091–1120.

Gifi, A. (1990), *Nonlinear Multivariate Analysis*, Chichester, U.K.: Wiley.

Hall, A. (1990), "Lagrange Multiplier Test for Normality Against Seminonparametric Alternatives," *Journal of Business and Economic Statistics*, 8, 417–426.

Hardle, W., and Stoker, T. M. (1989), "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association*, 84, 986–995.

Hastie, T. R., and Tibshirani, R. (1986), "Generalized Additive Models: Some Applications," *Journal of the American Statistical Association*, 82, 371–386.

——— (1987), "Generalized Additive Models," *Statistical Science*, 1, 197–318.

Horowitz, J. L. (1992), "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60, 505–531.

——— (1993), "Optimal Rates of Convergence of Parameter Estimators in the Binary Response Model With Weak Distributional Assumptions," *Econometric Theory*, 9, 1–18.

——— (1998), *Semiparametric Methods in Econometrics*, New York: Springer.

Hristache, M., Juditsky, A., Polzehl, J., and Spokoiny, V. (2001), "Structure Adaptive Approach for Dimension Reduction," *The Annals of Statistics*, 29 (6), 1537–1566.

Ichimura, H. (1993), "Semiparametric Least Squares (SLS) and Weighted SLD Estimation of Single-Index Models," *Journal of Econometrics*, 58, 71–120.

Ichimura, H., and Lee, L.-F. (1991), "Semiparametric Least Squares Estimation of Multiple Index Models: Single Equation Estimation," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, eds. W. A. Barnett, J. Powell, and G. Tauchen, Cambridge: Cambridge University Press, Chap. 1.

Kamakura, W., de Rosa, F., Wedel, M., and Mazzon, J. A. (2003), "Cross-Selling Financial Services With Database Marketing," *International Journal of Research in Marketing*, 20 (1), 45–65.

Kato, T. (1995), *Perturbation Theory for Linear Operators*, Berlin: Springer-Verlag.

Klein, R. W., and Spady, R. H. (1993), "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica*, 61 (2), 387–421.

Lemmens, A., and Croux, C. (2006), "Bagging and Boosting Classification Trees to Predict Churn," *Journal of Marketing Research*, 43 (2), 276–286.

Lewbel, A. (2000), "Semiparametric Qualitative Response Model Estimation With Unknown Heteroscedasticity and Instrumental Variables," *Journal of Econometrics*, 97, 145–177.

Li, B., Zha, H., and Chiaromonte, F. (2005), "Contour Regression: A General Approach to Dimension Reduction," *The Annals of Statistics*, 33 (4), 1580–1616.

Li, K.-C. (1991), "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*, 86, 316–342.

Li, K.-C., Aragon, Y., Shedden, K., and Agnan, C. T. (2003), "Dimension Reduction for Multivariate Response Data," *Journal of the American Statistical Association*, 98 (461), 99–109.

Li, Y., and Zhu, L. (2006), "Asymptotics for Sliced Average Variance Estimation," *The Annals of Statistics*, 35, 41–69.

Lin, W., and Kulasekara, K. B. (2007), "Idenitifiability of Single-Index Models and Additive-Index Models," *Biometrika*, 1–6.

Manski, C. F. (1975), "Maximum Score Estimator of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3, 205–228.

——— (1985), "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics*, 27, 313–333.

——— (1988), "Identification of Binary Response Models," *Journal of the American Statistical Association*, 83, 729–738.

Naik, P. A., and Tsai, C.-L. (2001), "Single-Index Model Selections," *Biometrika*, 88 (3), 821–832.

——— (2004), "Isotonic Single-Index Model for Database Marketing," *Computational Statistics and Data Analysis*, to appear.

——— (2005), "Constrained Inverse Regression for Incorporating Prior Information," *Journal of the American Statistical Association*, 100 (469), 204–211.

Neslin, S., Gupta, S., Kamakura, W., Lu, J., and Mason, C. (2006) "Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models," *Journal of Marketing Research*, 43 (2), 204–211.

Pagan, A., and Ullah, A. (1999), *Nonparametric Econometrics*, New York: Cambridge University Press.

Samarov, A. M. (1993), "Exploring Regression Structure Using Nonparametric Functional Estimation," *Journal of the American Statistical Association*, 88, 836–847.

Silverman, B. W. (1986), *Density Estimation*, London: Chapman & Hall.

Simonoff, J. S. (1996), *Smoothing Methods in Statistics*, New York: Springer.

Stock, J. H., and Watson, M. W. (2006), "Forecasting With Many Predictors," in *Handbook of Economic Forecasting*, Vol. 1, eds. G. Elliott, C. W. J. Granger, and A. Timmermann, Elsevier, Chap. 10.

Timmermann, A. (2006), "Forecast Combinations," in *Handbook of Economic Forecasting*, Vol. 1, eds. G. Elliott, C. W. J. Granger, and A. Timmermann, Elsevier, pp. 135–196.

Tyler, D. E. (1981), "Asymptotic Inference for Eigenvectors," *The Annals of Statistics*, 9 (4), 725–736.

Velilla, S. (1993), "A Note on Multivariate Box–Cox Transformations to Normality," *Statistics and Probability Letters*, 17, 315–322.

Winer, R. S. (2001), "A Framework for Customer Relationship Management," *California Management Review*, 43 (4), 89–105.

Xia, Y., Tong, H., Li, W. K., and Zhu, L.-X. (2002), "An Adaptive Estimation of Dimension Reduction Space," *Journal of the Royal Statistical Society*, Ser. B, 64 (3), 363–410.

Zhu, L., Miao, B., and Peng, H. (2006), "On Sliced Inverse Regression With High-Dimensional Covariates," *Journal of the American Statistical Association*, 101 (474), 630–643.

META DATA IN THE PDF FILE

Following information will be included as pdf file Document Properties:

Title : Multi-Index Binary Response Analysisof Large Data Sets
Author : Prasad A. Naik, Michel Wedel, Wagner Kamakura
Subject : Journal of Business \046 Economic Statistics, Vol.0, No.00, 0, 1-16
Keywords: Customer relationship marketing, Discrete choice, Factor model, Inverse regression, Semiparametric estimation, Sliced average variance estimation

THE LIST OF URI ADRESSES

Listed below are all uri addresses found in your paper. The non-active uri addresses, if any, are indicated as ERROR. Please check and update the list where necessary. The e-mail addresses are not checked – they are listed just for your information. More information can be found in the support page: *http://www.e-publications.org/ims/support/urihelp.html*.

```
--- mailto:panaik@ucdavis.edu [2:pp.1,1] Check skip
200 http://www.amstat.org [2:pp.1,1] OK
302 http://pubs.amstat.org/loi/jbes [2:pp.1,1] Found
200 http://dx.doi.org/10.1198/jbes.????.???? [2:pp.1,1] OK
404 http://repositories.cdlib.org/uclastat/papers/20005070801 [2:pp.14,14] Not Found
```