
**SCHOOL OF STATISTICS**
UNIVERSITY OF THE PHILIPPINES DILIMAN


# WORKING PAPER SERIES

## Semiparametric Probit Model for High-dimensional Clustered Data

by

**Daniel R. Raguindin**
**Joseph Ryan G. Lansangan**

UPSS Working Paper No. 2016-06
November 2016

School of Statistics
Ramon Magsaysay Avenue
U.P. Diliman, Quezon City
Telefax: 928-08-81
Email: updstat@yahoo.com

**Abstract**

In this paper, a semiparametric probit model for high dimensional clustered data and its estimation procedure are proposed. The proposed model is characterized by flexibility in the model structure through a nonparametric formulation of the effect of the predictors on the dichotomous response and a parametric specification of the inherent heterogeneity due to clustering. The predictive ability of the proposed model is further investigated by looking at possible factors such as dimensionality, presence of misspecification, clustering, and response distribution. Simulation studies illustrate the advantages of using the proposed model over the ordinary probit model even in low dimensional cases. High predictive ability is observed in high dimensional cases especially when the distribution of the response categories is balanced. Results show that cluster distribution and functional form of the response variable do not affect the performance of the proposed model. Also, the predictive ability of the proposed estimation increases as the number of clusters increases. Under the presence of misspecification, the predictive ability of the proposed model is little affected yet remains better than the ordinary probit model.

**Keywords:** *probit model, high dimensional data, backfitting algorithm, local scoring algorithm*

## 1. Introduction

Binary response variables are very common in fields of study such as in epidemiology, medicine, economics and social sciences (Agresti, 2002). A number of models may be considered when predicting a binary or dichotomous response variable with a set of quantitative and qualitative inputs. The use of generalized linear models, for one, has increased appreciably in the recent decades, and in the case of dichotomous response, logit and probit models have become the "de facto standards" (Zorn, 2005).

Logit and probit models have certain data and distributional assumptions. Aside from generalization of linearity, inputs must be nonstochastic (i.e., fixed under repeated samples) and the disturbances must be independent (Agresti, 2002). Such assumptions are in place to allow for consistency and/or efficiency of the model estimation. In many cases however, especially experiments in the natural sciences, observations are in fact conducted in several distinct batches, groups or clusters (Finney, 1952; Agresti and Gueorguieva, 2001). Samples may be taken from different points in time, from different groups of people, or from different geographical areas. Thus, several factors may implicitly govern the relationship of the dichotomous response and the set of inputs or predictor variables and may affect the model estimation. Cluster effect in the analysis may underestimate the true standard error of an estimated treatment or subject difference (Rao and Scott, 1992). Non-independence of the observations (disturbances) may make the standard logit/probit model problematic (Hedeker et al., 1994).

There may also exist correlations among the different predictor variables when clustering is present (Agresti and Gueorguieva, 2001). If two or more predictor variables are nearly perfectly correlated, then problems of computational imprecision, unstable estimates and large sampling error may occur. Hence, logit and probit models suffer the same problem of multicollinearity as in

ordinary least squares (Aldrich and Nelson, 1984). Accordingly, in the presence of a large number of predictor variables, the high dimensionality may pose problems of nonestimability (Dunteman, 1989; Aguilera et al., 2006; Aldrich et al., 1984).

In this study, a semiparametric probit model is postulated to address the issue on high dimensionality while accounting for the effect of data clustering. This model is characterized by flexibility in the model structure through a nonparametric formulation of the effect of the predictors on the dichotomous response and a parametric specification of the inherent (random) heterogeneity due to the clustering in the data. A dimension-reduction approach through principal components is induced to address high dimensionality.

In the following section, a brief discussion on probit analysis and use of principal components is presented. In Section 3, nonparametric regression and clustering considerations are given. The proposed model and its estimation are then presented in Section 4, and results of the simulation study are provided in Section 5. The last section provides concluding remarks.

## 2. Probit Model and Principal Components Approach

### *The Probit Model*

Probit or "probability unit" model is one of the common models used for dichotomous response. Suppose $Y$ is the response dichotomy $\{0,1\}$ variable and $\underline{X}$ is the set of inputs or predictor variables. The probability that $Y = 1$ is modeled as $P(Y = 1|\underline{X} = \underline{x}) = \Phi(\underline{xb}) = \Phi(z) = \pi(z)$ where $z$ is a single linear index, $\underline{b}$ a vector of coefficients, and $\Phi$ denote the cumulative probability function for the standard normal distribution. Here, the distribution is the normal distribution with

$$\pi(z) = \int_{-\infty}^{z} \frac{1}{\sigma\sqrt{2\pi}} exp\left[\frac{-1}{2}\left(\frac{s-\mu}{\sigma}\right)^2\right] ds = \Phi\left(\frac{z-\mu}{\sigma}\right)$$

For some real number $\mu$ and positive $\sigma$. Clearly, $\pi(z)$ has the form $\pi(z) = \Phi(\alpha + \beta z)$ with $\alpha = \frac{-\mu}{\sigma}$ and $\beta = \frac{1}{\sigma}$. So that $\Phi^{-1}[\pi(z)] = \alpha + \beta z$, or equivalently, reverting back to the original inputs $\underline{x}$, the stochastic model $y^* = \alpha + \underline{x}\underline{\beta}^* + \epsilon$ with $E(\epsilon) = 0$ is the linear probability or probit model (Agresti, 2002). The probit model is derived under the assumption that the predictor variables are exogenous (random or fixed, but independent) and the error term $\epsilon$ are normal and homoskedastic (Agresti, 2002; Train, 2002).

Although the logit and probit models are entirely different, a probit model may be used to approximate a logit model. Such is possible since the probit model formulation eliminates the integration problem in a logit model (Demidenko, 2013). In model selection, Chen and Tsurumi (2011) identified different measures to characterize and optimally choose between a probit model and a logit model. They noted that if the data is balanced (i.e., response categories are balanced), either model can perform well. But when data are unbalanced, a logit model is better when the data are generated by a leptokurtic distribution while probit model is preferred when the data are generated by a platykurtic distribution (Chen and Tsurumi, 2011).

### *Using Principal Components*

Principal component analysis (PCA) is a statistical technique that linearly transforms an original set of predictor variables into a substantially smaller set of uncorrelated variables (called principal components, or PCs) that represents most of the information in the original set of variables. The main goal of this technique is to reduce the dimensionality of the original data set while capturing the most information (Jolliffe, 2002; Dunteman, 1989). In a regression problem,

if multicollinearity exists, then PCA may help in the estimation of regression parameters by using the derived PCs as regressors. The approach of using PCs in regression is called principal component regression (PCR). Dunteman (1989) and Jolliffe (2002) illustrated that with an appropriate selection of a subset of PCs to be used as predictors in a desired regression model, PCR can still capture most of the variability in the data.

Aguilera et al. (2006) found that using the logistic regression in the framework of high dimensional binary response may lead to erroneous results. To improve the estimation under multicollinearity and to reduce the dimension, they proposed to use as covariates the set of optimum PCs of the original predictors. However, when using PCA especially for high dimensional data, identification of significant predictors and interpretability of the PCs may become difficult (Ma, 2013). One common solution is to induce sparsity of component loadings in the derivation of the optimal components.

Zou et al. (2006) introduced a new method called sparse principal component analysis (SPCA) using an elastic net approach to generate a modified set of principal components with sparse loadings. They first showed that PCA can be formulated as a regression-type optimization problem, and then they imposed a LASSO-type constrained (which they called elastic net) on the regression coefficients. The resulting sparse principal components (SPCs), unlike the ordinary PCs, are correlated. Also, at the expense of interpretability, the first k SPCs capture fewer variability than those of the first k PCs.

## 3. Nonparametric Regression and Clustered Data

### *Nonparametric Regression*

A nonparametric regression model generally assumes that the regression curve belongs to some infinite dimensional collection of functions. Thus, this method allows great flexibility in the possible form of the regression curve and it makes no assumptions about a parametric form (Eubank, 1999). The local scoring algorithm is used to estimate the functions $f_i(x_i)$ nonparametrically via smoothing the partial residuals derived using backfitting algorithm (Hastie and Tibshirani, 1986; Eubank, 1999). Hastie and Tibshirani (1986; 1987) applied the local scoring to probit and logit models.

In the case of high dimensional data, Umali and Barrios (2014) proposed nonparametric principal component regression to address not only the multicollinearity problem but also to minimize the specification bias brought by dimension reduction. They noted that the loss in predictive ability of the components due to dimension reduction may be minimized by relaxing the functional form of the predictor variables on the dependent variable in a nonparametric context. A two-step procedure was used to model a (continuous) response variable, first via finding principal components (PCs) for the inputs, then fitting the best smooth function for the response using the PCs.

### *Clustered Data*

The problem of multicollinearity is often associated in high dimensional data. As defined in the literature, multicollinearity problem exists when there are explanatory variables in a regression model that are highly correlated to each other (Belsley et al., 1980; Greene, 2000). This can also happen in clustered data due to interdependencies within cluster.

Complicated problems that may lead to a more complex modelling strategy may originate from the dependence structure between observations from same clusters. Graubard and Korn (1994) suggested some techniques to address this issue. One approach is to model the expectation of the

outcome as a function of the covariates and a cluster-specific term which can be fixed or random effect. Another is to model the expectation of the outcome of a randomly chosen individual from the population as a function of the covariates without regard to cluster membership and then to take into account the correlation through robust variance estimation of the regression coefficients, such is commonly referred to as population average modelling (Graubard and Korn, 1994).

Hedeker et al. (1994) described a random-effect regression model for analyzing clustered data that does not assume that each observation is independent but does assume data within clusters are dependent to some degree. Furthermore, Kang et al. (2005) compared a random effects model estimator called the hierarchical generalized linear model (HGLM) and the conditional likelihood estimator in the analysis of clustered binary data. As they pointed out, the HGLM procedure exploits information by using the distributional assumption about random effects. However, in binary response data, checking distributional assumptions can be difficult.

Demindenko (2013) proposed a probit model for clustered data. It is in the form of

$$\mathrm{P}\left(y_{ij}=1\middle|u_i\right)=\Phi\left(u_i+\beta' x_{ij}\right)$$

where $u_i$ is the cluster random intercept which are independent and normally distributed random variables with zero mean and unknown variance $\sigma^2$. As he noted, if the number of clusters is small but the sample size are relatively large it can be assumed that the $u_i$'s are fixed and unknown. However, if the number of clusters is large but the sample size is relatively small, the model with the random intercept is more appropriate.

## 4. The Proposed Model and Estimation Procedure

### *The Proposed Model*

It is proposed to transform the traditional probit model into an additive combination of parametric and nonparametric components. Let $Y_i^k$ be the categorical response variable that takes only two possible alternatives (e.g. 1 or 0), and let *p* variables be measured at *n* observations where $n > p$ or possibly $n \leq p$. Let

$$X^k=\begin{bmatrix} x_{11}^k & x_{12}^k\cdots & x_{1p}^k \\ x_{21}^k & x_{22}^k\cdots & x_{2p}^k \\ \vdots & \vdots\ \ddots & \vdots \\ x_{n1}^k & x_{n2}^k\cdots & x_{np}^k \end{bmatrix}$$

where $\underline{x}_i^k=(x_{i1}^k, x_{i2}^k, \cdots, x_{in}^k)^T$, $i = 1, 2, \ldots, p$, be the measurements in the *p* variables for the i$^{\text{th}}$ observation belonging to the k$^{\text{th}}$ cluster.

Suppose we want to model $Y_i^k$. Let $\pi\left(\underline{x}_i^k\right) = P\left(Y_i^k = 1\middle|\, \underline{X}_i = \underline{x}_i^k\right)$. Then, the proposed semiparametric model is given by

$$\Phi^{-1}\left[\pi\left(\underline{x}_i^k\right)\right]=\delta_k+f\{g_m\left(\underline{X}_i\right)\}+\varepsilon_i \quad \text{or} \quad \pi\left(\underline{x}_i^k\right)=\Phi\left(\delta_k+f\{g_m\left(\underline{X}_i\right)\}+\varepsilon_i\right) \qquad (1)$$

where $\pi\left(\underline{x}_i^k\right)$ is the probability of choosing $Y_i^k = 1$ for the i$^{\text{th}}$ observation belonging in the k$^{\text{th}}$ cluster, $\delta_k$ is a cluster-specific random intercept, $g_m(\cdot)$ is the m$^{\text{th}}$ component evaluated by either PCA or SPCA with $m = 1, 2, \ldots, q$ and $q << p$, $f(\cdot)$ is a smooth function of $Z$, and $\varepsilon_i$ is the error term.

From (1), the components $g_m$ are used instead of the original set of predictor variables to alleviate high-dimensionality or multicollinearity issues. Also, the nonparametric component $f(\cdot)$ is postulated to relax the assumption pertaining to the data and to capture the potential bias based on the information lost due to dimension reduction. The random component $\delta_k$ is the estimated k$^{\text{th}}$ cluster-specific random intercept that captures the effect of clustering of the data.

A three-step procedure is proposed to estimate the model. First is to choose the optimal components to be included in the model through SPCA (or PCA especially for non-high-dimensional cases). The second step is a rescaling of the dependent variable through a modified local scoring algorithm. The third step fits the transformed dependent variable with the best smooth function of the components and the cluster-specific random intercept through a backfitting algorithm. Details of which are presented in the following section.

***Modified Local Scoring***

Abe (1999) proposed a local scoring algorithm based on the original algorithm proposed by Hastie and Tibshirani (1990) for matched case-control data. This method incorporates a nonparametric additive utility function that generalizes the logistic regression of the general additive model (GAM) for a binary response to a qualitative variable that can assume more than two values. This algorithm is modified in the case of clustered data. The iterative procedure is described as follows:

Step 1: Obtain the initial estimates of the log odds $\eta_i$ for the i$^{\text{th}}$ observation by the linear model, using the components $C_m$ derived via SPCA or PCA.

$$\eta_i = log\left(\frac{P(Y=1)}{P(Y=0)}\right) = \beta_0 + \beta_1 C_1 + \beta_2 C_2 + \cdots + \beta_q C_q + \varepsilon$$

Step 2: Estimate the initial probability of the first category on the i$^{\text{th}}$ observation $\hat{\mu}_i$ by

$$\hat{\mu}_i = P_i(Y=1) = \frac{\text{number of observations take on first category}}{n}$$

and solve for the log-likelihood based on the initial values as

$$log - likelihood = \sum_{i=1}^{n} I(Y_i) P_i(Y=1)$$

Step 3: Define and compute the adjusted logit function $z_i$ as a function of some component $\eta_i$. This will serve as the dependent variable for the backfitting algorithm. Compute for the weight $w_i$.

$$z_i = \eta_i + \frac{Y_i - \hat{\mu}_i}{\hat{\mu}_i(1-\hat{\mu}_i)} \quad (2)$$

$$w_i = \hat{\mu}_i(1-\hat{\mu}_i) \quad (3)$$

Step 4: Use backfitting algorithm to obtain the smooth function of each of the component $C_m$. Apply the results of the backfitting procedure to update the value of $\eta_i$. Now we have

$$z_i = \eta_i + \epsilon_k \quad (4)$$

where $\epsilon_k$ is the partial residual for the k$^{\text{th}}$ cluster and

$$\eta_i = \sum_{m=1}^{q} f(C_m) \quad (5)$$

where $f(C_m)$ is the smooth function of the m$^{\text{th}}$ component. Update the value of $\eta_i$ in (2) using (5) to update response variable $z_i$ in Step 3. Then, repeat this step until the difference between $\eta_i$ in two consecutive iterations is sufficiently small ($\leq$0.001).

Step 5: After convergence, $\epsilon_k$ in (4) still contains the information about the cluster-specific intercept, $\hat{\delta}_k$, say $\epsilon_k = \hat{\delta}_k + \varepsilon$, it is then used to estimate $\hat{\delta}_k$ from a random effect formulation. The $z_i$ in (4) may then be viewed in the form

$$z_i = \hat{\delta}_k + \eta_i + \varepsilon_i \quad (6)$$

Step 6: Update the estimated probability that the response, *Y*, will take the first category using estimated $\hat{\delta}_k$ and the updated estimates of $\eta_i$ from (5). Update the log-likelihood value in Step 2 and update $z_i$ in (2) using the latest values of $\hat{\mu}_i$,

$$\hat{\mu}_i = \frac{1}{1 + e^{-(\eta_i + \hat{\delta}_k)}} \tag{7}$$

Step 7: Repeat steps 2 through 6 until the log-likelihood converges. That is, the difference in log-likelihood between two consecutive iterations is $\leq 0.001$.

***Backfitting the semiparametric model***

The parametric and nonparametric parts of the model are estimated separately in this procedure. The backfitting algorithm described by Hastie and Tibshirani (1990) is used in the estimation of the smooth function $f(C_m)$ and the cluster-specific random intercept, $\hat{\delta}_k$. The following iterative procedure is used in the estimation:

Step 1: Let the final value of $z_i$ estimated in Step 5 from the modified local scoring procedure be the dependent variable. Using smoothing splines, $f(PC_m)$ is estimated nonparametrically ignoring initially the contribution of the cluster effects using the original *m* PCs. The partial residual is then computed as

$$\epsilon_k = z_i - \sum_{m=1}^{q} f(PC_m) \tag{8}$$

Step 2: The partial residual, $\epsilon_k$, evaluated in (8) still contains information on the cluster effect. So that $\epsilon_k$ is based on previously identified cluster and is used to estimate the cluster-specific random intercept, $\hat{\delta}_k$ from a random effect formulation, say $\epsilon_k = \hat{\delta}_k + \varepsilon$. Thus, the model becomes

$$z_i = \hat{\delta}_k + \sum_{m=1}^{q} f(C_m) + \varepsilon_i \tag{9}$$

Step 3: Repeat Steps 1 through 2 until the difference between two consecutive iterations in $f(C_m)$ is sufficiently small ($\leq 0.001$). Hence, the probability that the $i^{th}$ respondent selects the first category follows the standard normal cumulative distribution function

$$P_i(Y = 1) = \Phi(z_i) = \int_{-\infty}^{z_i} \frac{1}{\sqrt{2\pi}} \; e^{\frac{-t^2}{2}} dt \tag{10}$$

Step 4: Finally, the a-priori knowledge on the probability that $Y = 1$ which is $\hat{\mu}_i$ in Step 2 of modified local scoring may serve as the cut-off value in transforming $P_i(Y = 1)$ in (10) into a binary response (1 or 0). If $P_i(Y = 1) \geq \hat{\mu}_i$, then $Y = 1$, and otherwise 0.

## 5. Simulation Study and Results

The proposed method is evaluated through a simulation study under various scenarios. Several factors were considered: dimensionality of the data, functional form of the data-generating model, presence of misspecification in the model, number of clusters, distribution of the clusters, and the distribution of the observations in the response category. Summaries of the percentage of correctly predicted responses for each category and across all categories are computed to assess the proposed estimation procedure. For comparison, standard probit modelling is performed for low dimensional data. The simulation settings are summarized in the following table.

**Table 1. Simulation Parameter Settings**

| Factor | Setting |
|---|---|
| Dimensionality | Number of input variables $p = 5,\ 200$<br>Number of observations $n = 100, 200$ |
| Response category | Balanced: 1:1 1-0 allocation<br>Moderately unbalance: 7:3<br>Severely unbalance: 9:1 |
| Input data | Low dimensional case:<br>$X_1, X_2, X_3, X_4, X_5$ are generated as $X_1 \sim U(a, b)$; $X_2 \sim N(\mu, \sigma^2)$; $X_3 = c * X_1 + d + \varepsilon_1$; $X_4 = e * X_2 + f + \varepsilon_2$; and $X_5 = g * X_1 + h * X_2 + \varepsilon_3$, where $c, d, e, f, g,$ and $h$ are fixed values and $\varepsilon_i \sim N(0, \sigma_i^2), i = 1, 2, 3.$ |
| | High dimensional case:<br>$X_1, X_2, \dots, X_{200}$ are generated as $X_1 \sim U(a, b)$; $X_{76} \sim N(\mu, \sigma^2)$; $X_i = c * X_1 + \varepsilon_i$ where *c* is a fixed value and $\varepsilon_i \sim N(0, \sigma_i^2)$, $i = 2, 3, \dots, 75$; $X_i = d * X_{41} + \varepsilon_i$ where *d* is a fixed value and $\varepsilon_i \sim N(0, \sigma_i^2)$, $i = 77, 78, \dots, 150$; and $X_i = e * X_1 + f * X_2 + \varepsilon_i$ where *e*, *f* area fixed values and $\varepsilon_i \sim N(0, \sigma_i^2)$, $i = 151,152, \dots, 200.$ |
| Functional form | Linear: $Y_i^{*k} = \delta_k + \beta_1 X_{i1}^k + \cdots + \beta_5 X_{i5}^k + w\varepsilon_i$<br>Nonlinear: $Y_i^k = \delta_k + \exp(\beta_1 X_{i1}^k) + \cdots + \exp(\beta_5 X_{i5}^k) + w\varepsilon_i$<br>(also apply for $p = 200$)<br>Lower values of $Y_i^{*k}$ corresponds to $Y = 0$,<br>and higher values corresponds to $Y = 1$;<br>with actual observations selected randomly |
| Misspecification | $w = 1, 10$ |
| Number of clusters | Number of clusters $k = 5, 10$ |
| Cluster distribution | Equal: each cluster with $n/k$ observations<br>Unequal: 16:15:8:6:5 allocation for 5 clusters;<br>allocation split to two for 10 clusters |

For the tables and discussions, the following notations are used: *Semiparametric Choice Model 1* or *Model 1* – the proposed method using ordinary PCA in reducing the dimension of the data; *Semiparametric Choice Model 2* or *Model 2* – the proposed method using SPCA in reducing the dimension of the data. Furthermore, percentage of perfect prediction rate is used to determine the level of separation problem for each scenario. This study considered those having more than 10% of the replicates with perfect prediction rate as a case on having problems in separation. Problems on separation of data are common for binary responses. For 2 × 2 table formed by a binary response, *complete*[1] separation occurs when only the two opposing diagonal cells contain data.

[1] Y can be perfectly predicted by X's across all observations. This leads to two results, a) there is no variability left to be explained by the other covariates in the model, so the corresponding parameter estimates for the remaining covariates is zero and b) likelihood is flat yielding infinite standard error estimates.

*Quasi complete*[2] separation, on the other hand, occurs when only one cell in the 2 × 2 table is "empty" (Zorn, 2005).

***Effects of Cluster Distribution and of Number of Clusters***

For $n > p$, when the number of observations among clusters is equal, the average predictive ability of Model 2 is higher than the traditional probit model (see Table 1 in Appendix). However, both the traditional probit model and Model 2 suffer from the problem of separation if the distribution of the response category becomes unbalanced. In this case, Model 1 has the advantage over the traditional model and Model 2. Similar results are observed when the number of observations among clusters is unequal (see Table 2 in Appendix). This implies that the cluster distribution for low dimensional cases does not affect the model estimation.

For high dimensional cases ($n \leq p$) with equal number of observations per cluster, the proposed Model 2 works well if the distribution of the response variable is balanced (see Table 3 through Table 6 in Appendix). Model 2 is at par with Model 1 when the distribution of the categories of the response variable in the data becomes unbalanced, but as it seems, Model 2 may suffer from the problem of separation.

Regardless whether the number of observations per cluster is equal or unequal, as the distribution of the response categories become more unbalanced, prediction ability is higher in low dimensional cases compared to high dimensional cases. In high dimensional cases, the predictive ability of the proposed estimation increases as the number of cluster increases. Also, both for low dimensional and high dimensional data, Model 2 provides the highest prediction ability when the distribution of the response categories is balanced regardless on the number of clusters. Model 1, however, provides good prediction ability when the distribution of the response category becomes unbalanced.

***Effects of the Functional Form and Misspecification in the Model***

If the dependent variable is generated from a linear combination of independent variables, Model 2 seems at par with Model 1 even under the severely unbalanced response categories. Model 2 in the low dimensional cases provides the highest prediction rate even if the dependent variable is generated nonlinearly (see Table 7 in Appendix). Under the high dimensional case, Model 2 performs well even when the dependent variable is generated nonlinearly and in the case where the distribution of response category is balanced (see Table 8 in Appendix).

Recall that misspecification in the model is characterized by multiplying a constant in the error terms, leading to residuals that contain large variances. Consequently, larger part on the variation on the dependent variable may not necessarily be explained by the covariates in the model. The proposed estimation (either Model 1 or Model 2) is generally advantageous over the traditional probit model when misspecification is present in the model under the low dimensional case and a balanced distribution in the response categories (see Table 9 in Appendix). However, as the number of clusters increases, the performance of the proposed estimation procedure (especially Model 2) slightly decreases (yet remains better than the ordinary probit model). In high dimensional cases, Model 2 still provides high prediction rate under the balanced distribution of the response categories but generally slightly deteriorates as the number of cluster increases (see Table 10 in Appendix).

---

[2] The parameter estimates for the separating variable X's and its standard errors will also be infinite, but the other covariates may remain relatively unaffected.

## 6. Conclusions

A semiparametric probit model is proposed for high dimensional clustered data. The nonparametric approach of using smooth functions of the components and the parametric approach of incorporating the cluster effects to model the response variable capture the relationship the response variable with respect to the original predictor variables.

The proposed semiparametric model (either with PCA or SPCA) generally provides better results compared to a traditional probit model even under the low dimensional cases (i.e., when the number of observations is much greater than the number of predictor variables). It provides relatively high predictive ability especially under a balanced distribution of the response categories. The cluster distribution and functional form of the response variable (relative to the original input variables) in general do not affect the performance of the estimated model. The proposed model is little affected by the presence of misspecification, yet over-all it remains better compared to the ordinary probit model under the low dimensional case.

## REFERENCES


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

AGRESTI, A., 2002, *Categorical Data Analysis*, 2nd Edition, Hoboken, New Jersey: John Wiley & Sons, Inc.

AGRESTI, A. and R. GUEORGUIEVA, 2001, A Correlated Probit Model for Joint Modelling for Clustered Binary Data and Continuous Responses, *Journal of the American Statistical Association*, 96(455): 1102-1112.

AGUILERA, A., M. ESCABIAS, and M. VALDERRAMA, 2006, Using Principal Components for Estimating Logistic Regression with High-Dimensional Multicollinear Data, *Computational Statistics and Data Analysis*, 50: 1905-1924.

ALDRICH, J.H. and F. NELSON, 1984, *Linear Probability, Logit, and Probit Models*, Sara Miller McCune, Sage Publications, Inc.

BELSLEY, D.A., E. KUH, and R.E. WELSCH, 1980, Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: John Wiley and Sons

CHEN, G. and H. TSURUMI, 2011, Probit and Logit Model Selection, *Communication in Statistics – Theory and Methods*, 40: 159-175.

DEMIDENKO, E., 2013, *Mixed Models: Theory and Applications with R*, 2nd Edition, Hoboken, New Jersey: John Wiley & Sons, Inc.

DUNTEMAN, G., 1989, *Principal Component Analysis*, Sage University Paper Series on Quantitative Applications on the Social Sciences, Series No. 07-069, Newbury Park, California, SAGE Publication, Inc.

EUBANK, R., 1999, *Nonparametric Regression and Spline Smoothing*, 2nd Edition, New York: Marcel Dekker, Inc.

FINNEY, D. J., 1952, *Probit Analysis*, 2nd Edition, Cambridge: Cambridge University Press.

GRAUBARD, B. and E. KORN, 1994, Regression Analysis with Clustered Data, *Statistics in Medicine*, 13: 509-522.

GREENE, W.H., 2000, Econometric Analysis (Fourth edition), Upper Saddle River, NJ: PrenticeHall.

HASTIE, T. and R. TIBSHIRANI, 1986, Generalized Additive Models, *Statistical Science*, 1(3):297-318.

HASTIE, T. and R. TIBSHIRANI, 1987, Generalized Additive Models: Some Applications, *Journal of the American Statistical Association*, 82(398): 371-386.

HASTIE, T. and R. TIBSHIRANI, 1990, *Generalized Additive Models,* New York: Chapman and Hall.

HEDEKER, D., S. McMAHON, L. JASON, and D. SALINA, 1994, Analysis of Clustered Data in Community Psychology: With an Example from a Worksite Smoking Cessation Project, *American Journal of Community Psychology*, 22(5): 595-615.

JOLLIFFE, I., 2002, *Principal Component Analysis*, 2nd Edition, New York: Springer – Verlag.

KANG, W., M.S. LEE, and Y. LEE, 2005, HGLM versus Conditional Estimators for the Analysis of Clustered Binary Data, *Statistics in Medicine*, 24: 741-752.

MA, Z., 2013, Sparse Principal Component Analysis and Iterative Thresholding, *The Annals of Statistics*, 41(2): 772-801.

RAO, J. N. and A. SCOTT, 1992, A Simple Method for the Analysis of Clustered Binary Data, *Biometrics*, 48 (2): 577-585.

TRAIN, K., 2002, *Discrete Choice Methods with Simulations*. Cambridge University Press.

UMALI, J. and E. BARRIOS, 2014, Nonparametric Principal Components Regression, *Communications in Statistics – Simulation and Computation*, 43: 1797-1810.

ZORN, C., 2005, A Solution to Separation in Binary Response Models, *Political Analysis*, 13: 157-170.

ZOU, H., T. HASTIE, and R. TIBSHIRANI, 2006, Sparse Principal Component Analysis, *Journal of Computational and Graphical Statistics*, 15(2): 256-286.

**APPENDIX: Summary of Simulation Study**

**Table 1.** Comparison of predictive ability for low dimensional case ($n > p$) with equal cluster size
(by the distribution of the dependent variable)

| | Ordinary Probit Regression | | | Semiparametric Choice Model 1 | | | Semiparametric Choice Model 2 | | |
|---|---|---|---|---|---|---|---|---|---|
| For *k = 5* | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall |
| ***Average*** | | | | | | | | | |
| The distribution of Y is balance | 60.16 | 60.14 | 60.15 | 55.68 | 56.44 | 56.06 | 93.50 | 92.42 | 92.96 |
| The distribution of Y is moderately unbalance | 99.97 | 0.30 | 70.19 | 63.93 | 52.90 | 60.62 | 89.86 | 97.60 | 92.18 |
| The distribution of Y is severely unbalance | 100.00 | 0.10 | 90.01 | 77.00 | 61.40 | 75.44 | 86.66 | 99.90 | 87.98 |
| ***Percentage of 100% prediction rate*** | | | | | | | | | |
| The distribution of Y is balance | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 4.00 | 1.00 | 0.00 |
| The distribution of Y is moderately unbalance | 98.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 50.00 | 0.00 |
| The distribution of Y is severely unbalance | 100.00 | 0.00 | 0.00 | 0.00 | 3.00 | 0.00 | 0.00 | 99.00 | 0.00 |
| | Ordinary Probit Regression | | | Semiparametric Choice Model 1 | | | Semiparametric Choice Model 2 | | |
| For *k = 10* | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall |
| ***Average*** | | | | | | | | | |
| The distribution of Y is balance | 59.76 | 60.02 | 59.89 | 56.40 | 57.46 | 56.93 | 92.76 | 92.90 | 92.83 |
| The distribution of Y is moderately unbalance | 99.93 | 0.77 | 70.18 | 64.43 | 52.60 | 60.88 | 88.76 | 96.30 | 91.02 |
| The distribution of Y is severely unbalance | 100.00 | 0.00 | 90.00 | 75.90 | 61.60 | 74.47 | 85.66 | 99.80 | 87.07 |
| ***Percentage of 100% prediction rate*** | | | | | | | | | |
| The distribution of Y is balance | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 2.00 | 1.00 | 0.00 |
| The distribution of Y is moderately unbalance | 95.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 34.00 | 0.00 |
| The distribution of Y is severely unbalance | 100.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 98.00 | 0.00 |

**Table 2.** Comparison of predictive ability for low dimensional case ($n > p$) with unequal cluster size
(by the distribution of the dependent variable)

| | Ordinary Probit Regression | | | Semiparametric Choice Model 1 | | | Semiparametric Choice Model 2 | | |
|---|---|---|---|---|---|---|---|---|---|
| For *k = 5* | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall |
| ***Average*** | | | | | | | | | |
| The distribution of Y is balance | 59.82 | 61.04 | 60.43 | 56.66 | 56.50 | 56.58 | 93.54 | 93.22 | 93.38 |
| The distribution of Y is moderately unbalance | 99.75 | 0.94 | 69.12 | 64.03 | 53.39 | 60.73 | 89.86 | 97.23 | 92.14 |
| The distribution of Y is severely unbalance | 100.00 | 0.00 | 90.00 | 76.07 | 61.10 | 74.57 | 86.81 | 100.00 | 88.13 |
| ***Percentage of 100% prediction rate*** | | | | | | | | | |
| The distribution of Y is balance | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 7.00 | 4.00 | 0.00 |
| The distribution of Y is moderately unbalance | 85.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 43.00 | 0.00 |
| The distribution of Y is severely unbalance | 100.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 100.00 | 0.00 |

| | Ordinary Probit Regression | | | Semiparametric Choice Model 1 | | | Semiparametric Choice Model 2 | | |
|---|---|---|---|---|---|---|---|---|---|
| For *k = 10* | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall |
| ***Average*** | | | | | | | | | |
| The distribution of Y is balance | 59.84 | 60.02 | 59.93 | 57.20 | 58.34 | 57.77 | 92.64 | 93.10 | 92.87 |
| The distribution of Y is moderately unbalance | 99.90 | 0.83 | 70.18 | 64.64 | 54.20 | 61.51 | 89.59 | 97.13 | 91.85 |
| The distribution of Y is severely unbalance | 100.00 | 0.00 | 90.00 | 84.63 | 60.17 | 73.23 | 87.10 | 99.92 | 88.64 |
| ***Percentage of 100% prediction rate*** | | | | | | | | | |
| The distribution of Y is balance | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 3.00 | 3.00 | 0.00 |
| The distribution of Y is moderately unbalance | 93.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 44.00 | 0.00 |
| The distribution of Y is severely unbalance | 100.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 99.00 | 0.00 |

**Table 3.** Comparison of predictive ability for high dimensional case ($n < p$) with equal cluster size
(by the distribution of the dependent variable)

| | Semiparametric Choice Model 1 | | | Semiparametric Choice Model 2 | | |
|---|---|---|---|---|---|---|
| For *k = 5* | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall |
| ***Average*** | | | | | | |
| The distribution of Y is balance | 53.44 | 53.84 | 53.64 | 94.06 | 93.76 | 93.91 |
| The distribution of Y is moderately unbalance | 54.60 | 54.87 | 54.68 | 90.34 | 97.47 | 92.48 |
| The distribution of Y is severely unbalance | 56.10 | 58.80 | 56.37 | 86.89 | 99.80 | 88.18 |
| ***Percentage of 100% prediction rate*** | | | | | | |
| The distribution of Y is balance | 0.00 | 0.00 | 0.00 | 5.00 | 3.00 | 0.00 |
| The distribution of Y is moderately unbalance | 0.00 | 0.00 | 0.00 | 1.00 | 53.00 | 1.00 |
| The distribution of Y is severely unbalance | 0.00 | 0.00 | 0.00 | 0.00 | 98.00 | 0.00 |
| | Semiparametric Choice Model 1 | | | Semiparametric Choice Model 2 | | |
| For *k = 10* | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall |
| ***Average*** | | | | | | |
| The distribution of Y is balance | 54.72 | 55.26 | 54.99 | 93.98 | 94.42 | 94.20 |
| The distribution of Y is moderately unbalance | 54.63 | 55.80 | 54.98 | 90.67 | 97.83 | 92.82 |
| The distribution of Y is severely unbalance | 55.86 | 58.20 | 56.09 | 87.32 | 99.70 | 88.56 |
| ***Percentage of 100% prediction rate*** | | | | | | |
| The distribution of Y is balance | 0.00 | 0.00 | 0.00 | 6.00 | 6.00 | 0.00 |
| The distribution of Y is moderately unbalance | 0.00 | 0.00 | 0.00 | 0.00 | 55.00 | 0.00 |
| The distribution of Y is severely unbalance | 0.00 | 1.00 | 0.00 | 0.00 | 97.00 | 0.00 |

**Table 4.** Comparison of predictive ability for high dimensional case ($n < p$) with unequal cluster size
(by the distribution of the dependent variable)

| | Semiparametric Choice Model 1 | | | Semiparametric Choice Model 2 | | |
|---|---|---|---|---|---|---|
| For *k = 5* | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall |
| ***Average*** | | | | | | |
| The distribution of Y is balance | 54.30 | 53.86 | 54.08 | 93.74 | 92.88 | 93.31 |
| The distribution of Y is moderately unbalance | 54.26 | 53.55 | 54.04 | 91.17 | 97.26 | 93.06 |
| The distribution of Y is severely unbalance | 56.20 | 58.70 | 56.45 | 86.49 | 99.80 | 87.82 |
| ***Percentage of 100% prediction rate*** | | | | | | |
| The distribution of Y is balance | 0.00 | 0.00 | 0.00 | 3.00 | 3.00 | 0.00 |
| The distribution of Y is moderately unbalance | 0.00 | 0.00 | 0.00 | 1.00 | 45.00 | 0.00 |
| The distribution of Y is severely unbalance | 0.00 | 0.00 | 0.00 | 0.00 | 99.00 | 0.00 |
| | Semiparametric Choice Model 1 | | | Semiparametric Choice Model 2 | | |
| For *k = 10* | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall |
| ***Average*** | | | | | | |
| The distribution of Y is balance | 53.70 | 54.54 | 54.12 | 94.28 | 93.58 | 93.93 |
| The distribution of Y is moderately unbalance | 56.17 | 57.00 | 56.42 | 90.69 | 97.73 | 92.80 |
| The distribution of Y is severely unbalance | 56.81 | 62.00 | 57.43 | 88.03 | 99.83 | 89.45 |
| ***Percentage of 100% prediction rate*** | | | | | | |
| The distribution of Y is balance | 0.00 | 0.00 | 0.00 | 8.00 | 1.00 | 0.00 |
| The distribution of Y is moderately unbalance | 0.00 | 0.00 | 0.00 | 1.00 | 54.00 | 0.00 |
| The distribution of Y is severely unbalance | 0.00 | 1.00 | 0.00 | 0.00 | 98.00 | 0.00 |

**Table 5.** Comparison of predictive ability for high dimensional case ($n < p$) with equal cluster size (by the distribution of the dependent variable)

| | Semi - parametric Choice Model 1 | | | Semi - parametric Choice Model 2 | | |
|---|---|---|---|---|---|---|
| For *k = 5* | Category 1 | Category 2 | Overall | Category 1 | Category 2 | Overall |
| ***Average*** | | | | | | |
| The distribution of Y is balance | 54.01 | 53.48 | 53.75 | 91.15 | 91.57 | 91.36 |
| The distribution of Y is moderately unbalance | 55.14 | 55.15 | 55.14 | 88.69 | 97.10 | 91.22 |
| The distribution of Y is severely unbalance | 55.76 | 57.50 | 55.94 | 85.84 | 99.95 | 87.26 |
| ***Percentage of 100% prediction rate*** | | | | | | |
| The distribution of Y is balance | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| The distribution of Y is moderately unbalance | 0.00 | 0.00 | 0.00 | 0.00 | 16.00 | 0.00 |
| The distribution of Y is severely unbalance | 0.00 | 0.00 | 0.00 | 0.00 | 99.00 | 0.00 |
| | Semi - parametric Choice Model 1 | | | Semi - parametric Choice Model 2 | | |
| For *k = 10* | Category 1 | Category 2 | Overall | Category 1 | Category 2 | Overall |
| ***Average*** | | | | | | |
| The distribution of Y is balance | 54.72 | 55.26 | 54.99 | 93.98 | 94.42 | 94.20 |
| The distribution of Y is moderately unbalance | 54.63 | 55.80 | 54.98 | 90.67 | 97.83 | 92.82 |
| The distribution of Y is severely unbalance | 55.86 | 58.20 | 56.09 | 87.32 | 99.70 | 88.56 |
| ***Percentage of 100% prediction rate*** | | | | | | |
| The distribution of Y is balance | 0.00 | 0.00 | 0.00 | 6.00 | 6.00 | 0.00 |
| The distribution of Y is moderately unbalance | 0.00 | 0.00 | 0.00 | 0.00 | 55.00 | 0.00 |
| The distribution of Y is severely unbalance | 0.00 | 1.00 | 0.00 | 0.00 | 97.00 | 0.00 |

**Table 6.** Comparison of predictive ability for high dimensional case ($n < p$) with unequal cluster size
(by the distribution of the dependent variable)

| | Semiparametric Choice Model 1 | | | Semiparametric Choice Model 2 | | |
|---|---|---|---|---|---|---|
| For *k = 5* | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall |
| ***Average*** | | | | | | |
| The distribution of Y is balance | 53.88 | 53.33 | 53.61 | 90.81 | 91.26 | 91.04 |
| The distribution of Y is moderately unbalance | 54.64 | 53.66 | 54.34 | 88.95 | 96.00 | 91.14 |
| The distribution of Y is severely unbalance | 56.05 | 56.50 | 56.10 | 85.42 | 99.90 | 86.87 |
| ***Percentage of 100% prediction rate*** | | | | | | |
| The distribution of Y is balance | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| The distribution of Y is moderately unbalance | 0.00 | 0.00 | 0.00 | 0.00 | 12.00 | 0.00 |
| The distribution of Y is severely unbalance | 0.00 | 0.00 | 0.00 | 0.00 | 98.00 | 0.00 |
| | Semiparametric Choice Model 1 | | | Semiparametric Choice Model 2 | | |
| For *k = 10* | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall |
| ***Average*** | | | | | | |
| The distribution of Y is balance | 53.70 | 54.54 | 54.12 | 94.28 | 93.58 | 93.93 |
| The distribution of Y is moderately unbalance | 56.17 | 57.00 | 56.42 | 90.69 | 97.73 | 92.80 |
| The distribution of Y is severely unbalance | 56.81 | 62.00 | 57.43 | 88.03 | 99.83 | 89.45 |
| ***Percentage of 100% prediction rate*** | | | | | | |
| The distribution of Y is balance | 0.00 | 0.00 | 0.00 | 8.00 | 1.00 | 0.00 |
| The distribution of Y is moderately unbalance | 0.00 | 0.00 | 0.00 | 1.00 | 54.00 | 0.00 |
| The distribution of Y is severely unbalance | 0.00 | 1.00 | 0.00 | 0.00 | 98.00 | 0.00 |

**Table 7.** Comparison of predictive ability for low dimensional case ($n > p$) with equal cluster size
(by functional form of the data-generating model)

| | Ordinary Probit Regression | | | Semiparametric Choice Model 1 | | | Semiparametric Choice Model 2 | | |
|---|---|---|---|---|---|---|---|---|---|
| For *k = 5* | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall |
| ***Average*** | | | | | | | | | |
| Y is generated linearly | 60.16 | 60.14 | 60.15 | 55.68 | 56.44 | 56.06 | 93.50 | 92.42 | 92.96 |
| Y is generated nonlinearly | 60.36 | 59.20 | 59.78 | 55.76 | 56.08 | 55.92 | 93.10 | 93.96 | 93.53 |
| ***Percentage of 100% prediction rate*** | | | | | | | | | |
| Y is generated linearly | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 4.00 | 1.00 | 0.00 |
| Y is generated nonlinearly | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 4.00 | 2.00 | 0.00 |
| | Ordinary Probit Regression | | | Semiparametric Choice Model 1 | | | Semiparametric Choice Model 2 | | |
| For *k = 10; Balanced* | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall |
| ***Average*** | | | | | | | | | |
| Y is generated linearly | 59.76 | 60.02 | 59.89 | 56.40 | 57.46 | 56.93 | 92.76 | 92.90 | 92.83 |
| Y is generated nonlinearly | 61.62 | 60.26 | 60.94 | 56.90 | 56.52 | 56.71 | 93.46 | 93.80 | 93.63 |
| ***Percentage of 100% prediction rate*** | | | | | | | | | |
| Y is generated linearly | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 2.00 | 1.00 | 0.00 |
| Y is generated nonlinearly | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 3.00 | 4.00 | 0.00 |
| | Ordinary Probit Regression | | | Semiparametric Choice Model 1 | | | Semiparametric Choice Model 2 | | |
| For *k = 10; Severely Unbalanced* | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall |
| ***Average*** | | | | | | | | | |
| Y is generated linearly | 100.00 | 0.00 | 90.00 | 75.90 | 61.60 | 74.47 | 85.66 | 99.80 | 87.07 |
| Y is generated nonlinearly | 100.00 | 0.00 | 90.00 | 75.13 | 60.20 | 73.64 | 85.73 | 100.00 | 87.16 |
| ***Percentage of 100% prediction rate*** | | | | | | | | | |
| Y is generated linearly | 100.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 98.00 | 0.00 |
| Y is generated nonlinearly | 100.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 100.00 | 0.00 |

**Table 8.** Comparison of predictive ability for high dimensional case ($n < p$) with equal cluster size (by functional form of the data-generating model)

| | Semiparametric Choice Model 1 | | | Semiparametric Choice Model 2 | | |
|---|---|---|---|---|---|---|
| For *k = 5* | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall |
| ***Average*** | | | | | | |
| Y is generated linearly | 53.44 | 53.84 | 53.64 | 94.06 | 93.76 | 93.91 |
| Y is generated nonlinearly | 97.38 | 97.32 | 97.35 | 93.00 | 94.62 | 93.81 |
| ***Percentage of 100% prediction rate*** | | | | | | |
| Y is generated linearly | 0.00 | 0.00 | 0.00 | 5.00 | 3.00 | 0.00 |
| Y is generated nonlinearly | 79.00 | 73.00 | 59.00 | 6.00 | 7.00 | 0.00 |
| | Semiparametric Choice Model 1 | | | Semiparametric Choice Model 2 | | |
| For *k = 10; Balanced* | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall |
| ***Average*** | | | | | | |
| Y is generated linearly | 54.72 | 55.26 | 54.99 | 93.98 | 94.42 | 94.20 |
| Y is generated nonlinearly | 94.66 | 95.08 | 94.87 | 93.54 | 95.34 | 94.44 |
| ***Percentage of 100% prediction rate*** | | | | | | |
| Y is generated linearly | 0.00 | 0.00 | 0.00 | 6.00 | 6.00 | 0.00 |
| Y is generated nonlinearly | 2.00 | 5.00 | 0.00 | 1.00 | 12.00 | 0.00 |
| | Semiparametric Choice Model 1 | | | Semiparametric Choice Model 2 | | |
| For *k = 10; Severely Unbalanced* | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall |
| ***Average*** | | | | | | |
| Y is generated linearly | 55.86 | 58.20 | 56.09 | 87.32 | 99.70 | 88.56 |
| Y is generated nonlinearly | 73.74 | 97.90 | 76.16 | 86.04 | 99.90 | 87.43 |
| ***Percentage of 100% prediction rate*** | | | | | | |
| Y is generated linearly | 0.00 | 1.00 | 0.00 | 0.00 | 97.00 | 0.00 |
| Y is generated nonlinearly | 0.00 | 79.00 | 0.00 | 0.00 | 99.00 | 0.00 |

**Table 9.** Comparison of predictive ability for low dimensional case ($n > p$) with equal cluster size
(by presence or absence of misspecification in the model)

| | Ordinary Probit Regression | | | Semiparametric Choice Model 1 | | | Semiparametric Choice Model 2 | | |
|---|---|---|---|---|---|---|---|---|---|
| For *k = 5* | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall |
| ***Average*** | | | | | | | | | |
| Misspecification is absent | 60.16 | 60.14 | 60.15 | 55.68 | 56.44 | 56.06 | 93.50 | 92.42 | 92.96 |
| Misspecification is present | 60.98 | 60.86 | 60.92 | 56.90 | 55.92 | 56.41 | 92.48 | 92.12 | 92.30 |
| ***Percentage of 100% prediction rate*** | | | | | | | | | |
| Misspecification is absent | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 4.00 | 1.00 | 0.00 |
| Misspecification is present | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 3.00 | 0.00 |
| | Ordinary Probit Regression | | | Semiparametric Choice Model 1 | | | Semiparametric Choice Model 2 | | |
| For *k = 5; Balanced Response* | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall |
| ***Average*** | | | | | | | | | |
| Misspecification is absent | 59.82 | 61.04 | 60.43 | 56.66 | 56.50 | 56.58 | 93.54 | 93.22 | 93.38 |
| Misspecification is present | 60.10 | 61.38 | 60.74 | 56.76 | 56.80 | 56.78 | 91.20 | 92.28 | 91.74 |
| ***Percentage of 100% prediction rate*** | | | | | | | | | |
| Misspecification is absent | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 7.00 | 4.00 | 0.00 |
| Misspecification is present | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 3.00 | 0.00 |
| | Ordinary Probit Regression | | | Semiparametric Choice Model 1 | | | Semiparametric Choice Model 2 | | |
| For *k = 10; Balanced Response* | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall |
| ***Average*** | | | | | | | | | |
| Misspecification is absent | 59.76 | 60.02 | 59.89 | 56.40 | 57.46 | 56.93 | 92.76 | 92.90 | 92.83 |
| Misspecification is present | 59.14 | 59.62 | 59.38 | 56.24 | 56.74 | 56.49 | 91.74 | 91.54 | 91.64 |
| ***Percentage of 100% prediction rate*** | | | | | | | | | |
| Misspecification is absent | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 2.00 | 1.00 | 0.00 |
| Misspecification is present | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 3.00 | 0.00 |

**Table 10.** Comparison of predictive ability for high dimensional case ($n < p$) with equal cluster size (by presence or absence of misspecification in the model)

| | Semiparametric Choice Model 1 | | | Semiparametric Choice Model 2 | | |
|---|---|---|---|---|---|---|
| For $k = 5$ | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall |
| ***Average*** | | | | | | |
| Misspecification is absent | 53.44 | 53.84 | 53.64 | 94.06 | 93.76 | 93.91 |
| Misspecification is present | 54.18 | 54.50 | 54.34 | 94.16 | 93.88 | 94.02 |
| ***Percentage of 100% prediction rate*** | | | | | | |
| Misspecification is absent | 0.00 | 0.00 | 0.00 | 5.00 | 3.00 | 0.00 |
| Misspecification is present | 0.00 | 0.00 | 0.00 | 7.00 | 5.00 | 0.00 |
| | Semiparametric Choice Model 1 | | | Semiparametric Choice Model 2 | | |
| For *k = 5; Balanced Response Categories* | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall |
| ***Average*** | | | | | | |
| Misspecification is absent | 54.30 | 53.86 | 54.08 | 93.74 | 92.88 | 93.31 |
| Misspecification is present | 53.84 | 54.28 | 54.06 | 93.62 | 93.60 | 93.61 |
| ***Percentage of 100% prediction rate*** | | | | | | |
| Misspecification is absent | 0.00 | 0.00 | 0.00 | 3.00 | 3.00 | 0.00 |
| Misspecification is present | 0.00 | 0.00 | 0.00 | 4.00 | 1.00 | 0.00 |
| | Semiparametric Choice Model 1 | | | Semiparametric Choice Model 2 | | |
| For *k = 10; Balanced Response Categories* | Y=1 | Y=0 | Overall | Y=1 | Y=0 | Overall |
| ***Average*** | | | | | | |
| Misspecification is absent | 54.72 | 55.26 | 54.99 | 93.98 | 94.42 | 94.20 |
| Misspecification is present | 53.82 | 54.86 | 54.34 | 94.06 | 93.48 | 93.77 |
| ***Percentage of 100% prediction rate*** | | | | | | |
| Misspecification is absent | 0.00 | 0.00 | 0.00 | 6.00 | 6.00 | 0.00 |
| Misspecification is present | 0.00 | 0.00 | 0.00 | 5.00 | 1.00 | 0.00 |