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

# WORKING PAPER SERIES

**High Dimensional Nonparametric Discrete Choice Model**

by

Maureen Dinna D. Giron
and
Erniel B. Barrios

UPSS Working Paper No. 2010-09
August 2009

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

**High Dimensional Nonparametric Discrete Choice Model**

Maureen Dinna D. Giron[a] and Erniel B. Barrios[b]

**Abstract**

The functional form of a model can be a constraint in the correct prediction of discrete choices. The flexibility of a nonparametric model can increase the likelihood of correct prediction. The likelihood of correct prediction of choices can further be increased if more predictors are included, but as the number of predictors approaches or exceeds the sample size, more serious complications can be generated than the improvement in prediction. With high dimensional predictors in discrete choice modeling, we propose a generalized additive model (GAM) where the predictors undergo dimension reduction prior to modeling. A nonparametric link function is proposed to mitigate the deterioration of model fit as a consequence of dimension reduction. Using simulated data with the dependent variable having two or three categories, the method is comparable to the ordinary discrete choice model when the sample size is sufficiently large relative to the number of predictors. However, when the number of predictors exceeds substantially the sample size, the method is capable of correctly predicting the choices even if the components included in the model account for only 20% of the total variation in all predictors.

**Keywords:** discrete choice model, generalized additive model, high dimensional data, nonparametric model.

## 1. Introduction

The multinomial logit (MNL) model is used in predicting a categorical response variable conditional on a set of exogenous predictors. The model is widely used among modelers of choice or classification since it is computationally efficient and results are easy to interpret. However, just like any other models, the MNL model will provide reliable estimates only if certain assumptions are met, e.g. homogeneity among decision makers and independence of irrelevant alternatives. Furthermore, the linear-in-parameters constraint prevents it from being useful when the underlying relationship between the predictor variables and the utility function of the alternatives for the response variable is nonlinear.

The simplest way to add flexibility to any generalized linear model, including the MNL model described above, would be to replace the linear function of predictors with a nonparametric one, e.g., the nonlinear surface smoothers discussed by Hastie and Tibshirani (1990). However, doing so often results in complications in the interpretability of the model and in the curse of dimensionality. Several techniques have been proposed in the literature to prevent these difficulties. These techniques can be roughly classified into two types: those that aim to avoid the curse of dimensionality directly by reducing the number of dimensions such as Single Index Models (Hall et al, 1993); and those that try to alleviate the curse of dimensionality by simplifying the structure of the final model such as the Generalized

[a]Instructor, School of Statistics University of the Philippines, Diliman Quezon City, Philippines
[b]Professor, School of Statistics University of the Philippines, Diliman Quezon City, Philippines

Additive Models of Hastie and Tibshirani (1990) and Projection Pursuit Regression by Friedman and Stuetzle (1981).

Abe (1999) proposed a Generalized Additive Model (GAM) for discrete choice data which incorporates an additive rather than linear predictor index in the MNL model; relaxing the linear-in-parameter constraint of the MNL model while circumventing the curse of dimensionality that is the drawback of fully non-parametric multivariate MNL models. This model adapts the framework developed for matched case-control studies discussed by Hastie and Tibshirani (1990) and extends the algorithm to the multinomial case. Trial runs made on simulated data and a real-life application showed that the proposed GAM not only exhibited better fit compared to the Generalized Linear Model. Abe (1999) further illustrated that the method is robust even when no assumptions are made regarding the distribution of the random term, limitations are included as well. First, the method requires a large amount of data for the estimates to be reliable. Second, an essential assumption of generalized additive models, additive separability in covariates, may not be satisfied. Lastly, on the occasion that the number of predictors exceeds the number of observations, the viability of the method is questionable.

This study aims to assess the feasibility and efficiency of using principal components of the predictors in the generalized additive model (GAM) for discrete-choice data and determine its capacity to manage the drawbacks of high-dimensional data while satisfying the assumption of generalized additive models. It also aims to determine the effects of response category distribution and significant correlations between the original variables on the predicted output using the proposed model.

## 2. High Dimensional data

When the number of explanatory variables, $p$, is large and a single non-parametric function of the explanatory variables is used to construct the net utility, the model is likely to suffer from the curse of dimensionality (Bellman, 1961), i.e., an increase in variance that requires an exponential increase in the sample size requirement for the model to be reliable as the number of dimensions increases.

In this study, to deal with the curse of dimensionality, a technique that uses a combination of both data reduction and function approximation is to be explored. The use of principal components analysis and careful selection of the number of principal components to use will directly reduce the dimensionality of the data while minimizing the loss of information (variability), possibly addressing the limitations of the GAM proposed by Abe (1999). The reduced dimensionality will to the reduction of the sample size requirement due to the nonparametric model specification, but still correctly capturing the effect of the predictors on the response variable. The method of model-building proposed here makes it possible to conduct analysis of categorical data involving a substantial number of predictors, even when the number of predictors exceeds the sample size.

### 2.1 Additive Separability of Covariates

The fundamental assumption of GAMs – additive separability of covariates - alleviates the curse of dimensionality, since, instead of a single function, the utility is the sum of $p$ non-parametric functions. Additive separability exists when the removal of one or more explanatory variables from the model does not alter the contribution of the other variables on the predicted response variable, i.e. the nonparametric

functions of the other variables do not change, (Abe, 1999). Another advantage of this assumption is the ease of interpretation. Linear Regression Models and GAMs are similar with respect to one important characteristic: they are both additive in predictor effects. This implies that once we have fitted the model to the data, it is possible to determine the impact of a single factor on the choice being made when all other variables are fixed and without interactions among the variables, (Hastie and Tibshirani, 1990).

Additive separability requires that correlations between explanatory factors be ignored, when in reality, associations between different variables are part of the multidimensional framework (Weymark, 2005). When predictors are significantly correlated, the influence of one covariate on the model will shift to the variable/s to which it is correlated once it is removed, violating the assumption of additive separability. To address this issue, the proposed model uses principal components, which, being uncorrelated, have no mutual linear dependence on one another and automatically satisfies this assumption.

### 2.2 Principal Components Regression

Principal components analysis (PCA) prior to regression has been used in linear regression in mitigating the ill-effects of multicollinearity among the predictor variables into modeling. When two predictors are moderately to highly correlated with one another, linear regression of the original predictor variables will result in unstable and sometimes deceptive parameter estimates with significantly large variances. If PCA is used in pre-analyzing as regression model, the multicollinearity manifests itself as principal components (PC) with very small variance. Since the choice of the components to be included in the model are dependent upon the proportion of original variance that they can explain, that some PCs will be dropped necessarily. The PCs that included in the regression model are orthogonal, thus, and are already free from implications on the multicollinearity problem. The estimators of regression coefficient from this method are usually biased, but have variances that are far smaller than the variances of estimators produced by ordinary linear regression. (Joliffe, 2002).

## 3. Methodology

The data consists of a single categorical response variable, Y, which can take on $r + 1$ possible alternatives, and $p$ predictor variables, $X_1, X_2, ..., X_p$, measured on $n$ subjects, $\mathbf{X}=\begin{bmatrix} x_{11} & x_{12} & \text{K} & x_{1p} \\ x_{21} & x_{22} & \text{K} & x_{2p} \\ \text{M} & \text{M} & \text{O} & \text{M} \\ x_{n1} & x_{n2} & \text{K} & x_{np} \end{bmatrix}$

with sample covariance matrix $\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})^T$

The following notations will be used throughout this paper:

| | |
|---|---|
| n | sample size |
| p | number of predictors |
| r | number of categories of the response variable |
| q | number of principal components selected |
| **X** | the matrix containing the predictors |
| Y | the response variable, with possible values $Y_1, Y_{2, ..., } Y_n$ |

### 3.1 Principal Components Analysis on the p Predictors

Principal components analysis is used a pre-analysis scheme that will both mitigate high dimensionality and induce additive effects of the predictors. The number of principal components used in the model can significantly influence the attainment of the goal of the pre-analysis as well as the achievement of goals in model-building.

In the conventional case where the sample size is reasonably larger than the number of predictors, the first q PCs such that the cumulative percentage of variation contributed by the components is at least 80% or such that the percentage of total variation contributed by the (q+1)th PC is less than or equal to 2% is recommended by (Joliffe, 2002). When n<p, we propose that fewer PCs be included, say, that accounting for 20% of the total variation only.

### 3.2 Local Scoring Algorithm

The local scoring algorithm follows the algorithm proposed by Abe (1999), based on the original algorithm described by Hastie and Tibshirani (1990) for matched case-control data, modified in accordance to a generalized logit model using the last alternative as the baseline category. In the multinomial logit model, the logit of each response category *k*, $\log\left(\frac{\pi_k}{\pi_r}\right)$, rather than the original variable Y is modeled. The local scoring algorithm must be performed for each of the $1^{st}$ *r – 1* categories of the response variable. The estimated values of the logit function are then used to estimate the probabilities associated with the *r* response categories.

*Step 1:* Obtain the preliminary estimates for the probability ($\hat{\mu}_{ik}$) and logit (η) of each response category. For the multinomial logit model with r categories, we use the last ($r^{th}$) category as the baseline. Solve for the log-likelihood based on the initial values.

$$\hat{\mu}_{ik} = P(Y = k) = \frac{\sum_{i=1}^{n} Y_{ki}}{n}, \qquad k = 1, \ldots, r \tag{3.1}$$

$$\eta\,(\underline{C}_{ik}) = \log\left(\frac{P(Y=k)}{P(Y=r)}\right) \qquad \forall\ i = 1, 2, \ldots, n; \qquad k = 1, 2, \ldots, r-1 \tag{3.2}$$

$$\log\text{–likelihood} = \sum_{k=1}^{r}\sum_{i=1}^{n} I_{\{k\}}(Y_i)P(Y=k) \tag{3.3}$$

*Step 2*: Define the variable $z_{ik}$ as adjusted logit function at the $k^{th}$ response. This will be the response variable in the backfitting algorithm. Compute for the weight using the predicted probability of selecting the $k^{th}$ response.

$$z_{ik} = \eta(c_{ik}) + \frac{y_{ik} - \hat{\mu}_{ik}}{\hat{\mu}_{ik}(1-\hat{\mu}_{ik})} \tag{3.4}$$

$w_{ik}$

$$= \hat{\mu}_{ik}(1-\hat{\mu}_{ik}) \tag{3.5}$$

*Step 3*: Use the *backfitting algorithm* to model the response variable $z_{ik}$ and obtain a non-parametric function for each of the covariates. Apply the results of the backfitting iterations to update the value of η ($\underline{C}_{ik}$)

$$\eta(C_{ik}) = \log\left(\frac{P(Y=k \mid \underline{c})}{P(Y=r+1 \mid \underline{c})}\right) = \beta_o + \beta_1 C_1 + \beta_2 C_2 + ... + \beta_q C_q \qquad (3.6)$$

where $C_i$ is the ith principal component

*Step 4*: After estimating the value of η($\underline{C}_{ik}$) for all k, compute for the estimated probability that response variable Y will take on the category k. The category with the highest probability will be the predicted value of Y for the $i^{th}$ observation. Using the latest values of $\hat{\mu}_{ik}$, we compute the current log-likelihood value using equation 3.3.

$$\hat{\mu}_{ir} = \frac{1}{1 + \sum_{k=1}^{r-1} e^{\eta_{ik}(c_{ik})}} \qquad (3.7)$$

$$\hat{\mu}_{ik} = \hat{\mu}_{ir} e^{\eta_{ik}(c_{ik})} \quad , k = 1, 2, \ldots, r-1 \qquad (3.8)$$

Repeat Steps 2 through 4 until the log-likelihood converges, i.e, until the difference in the log-likelihood between two consecutive iterations is smaller than 0.01.

### 3.3 The Backfitting Algorithm

The backfitting algorithm proposed by Hastie and Tibshirani (1990) is modified below:

*Step 1:* Initialization (at m = 0)

In step 2 of the local scoring algorithm, the logit function, $z_{ik}$ is obtained using equation 3.4. Let this $z_{ik}$ be the dependent variable. Set the intercept to be equal to the mean of the dependent variable and the preliminary functions of the principal components to be zero.

$s_o = E(Z_{ik})$ ; $s_1^0(x_1) = s_2^0(x_2) = \ldots = s_q^0(x_q) = 0$

*Step 2:* First iteration (m = 1), obtaining the sequence of entry of the components

Perform a correlation analysis between the response variable $z_{ik}$ and the predictors $C_1, \ldots, C_q$. The principal component with the highest absolute correlation will be the first predictor included ($x_1$). Adjust the response variable $z_{ik}$, use the adjusted response to estimate the regression model with predictor $x_1$.

$$R_1 = z_{ik} - s_o$$

$s_1^1(x_1) = E(R_1 \mid x_1)$

→ perform weighted (weight = $w_k$) simple linear regression on $R_1$ without intercept, $x_1$ as the predictor variable

Perform a correlation analysis between residual obtained from the regression procedure and the remaining components. Again, select the component with the largest absolute correlation, make this $x_2$. Adjust the response variable $z_{ik}$,

$$R_2 = z_{ik} - s_o - s_1(x_1)$$

$s_2^1(x_2) = E(R_2 \mid x_2)$

→ perform weighted (weight = $w_k$) simple linear regression on $R_2$ without intercept, $x_2$ as the predictor variable

Repeat the correlation analysis and regression procedure in a similar fashion until all the *q* components have been regressed on the adjusted dependent variable.

*Step 3:* Iterate (m = m + 1), following the sequence of entry obtained in Step 2

For j = 1, …, q

$$R_j = z_{ik} - s_o - \sum_{k<j} s_k{}^{m}(x_k) - \sum_{k>j} s_k{}^{m-1}(x_k)$$

$s_j(x_j) = E(R_j \mid x_j) \rightarrow$ perform weighted regression of $R_j$ without intercept and $x_j$ as the single predictor variable.

Repeat Step 2 until the difference in $s_j(x_j)$ for all j = 1, .., q, between consecutive iterations is sufficiently small (< 0.01). The deterministic component in the utility function that is used to construct the multinomial logit model is then a non-linear function of the principal components, such that the probability that the *ith* respondent selects the *jth* alternative follows the equation,

$$P(y = j \mid \underline{x}) = \frac{e^{v_{ij}}}{\sum_{l=1}^{r} e^{v_{il}}} \quad , \quad v_{ij} = \sum_{k=1}^{m} s(c_{ijk})$$

### 3.4 Simulation Study

Evaluation of the proposed method was conducted on simulated data under various scenarios and based on the predictive ability of the model. A correctly predicted response was defined as an individual with coinciding observed and predicted values. Summaries of the percentage of correctly predicted responses for each category and across all categories were computed for each set of replicates. For comparison purposes, the replicates in the large sample (n > p) scenarios were also modeled using the linear multinomial logit model and the generalized additive model for discrete choice data. Summary statistics were then used to determine which among the different models has better predictive ability.

Samples of dichotomous response and 3-category response variables were simulated according to the scenarios in Table 1.

Table 1. Simulation Settings

| | *Dichotomous Response* | *3-category Response* |
|---|---|---|
| Relationship between predictors | Uncorrelated, Correlated | Uncorrelated, Correlated |
| Sample size (n) and number of predictors (p) | n > p, n < p, n = p | N > p, n < p, n = p |
| Distribution of the observations in the response category | (Category 1; Category 2)<br>Balanced (50%, 50%)<br>Moderately Unbalanced (70%, 30%)<br>Severely Unbalanced (10%, 90%) | (Cat 1, Cat 2, Cat 3)<br>Balanced (33%, 33%, 34%)<br>Moderately Unbalanced - Type 1 (20%, 40%, 40%)<br>Moderately Unbalanced - Type 2 (60%, 20%, 20%)<br>Severely Unbalanced (70%, 20%, 10%) |

For the cases with uncorrelated predictors and sample size substantially larger than the number of predictors, the predictors are simulated as from $X_i \sim$ Uniform $(a_i, b_i)$ , i = 1,2 and $X_i \sim$ Normal $(\mu_i, \sigma^2_i)$, i = 3, 4, 5. For data with correlated predictors and sample size substantially larger than number of predictors, $X_1 \sim$ Uniform (a, b) and $X_2 \sim$ Normal$(\mu, \sigma^2)$. Then other predictors are derived from the first two, i.e.,

$X_3 = dX_1 + e + \varepsilon_3$ , where $\varepsilon_3 \sim$ Normal $(0, \sigma_1^2)$
$X_4 = fX_2 + g + \varepsilon_4$ , where $\varepsilon_4 \sim$ Normal $(0, \sigma_2^2)$
$X_5 = h X_1 + kX_2 + \varepsilon_5$ , where $\varepsilon_5 \sim$ Normal $(0, \sigma_3^2)$
d, e, f, g, h and k are constants

For uncorrelated data with fewer observations ($n$ = 30) than predictors ($p$ = 40), $X_1$, $X_2$, …, $X_{10}$ are normally distributed and $X_{11}$, $X_{12}$, …, $X_{20}$ follow a continuous uniform distribution. The twenty-first to thirtieth are also normally distributed while the last 10 are uniformly distributed. For correlated data with fewer observations ($n$ = 30) than predictors ($p$ = 40), the predictors $X_1$, $X_2$, …, $X_8$ are normally distributed, $X_9$, $X_{10}$, …, $X_{16}$ follow a continuous uniform distribution. The 8 predictors that follow linear combinations of the normally distributed predictors combined with a random error, the $25^{th}$ to $32^{nd}$ are linear combinations of the uniformly distributed variables combined with a random error. The last 8 variables are linear functions of pairs of normal and uniform random variables combined with a normally distributed random error.

## 4. Results and Discussion

In the traditional case where n>p, the proportions of correctly predicted responses of the proposed model are comparable to those of both the multinomial logistic regression model and the GAM for discrete choice data using the original variables, regardless of the distribution of the observations between the response categories. The results give evidence that the predictive power of the procedure is highly dependent upon the distribution of the observations. Data with more information on a single category leads to predicted choices that are biased in favor of that category. As the number of observations on a category increases, the procedure tends to assign more of the observations into the category endowed with more observation, to the point where all observations are predicted to fall into only a single category only.

When there are at least twice as many observations as there are predictors, the performance of the proposed model is at most comparable to that of logistic regression, even when correlation exists between the predictor variables. The rate of correct prediction, however, is more stable using the proposed method than GAM for discrete choice data. Like the uncorrelated case, all three models perform poorly in predicting choices when the data are biased towards one of the choices. Results for the correlated case and uncorrelated case are similar, this is not surprising since past studies had shown that multicollinearity of predictors affects the coefficient estimates but not necessarily the predicted values.

**Figure 1** - Proportion of correctly predicted observations by method when data has uncorrelated predictors

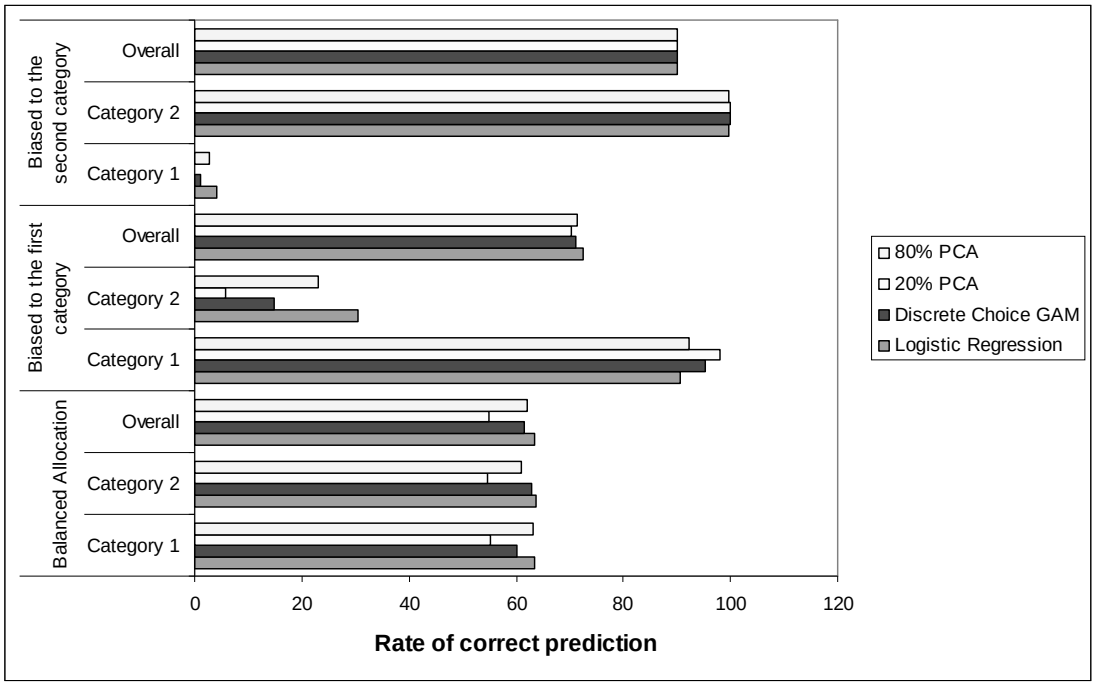
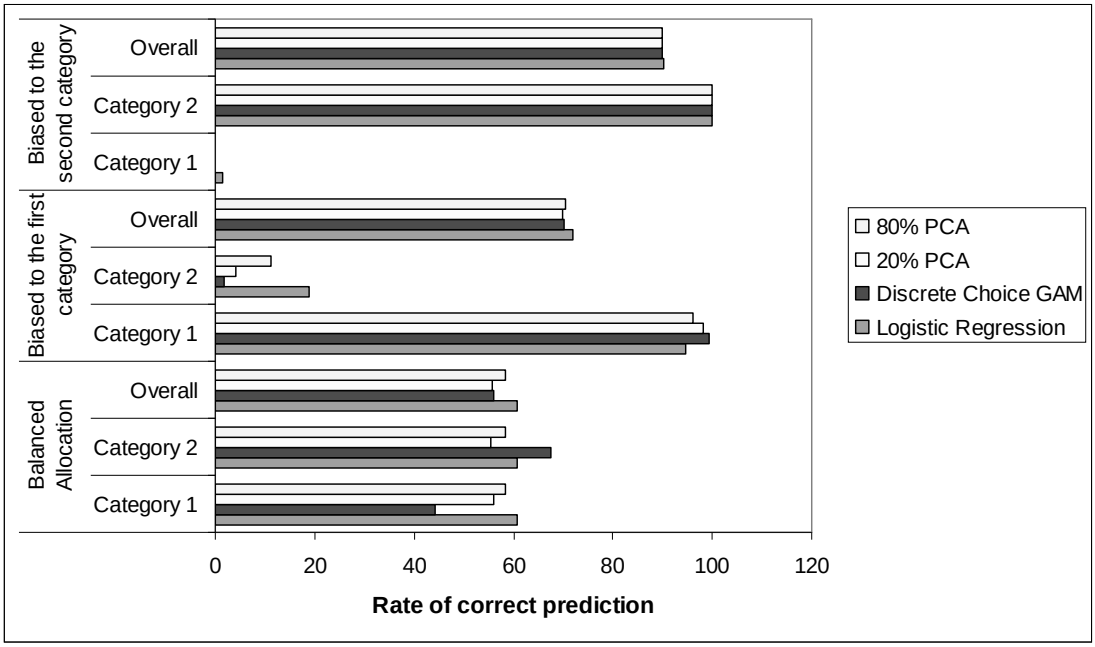



**Figure 2** - Proportion of correctly predicted observations by method when data has correlated predictors




Across all tables presented in the next section, statistics in each cell are mean and those enclosed in parenthesis are the minimum and maximum proportion of correctly classified observation. The following notations are used:

Logistic Regression – multinomial logit model using a linear relation

Discrete Choice GAM – generalized additive model for discrete choice by Abe

PC–GAM 1 – the proposed method using at PCs accounting for at least 20% of the cumulative variance explained.
PC-GAM 2 – the proposed method using at PCs accounting for at least 60% of the cumulative variance explained or variance explained by the next PC less than 2%.

### 4.1 Dichotomous Response Variable with n<p

The proposed method is advantageous when predicting choices from a dichotomous response variable if there are more predictors than observations. Similar is true when there are as many observations as the number of predictors (see Tables 2 and 3). Similar is true when there are as many observations as the number of predictors (see Tables 4 and 5). In the uncorrelated case, the model correctly predicted the response for up to 70% of the observations, even when the distribution of the sample was moderately imbalanced (up to 70% in one category). When the case is severely unbalanced, though, the capability of the model to correctly predict the response worsens.

In the correlated case, when the data are severely unbalanced and the number of predictors is at least equal to the sample size, the method encounters a problem similar to that in logistic regression (complete separation of data points), wherein a vector $\boldsymbol{b}$ exists that correctly allocates all observations to their response groups. When complete separation of data points occurs, one or more of the estimated parameter coefficients does not converge, even as the log-likelihood converges to zero. Heinz and Schemper (2002) noted that this estimation problem occurs often in severely disproportionate data with few observations, and whose response variable is strongly correlated to the explanatory variables. Complete separation in additive models also occurs when the data is severely unbalanced.

**Table 2 -** Proportion of correctly predicted observations by data distribution for data with uncorrelated predictors (p>n)

| | Balanced | | Moderately Unbalanced | | Severely Unbalanced | |
|---|---|---|---|---|---|---|
| | PC-GAM 1 | PC-GAM 2 | PC-GAM 1 | PC-GAM 2 | PC-GAM 1 | PC-GAM 2 |
| Choice 1 | 49.6<br>( 0, 80) | 69.7<br>( 0, 93.3) | 51.2<br>( 0, 100) | 51.6<br>( 0, 100) | 61.6<br>( 0, 100) | 61<br>( 0, 100) |
| Choice 2 | 77.9<br>( 46.7, 100) | 71.1<br>( 46.7, 100) | 69.7<br>( 0, 100) | 69.4<br>( 0, 100) | 38.4<br>( 0, 100) | 39<br>( 0, 100) |
| Overall | 63.8<br>( 50, 76.7) | 70.4<br>( 50, 86.7) | 62.9<br>( 50, 80) | 63.1<br>( 50, 80) | 40.7<br>( 10, 90) | 41.2<br>( 10, 90) |

**Table 3** - Proportion of correctly predicted observations by data distribution for data with correlated predictors (p>n)

| | Balanced | | Moderately Unbalanced | | Severely Unbalanced | |
|---|---|---|---|---|---|---|
| | PC-GAM 1 | PC-GAM 2 | PC-GAM 1 | PC-GAM 2 | PC-GAM 1 | PC-GAM 2 |
| Choice 1 | 71.4 ( 46.7, 93.3) | 71.6 ( 46.7, 93.3) | 90.8 ( 81, 100) | 85.7 ( 0, 95.2) | - | - |
| Choice 2 | 71.5 ( 13.3, 93.3) | 71.4 ( 13.3, 93.3) | 49 ( 0, 100) | 43.8 ( 22.2, 100) | - | - |
| Overall | 71.4 ( 33.3, 86.7) | 71.5 ( 33.3, 86.7) | 78.2 ( 60, 100) | 73.1 ( 30, 80) | - | - |

**Table 4** - Proportion of correctly predicted observations by data distribution using uncorrelated predictors (p=n)

| | Balanced | | Moderately Unbalanced | | Severely Unbalanced | |
|---|---|---|---|---|---|---|
| | PC-GAM 1 | PC-GAM 2 | PC-GAM 1 | PC-GAM 2 | PC-GAM 1 | PC-GAM 2 |
| Choice 1 | 71.3 (46.7 - 93.3) | 71.3 (46.7 - 93.3) | 89.7 (0 - 100) | 89.7 (0 - 100) | - | - |
| Choice 2 | 70.5 (13.3 - 100) | 70.7 (26.7 - 100) | 47.7 (0 - 100) | 47.9 (0 - 100) | - | - |
| Overall | 70.9 (30 - 93.3) | 71 (36.7 - 93.3) | 77.1 (30 - 96.7) | 77.1 (30 - 96.7) | - | - |

**Table 5** - Proportion of correctly predicted observations by data distribution using correlated predictors (p=n)

| | Balanced | | Moderately Unbalanced | | Severely Unbalanced | |
|---|---|---|---|---|---|---|
| | PC-GAM 1 | PC-GAM 2 | PC-GAM 1 | PC-GAM 2 | PC-GAM 1 | PC-GAM 2 |
| Choice 1 | 59.1 (40, 86.7) | 68.7 (0, 93.3) | 95.3 (85.7, 100) | 89.9 (0, 100) | - | - |
| Choice 2 | 61.3 (40, 80) | 72.4 (20, 100) | 21.9 (0, 66.7) | 55.4 (11.1, 100) | - | - |
| Overall | 60.2 (43.3, 76.7) | 70.6 (43.3, 86.7) | 73.3 (63.3, 86.7) | 79.6 (30, 100) | - | - |

### 4.2 Three Choice Response Variable with n>p

Tables 6-13 shows that generally, the multinomial logit model performs better compared to the proposed model and the GAM for discrete choice when the sample size is sufficiently larger than the number of predictors. It also tends to allocate the observations into the different categories in a more balanced manner. When predicting choices with more than 2 response choices, all models tend to perform poorly in the large-sample case, with the highest percentage of correctly predicted responses only slightly above 70%.

**Table 6** - Proportion of correctly predicted observations by method for balanced data with uncorrelated predictors

| | Logistic Regression | Discrete Choice GAM | PC-GAM 1 | PC-GAM 2 |
|---|---|---|---|---|
| Choice 1 | 74.7 (60.6, 93.9) | 84.2 (66.7, 97) | 59.2 (0, 97) | 82.2 (39.4, 97) |
| Choice 2 | 59.8 (36.4, 84.9) | 33.9 (6.1, 57.6) | 13.7 (0, 69.7) | 31.1 (9.1, 54.6) |
| Choice 3 | 85.8 (64.7, 94.1) | 78.9 (14.7, 100) | 59.2 (0, 91.2) | 69 (14.7, 97.1) |
| Overall | 73.6 (64, 87) | 65.8 (49, 76) | 44.2 (32, 58) | 60.9 (44, 72) |

**Table 7** – Proportion of correctly predicted observations by method for balanced data with correlated predictors

| | Logistic Regression | Discrete Choice GAM | PC-GAM 1 | PC-GAM 2 |
|---|---|---|---|---|
| Choice 1 | 48.4 (9.1, 69.7) | 23.6 (0, 100) | 53.3 (0, 78.8) | 53.2 (0, 78.8) |
| Choice 2 | 40 (9.1, 63.6) | 19.4 (0, 69.7) | 21.5 (0, 63.6) | 32.3 (0, 78.8) |
| Choice 3 | 51.6 (32.4, 76.5) | 72.7 (0, 100) | 44.1 (0, 82.4) | 40.3 (0, 76.5) |
| Overall | 46.7 (34, 59) | 38.9 (30, 54) | 39.7 (30, 50) | 41.9 (31, 53) |

**Table 8** - Proportion of correctly predicted observations by method for the first case of moderately unbalanced data with uncorrelated predictors

| | Logistic Regression | Discrete Choice GAM | PC-GAM 1 | PC-GAM 2 |
|---|---|---|---|---|
| Choice 1 | 35.1 (0, 90) | 45 (0, 85) | 12.2 (0, 70) | 51.2 (0, 100) |
| Choice 2 | 74.4 (55, 90) | 56.5 (32.5, 77.5) | 52.6 (17.5, 72.5) | 51.9 (32.5, 77.5) |
| Choice 3 | 85.9 (65, 95) | 85.7 (62.5, 100) | 65.1 (30, 87.5) | 79.3 (47.5, 95) |
| Overall | 71.2 (62, 85) | 65.9 (52, 78) | 49.5 (37, 60) | 62.7 (50, 75) |

**Table 9** - Proportion of correctly predicted observations by method for the first case of moderately unbalanced data with correlated predictors

| | Logistic Regression | Discrete Choice GAM | PC-GAM 1 | PC-GAM 2 |
|---|---|---|---|---|
| Choice 1 | 16.9 (0, 55) | 0.9 (0, 87.9) | 9.3 (0, 50) | 12.7 (0, 65) |
| Choice 2 | 58 (40, 75) | 50.3 (0, 100) | 53 (37.5, 70) | 54.5 (25, 75) |
| Choice 3 | 59.1 (42.5, 75) | 60.5 (2.5, 100) | 53.9 (12.5, 75) | 55.7 (20, 77.5) |
| Overall | 50.2 (41, 63) | 44.6 (36, 53) | 44.6 (31, 56) | 46.6 (37, 61) |

**Table 4.10** - Proportion of correctly predicted observations by method for the second case of moderately unbalanced data with uncorrelated predictors

| | Logistic Regression | Discrete Choice GAM | PC-GAM 1 | PC-GAM 2 |
|---|---|---|---|---|
| Choice 1 | 94.2 (88.3, 98.3) | 95.4 (86.7, 100) | 95.1 (83.3, 100) | 91.2 (85, 98.3) |
| Choice 2 | 34.7 (0, 75) | 36.3 (5, 75) | 15.9 (0, 55) | 46.3 (10, 90) |
| Choice 3 | 79.3 (40, 100) | 6.8 (0, 65) | 1 (0, 35) | 11.4 (0, 50) |
| Overall | 79.3 (66, 90) | 65.9 (59, 75) | 60.4 (55, 66) | 66.2 (56, 80) |

**Table 4.11** - Proportion of correctly predicted observations by method for the second case of moderately unbalanced data with correlated predictors

| | Logistic Regression | Discrete Choice GAM | PC-GAM 1 | PC-GAM 2 |
|---|---|---|---|---|
| Choice 1 | 95 (85, 100) | 99 (3.3, 100) | 97.8 (88.3, 100) | 96.9 (86.7, 100) |
| Choice 2 | 7.9 (0, 35) | 0.7 (0, 55) | 6.1 (0, 35) | 7.9 (0, 40) |
| Choice 3 | 14.5 (0, 45) | 0.4 (0, 40) | 0.1 (0, 5) | 0.3 (0, 5) |
| Overall | 61.5 (54, 67) | 59.6 (21, 61) | 59.9 (57, 64) | 59.8 (56, 65) |

**Table 4.12** - Proportion of correctly predicted observations by method for severely unbalanced data with uncorrelated predictors

| | Logistic Regression | Discrete Choice GAM | PC-GAM 1 | PC-GAM 2 |
|---|---|---|---|---|
| Choice 1 | 94.9 (90, 98.6) | 95.7 (90, 100) | 95.4 (84.3, 100) | 90.8 (84.3, 97.1) |
| Choice 2 | 48.6 (5, 80) | 43.8 (5, 85) | 16.6 (0, 55) | 60 (10, 95) |
| Choice 3 | 71 (0, 100) | 0 (0, 0) | 0 (0, 0) | 0 (0, 0) |
| Overall | 83.3 (70, 94) | 75.7 (68, 84) | 70.1 (65, 75) | 75.5 (65, 87) |

**Table 4.13** - Proportion of correctly predicted observations by method for the severely unbalanced data with correlated predictors

| | Logistic Regression | Discrete Choice GAM | PC-GAM 1 | PC-GAM 2 |
|---|---|---|---|---|
| Choice 1 | 97.7 (90, 100) | 100 (98.6, 100) | 98.6 (91.4, 100) | 97.9 (87.1, 100) |
| Choice 2 | 8 (0, 35) | 0.3 (0, 20) | 3.8 (0, 25) | 6.2 (0, 35) |
| Choice 3 | 6.1 (0, 40) | 0 (0, 0) | 0 (0, 0) | 0 (0, 0) |
| Overall | 70.6 (67, 76) | 70 (70, 73) | 69.8 (66, 74) | 69.8 (65, 73) |

## 4.3 Three Choice Response Variable with n<p

When the number of predictors is greater than or equal to the number of observations, the performance of the proposed method fare better, especially when the predictor variables are correlated. In Tables 14 and 16, the uncorrelated cases, the proposed model using PCs selected based on 80% explained variance only yield better prediction compared to the model using only 20% explained variance. In tables 15 and 17, the prediction proportions are the similar, although the reason behind this is that both models used the same number of components.

In the three-choice response case, the proposed method also performs better when the data is uncorrelated. This is due to the number of predictors included in the model. In the correlated case, a large part of the overall variance is captured by the first few (usually 3-5) components; while in the uncorrelated case, the predictors included in the sample ranged from 7 to 9 components. Though they may contain the same amount of cumulative variance explained, the relationships between the components and the response variable are not the same.

**Table 14** - Proportion of correctly predicted observations by data distribution for data with uncorrelated predictors (p>n)

| | Balanced | | Moderately Unbalanced 1 | | Moderately Unbalanced 2 | |
|---|---|---|---|---|---|---|
| | PC-GAM 1 | PC-GAM 2 | PC-GAM 1 | PC-GAM 2 | PC-GAM 1 | PC-GAM 2 |
| Choice 1 | 59.2 (0, 97) | 82.2 (39.4, 97) | 12.2 (0, 70) | 51.2 (0, 100) | 95.1 (83.3, 100) | 91.2 (85, 98.3) |
| Choice 2 | 13.7 (0, 69.7) | 31.1 (9.1, 54.6) | 52.6 (17.5, 72.5) | 51.9 (32.5, 77.5) | 15.9 (0, 55) | 46.3 (10, 90) |
| Choice 3 | 59.2 (0, 91.2) | 69 (14.7, 97.1) | 65.1 (30, 87.5) | 79.3 (47.5, 95) | 1 (0, 35) | 11.4 (0, 50) |
| Overall | 44.2 (32, 58) | 60.9 (44, 72) | 49.5 (37, 60) | 62.7 (50, 75) | 60.4 (55, 66) | 66.2 (56, 80) |

**Table 15** - Proportion of correctly predicted observations by data distribution for data with correlated predictors (p>n)

| | Balanced | | Moderately Unbalanced 1 | | Moderately Unbalanced 2 | |
|---|---|---|---|---|---|---|
| | PC-GAM 1 | PC-GAM 2 | PC-GAM 1 | PC-GAM 2 | PC-GAM 1 | PC-GAM 2 |
| Choice 1 | 90.9 (40, 100) | 90.9 (40, 100) | 89.6 (41.7, 100) | 89.4 (41.7, 100) | 90.8 (16.7, 100) | 90.8 (16.7, 100) |
| Choice 2 | 64.1 (20, 100) | 64.1 (20, 100) | 66.7 (33.3, 91.7) | 66.7 (33.3, 91.7) | 74.3 (0, 100) | 74.3 (0, 100) |
| Choice 3 | 57.4 (53.3, 100) | 57.4 (53.3, 100) | 70.6 (50, 100) | 70.9 (50, 100) | 40.2 (63.3, 100) | 40.2 (63.3, 100) |
| Overall | 70.8 (0, 90) | 70.8 (0, 90) | 72.9 (9, 93.3) | 72.9 (9, 93.3) | 77.4 (0, 96.7) | 77.4 (0, 96.7) |

**Table 16** - Proportion of correctly predicted observations by data distribution for data with uncorrelated predictors (p=n)

| | Balanced | | Moderately Unbalanced 1 | | Moderately Unbalanced 2 | |
|---|---|---|---|---|---|---|
| | PC-GAM 1 | PC-GAM 2 | PC-GAM 1 | PC-GAM 2 | PC-GAM 1 | PC-GAM 2 |
| Choice 1 | 57.3 (0, 100) | 72.6 (10, 100) | 26 (0, 100) | 52 (0, 100) | 94.3 (77.8, 100) | 88.7 (72.2, 100) |
| Choice 2 | 36.7 (0, 80) | 49.6 (0, 90) | 51.2 (8.3, 83.3) | 55 (25, 83.3) | 18 (0, 83.3) | 42.2 (0, 100) |
| Choice 3 | 34 (0, 100) | 37.2 (0, 90) | 55.3 (8.3, 91.7) | 59.9 (0, 91.7) | 2.7 (0, 66.7) | 10.2 (0, 66.7) |
| Overall | 42.7 (30, 56.7) | 53.1 (36.7, 70) | 47.8 (23.3, 63.3) | 56.4 (33.3, 83.3) | 60.7 (53.3, 73.3) | 63.7 (50, 83.3) |

**Table 17** - Proportion of correctly predicted observations by data distribution for data with correlated predictors (p=n)

| | Balanced | | Moderately Unbalanced 1 | | Moderately Unbalanced 2 | |
|---|---|---|---|---|---|---|
| | PC-GAM 1 | PC-GAM 2 | PC-GAM 1 | PC-GAM 2 | PC-GAM 1 | PC-GAM 2 |
| Choice 1 | 75.1 (40, 100) | 75.4 (40, 100) | 68.3 (16.7, 100) | 68.3 (16.7, 100) | 88.1 (72.2, 100) | 88.2 (72.2, 100) |
| Choice 2 | 61.7 (20, 100) | 62 (20, 100) | 68.9 (41.7, 91.7) | 68.9 (41.7, 91.7) | 62 (16.7, 100) | 62.5 (16.7, 100) |
| Choice 3 | 53.8 (20, 90) | 54.2 (20, 100) | 63.9 (33.3, 91.7) | 63.9 (33.3, 91.7) | 37 (0, 83.3) | 37.3 (0, 100) |
| Overall | 63.5 (43.3, 83.3) | 63.9 (43.3, 100) | 66.8 (46.7, 83.3) | 66.8 (46.7, 83.3) | 72.7 (56.7, 90) | 72.9 (56.7, 100) |

## 5. Conclusions and Recommendations

In discrete choice models with high dimensional predictors, principal components analysis can be used in dimension reduction. This will further abate the potential curse of dimensionality when nonparametric models are fitted using these high dimensional predictors. Although dimension reduction minimizes the loss in variance due from the ignored components, model fit necessarily suffers (e.g., bias introduction, inferior model fit).

In the dichotomous case with sufficiently large sample size, the generalized additive model using principal components yields comparable results to both the original multinomial logit model and the generalized additive model for discrete choice data. However, when there are three response categories, and sufficiently large sample size, it is better to use the original multinomial logit model.

In the dichotomous response case, correlation between variables does not strongly affect the predictive power of the proposed model. However, the predictive performance of the model when the covariates are correlated is poorer than for data with independent covariates for the three-choice setting. This is an effect of the number of components selected; models using uncorrelated covariates will utilize more principal components, thus providing better fit. Also, correlation between predictors increases the amount of computing time needed to build a model using the generalized additive model for discrete-choice data. Thus, using principal components of covariates in the model will result to faster computation and subsequent convergence.

The generalized additive model using principal components yields good predictive ability when the data are high-dimensional, especially when the number of predictors exceeds the number of observations in the sample. Similar to other methods of predicting discrete choices, the generalized additive model using PCs is not practical for use with highly unbalanced data. In general, when approximately 20% of the observed responses belong to one category, the model can correctly predict 10% - 40% only of the observations into that category.

## References


Abe, M. (1999) *A Generalized Additive Model for Discrete Choice Data*. Journal of Business and Economic Statistics, Vol. 17, No. 3, pp. 271-284.

Agresti, A. (2007) *An Introduction to Categorical Data Analysis.* 2nd Edition. Hoboken, NJ: John Wiley & Sons, Inc.

Aguilera, A., Escabias, M. and Valderrama, M. (2006) *Using principal components for estimating logistic regression with high-dimensional multicollinear data.* Computational Statistics & Data Analysis, 50, pp. 1905 – 1924.

Bellman, R.E. (1961) *Adaptive Control Processes.* Princeton, NJ: Princeton University Press.

Borooah, V. (2002) *Logit and Probit. Ordered and Multinomial Models.* Sage University Paper series on Quantitative Applications on the Social Sciences, series no. 07-138. Beverly Hills, CA: Sage.

Dunteman, G. (1989) *Principal Components Analysis.* Sage University Paper series on Quantitative Applications on the Social Sciences, series no. 07-069. Beverly Hills, CA: Sage.

Friedman, J. H. and Stuetzle, W. (1981). *Projection pursuit regression*. Journal of the American Statistical Association. 76 817-823.

Fukuda, D. and Yai, T. (2010) *Semiparametric specification of the utility function in a travel mode choice model*. Transportation. Vol 37, No. 2, pp. 221-238.

Hall, P., Hardle, W., and Ichimura, H. (1993) *Optimal Smoothing in Single Index Models.* The Annals of Statistics, Vol. 21, No. 1, pp. 157-178.

Hardle, W. (1990). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

Hastie, T., and Tibshirani, R. (1990) *Generalized Additive Models*. New York: Chapman and Hall.

Heinze, G., and Schemper, M. (2002) *A solution to the problem of separation in logistic regression.* Statistics in Medicine, 21, pp. 2409-2419.

Heneman, R.L., Porter, G., Greenberger, D. and Strasser, S. (1997) *Modeling the relationship between pay level and pay satisfaction.* Journal of Business and Psychology, 12(2), pp. 147 – 158.

Jolliffe, I.T., (2002). *Principal Components Analysis*. New York: Springer-Verlag.

Marx, B.D. and Smith, E.P. (1990) *Principal Component Estimation for Generalized Linear Regression*. Biometrika, Vol. 77, No. 1, pp. 23-31.

Vines, S. K. (2000). *Simple Principal Components* . Applicational Satistics, Vol. 49, Part 4, pp. 441-451.

Weymark, J.A. (2005). *Measurement theory and the foundations of utilitarianism.* Social Choice and Welfare, vol. 25, issue 2, pp. 527 – 555.

Wang, L., and Yang, L. (2007). *Spline-backfitted kernel smoothing of nonlinear additive autoregression model*. Annals of Statistics, Vol. 35, Number 6, pp. 2474-2503.