14th National Convention on Statistics (NCS)
Crowne Plaza Manila, Quezon City
October 1-3, 2019

INFERENCE ON HIGH DIMENSIONAL DISCRETE CHOICE MODEL

by

Danilo R. Si
Joseph Ryan G. Lansangan, Ph.D.
Erniel B. Barrios, Ph.D.

For additional information, please contact:

Author's name	Danilo R. Si
Designation	Ateneo Senior High School faculty
Author's name	Joseph Ryan G. Lansangan, Ph.D.
Designation	UP School of Statistics Faculty
Author's name	Erniel B. Barrios, Ph.D.
Designation	UP School of Statistics Faculty
Affiliation	UP School of Statistics
Address	University of the Philippines Diliman, Quezon City
Tel. no.	426-6001 loc 6068
E-mail	dsi@ateneo.edu, jglansangan@up.edu.ph, ebbarrios@up.edu.ph

INFERENCE ON HIGH DIMENSIONAL DISCRETE CHOICE MODEL

by

Danilo R. Si
Joseph Ryan Lansangan, Ph.D.
Erniel B. Barrios, Ph.D.

ABSTRACT

Discrete choice models are commonly applied in decision making contexts. But problems in predictive performance, variable selection and/or model interpretation arise when the number of predictors greatly exceeds the sample size. In this study, such problems are mitigated via a heuristic approach combining dimension reduction and multiple comparisons correction. As a pre-process for variable selection, dimension reduction on high-dimensional inputs of the discrete choice model is conducted. Variables that were selected are then subjected to a local scoring algorithm with backfitting. To lower the false discovery rate, the Benjamini-Hochberg (BH) procedure is then implemented on the significant predictors resulting from the backfitting. Simulation studies show that most of the balanced cases are correctly sized, and consistently, the proposed test procedure is more powerful than the ordinary Bonferroni multiple comparisons testing procedure.

Keywords: *discrete choice model, backfitting algorithm, high dimensional data, dimension reduction, variable selection, multiple testing, false discovery rate, BH procedure*

1. INTRODUCTION

Discrete choice models are presented as development and renovation of the classical choice theory. These models have overcome the rigidities and inadequacies of consumer behavior study by mentioning the problems of economic agent choices in random specific environment for each situation involving the choice between mutually exclusive alternatives (Aloulou, 2018). The model has been used extensively to discrete choice processes in different fields such as econometrics (McFadden, 1974; Manski and McFadden, 1981) and transportation (Ben-Akiva and Lerman, 1985) to name some. These resulted with great success because of the model's analytical and computational tractability (Abe, 1999).

Abe (1999) developed a methodology for discrete choice data using the generalized additive model (GAM). This method incorporates an additive predictor instead of a linear predictor for the Multinomial logit (MNL) model. This relaxes the linear-in parameter constraint of the MNL model while circumventing the curse of dimensionality which is the drawback of fully nonparametric multivariate MNL models (Giron, 2010).

Giron (2010) assessed the feasibility and efficiency of using principal components of the predictors in the generalized additive mode (GAM) for discrete choice data and determine its capacity to manage drawbacks of high-dimensional data. Giron (2010) showed that the principal component analysis could be used in dimension reduction for high-dimensional discrete choice data and resulted with comparable results with the original multinomial logit model and the generalized additive model for discrete choice data.

Torres (2013) extended the work of Giron (2010) by using generalized adaptive sparse-PCA (GAS-PCA) as a data reduction tool and used GAM for high dimensional discrete data. Torres (2013) compared the results to PCA for estimating high-dimensional discrete choice data and yielded comparable predictive ability. Also, GAS-PCA yielded very sparse PC loadings compared to the generalized additive model using principal component analysis (Torres, 2013).

A variable selection procedure is being proposed using two methods of dimension reduction through sparse principal components: The Penalized Matrix Decomposition and the Generalized Adaptive Sparse Principal Component analysis.

This study aims to conduct variable selection process on the explanatory variables using two methods of sparse principal component analysis. The selected variables from the above method will be tested for significance thru the local scoring algorithm and the backfitting algorithm. Lastly, the proposed test aims to mitigate the false discovery rate which is very common in high-dimensional discrete choice data.

2. METHODOLOGY

The methodology starts with the data generating process to create a high dimensional data for discrete choice model. The proposed estimation proposed procedure composed of the following: Principal Component Analysis, Sparse Principal Component Analysis using Penalized Matrix Decomposition Sparse Principal Component (PMD-SPC) and General Adaptive Sparse Principal Component Analysis (GAS-PCA), Local Scoring Algorithm and Backfitting Algorithm. The proposed estimation procedure is then followed by the hypothesis testing on significant predictors, controlling false discovery rate and power and size computation. The design of the simulation study is stated also including all settings for each scenario.

2.1 Data Generating Process

Let the postulated model be

$$P\left(Y = y_j | \underline{X} = \underline{x}_j^*\right) = \frac{exp\left(\underline{X_j^*}\beta_p\right)}{\sum_{i=1}^{J} exp\left\{\underline{X_i^*}\beta_p + \epsilon_i\right\}} \quad i = 1,2,..J$$

Where,

$\underline{X_j^*}$	matrix of predictors/characteristics of individual j
Y	the response variables, with possible values, Y_1, Y_2, …, Y_r
β_p	beta coefficients
ϵ_i	random error

The data consists of a single categorical response variable, Y, which can take on j possible alternatives, and p predictor variables, $x_1, x_2, \dots, x_p$, measured on n subjects, with population covariance matrix $S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})^T$.

The following notations will be used throughout this study:

n	sample size
p	number of dimensions/ predictors
j	number of categories of the response variable
q	number of sparse principal components selected

2.2 Proposed Estimation Method

2.2.1 Principal Component Analysis on the p Predictors

Principal component analysis is performed to reduce the number of predictors. Principal Component analysis will be performed on the p predictors. A number of q principal components was selected according to the following guidelines (Giron, 2010) and (Joliffe, 2002):

For the cases where n=p and n<p using PMD-SPC:

1. Include the first q PCs wherein the cumulative percentage of the total variation contributed by these PCs is 40%.
2. Include the first q PCs wherein the cumulative percentage of the total variation contributed by these PCs is at least 60%

For the cases where n=p and n<p using GAS-PCA:

3. Include the first q PCs wherein the cumulative percentage of the total variation contributed by these PCs is 90%.
4. Include the first q PCs wherein the cumulative percentage of the total variation contributed by these PCs is at least 95%

2.2.2 Sparse Principal Component Analysis on the p Predictors

The algorithm proposed by Torres (2013) on the procedure of finding the inverse covariance matrix will be used for the GAS-PCA for high dimensional predictors. Since the covariance matrix for high dimensional predictors will tend to be singular due to collinearity, the inverse covariance matrix will be impossible to determine. The sparse inverse covariance estimate (SICE) method will be used to get an estimate of the inverse covariance matrix. (Torres, 2013).

The process of Witten et al. in computing the sparse principal components using Penalized Matrix Decomposition will be utilized as a process of dimension reduction similar to the GAS-PCA. Such method will be called PMD-SPC.

Since the sparse principal components are formed as a sparsely weighted linear combination of the observed variables $\boldsymbol{Z} = \mathbf{XB}$**.** The data can be approximately rotated back as $\widetilde{\boldsymbol{X}} = \mathbf{ZA^T}$ (Erichson et al., 2018)**.** These X's are now the reduced original predictors that will be used in the local scoring algorithm and backfitting algorithm. A more detailed procedure of the local scoring algorithm and backfitting algorithm is provided in Giron (2010) and Torres (2013).

2.3 Hypothesis Test on Significant predictors

Given the postulated model on Xs, we test the following hypothesis:

$$H_o^{(i)}: \beta_i = 0 \text{ vs } H_a^{(i)}: \beta_i \neq 0$$
$$\text{For } i = 1,2,\dots,p$$

where β_i are the estimated parameter values from the model.

For testing the above hypothesis, the p-values of the significant predictors will be computed. The p-values are computed using the F-test. Each term in the model are separated into two: projection part and the nonparametric part. This causes two ANOVA objects in which one is the linear component and the other is the nonparametric component. A type of score test is performed in each of the nonparametric terms. The nonparametric component is set to zero, and the linear part is updated, holding the other nonparametric terms fixed. This is done efficiently and simultaneously for all terms (Hastie, 1991).

2.3.1 False Discovery Rate Controlling Procedure

This follows from the false discovery rate procedure proposed by Benjamini and Hochberg (1995). A more detailed procedure can be found in Benjamini and Hochberg (1995). The procedure has the advantage of not assuming any parametric model on the data and also controls the false discovery rate which is common to high dimensional data.

2.3.2 Power and Size Computation

Power is computed as the number of times a significant predictor was considered as significant in 200 replicates for the n=p (NHD) cases and 100 replicates for the n<<p (HD) cases. Power is computed by getting the proportion of replicates wherein the significant predictors were considered significant or the number of times the null hypothesis was rejected in the assigned significant predictors. The design assigns the first 8 predictors for L_1, L_2 and L_3 latent factors and first 4 in L_4 and L_5 latent factors for NHD cases. Whereas for the HD cases, the first 8 predictors were assigned to be significant for all latent factors.

Size is computed as the number of times a nonsignificant predictor was considered as significant in 200 replicates for the NHD cases and 100 replicates for the HD cases. This is

computed by getting the proportion of replicates wherein the nonsignificant predictors were considered as significant. This is computed individually for each variable not assigned as significant i.e., all variables except the first 8 variables of each latent factor.

2.4 Design of Simulation Study

The data generation process involves five hidden factors that will generate 5 groups of strongly correlated predictors. The simulation study is similar to that of Zou and Hastie (2005), Lansangan and Barrios (2017) and Supranes and Lansangan (2017). The latent factors L_1, L_2, L_3, L_4 and L_5 represent the "important drivers" that predict the response variable. The latent factors were generated such that:

$$L_1 \sim N(50,40^2),\ L_2 \sim N(50,20^2),\ L_3 \sim N(50,10^2),\ L_4 \sim N(50,5^2),\ L_5 \sim N(50,5^2)$$

$$cor(L_1, L_2, L_4, L_5 | L_3) = \begin{bmatrix} 1 & 0 & 0.6 & 0 & 0 \\ 0 & 1 & -0.5 & 0 & 0 \\ 0.6 & -0.5 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Given those specifications, L_1 and L_2 are independent of each other, L_3 is moderately correlated to L_1 and L_2. On the other hand, L_4 and L_5 are independent on the other hidden factors.

Each Xij represents a measurable manifestation of a latent factor. Since $X_{ij}s$ are linear functions of a latent factor, variable generated from the same latent factor are strongly correlated with each other. Predictors from the first two latent factors (L_1 and L_2) are moderately correlated to predictors from L_3. Predictors from L_4 and L_5 are independent from the others.

Since the groups $L_1 - L_5$ create the total population, each group will have different variances. All groups have a mean of 50 which is necessary to avoid negative values and for the Cholesky decomposition to be performed.

Two different scenarios in terms of the number of variables are considered: NHD (n=p) case and HD (n<<p) case. Specifically, 200 samples with 200 predictors (NHD) and 200 samples with 2000 predictors (HD) for the dichotomous response and 210 samples with 210 predictors (NHD) and 210 samples with 2100 predictors (HD) for 3-category response. Tables 1 and 2 shows the simulated relative contribution of each group for each scenario. The relative contribution is the percentage of the number of predictors that came from each latent factor.

Table1. Scenario Settings for Dichotomous Response Simulation Study

Case	Relative Contribution by Group				
	Group 1	Group 2	Group 3	Group 4	Group 5
200 – 200	10% (20)	15% (30)	15% (30)	25% (50)	35% (70)
200 – 2000	1% (20)	1.5% (30)	1.5% (30)	48% (960)	48% (960)

Table 2. Scenario Settings for 3-category Response Simulation Study

Case	Relative Contribution by Group				
	Group 1	Group 2	Group 3	Group 4	Group 5
210 – 210	10% (21)	10% (21)	20% (42)	30% (63)	30% (63)
210 – 2100	1% (21)	1% (21)	2% (42)	48% (1008)	48% (1008)

For NHD scenarios, L_1, L_2 and L_3 will each have 8 significant predictors while L_4 and L_5 will each have 4 significant predictors. For HD scenarios, all 5 groups will have 8 significant predictors each. L_1 to L_3 will either follow U(8,12) or U(20,30) while L_4 and L_5 will either follow U(2,4) or U(5,10).

We then generate 5000 observations from each predictor and compute for $Y^* = \alpha + \sum_i f_i(X_i) + \epsilon$ where ϵ is from a Gumbel distribution. We use a Gumbel distribution for the error term as the generalized extreme value distribution which is common in discrete choice model. We arrange the data from smallest to largest value with respect to Y^*. For the dichotomous category, the following are the setting for each distribution:

Balanced (50-50) Let

$$Y^*_{(i)} = \begin{cases} 1, & for\ i = 1,2,\dots,2400 \\ 2, & for\ i = 2601, 2602, \dots, 5000 \end{cases}$$

Moderately Unbalanced (70-30) Let

$$Y^*_{(i)} = \begin{cases} 1, & for\ i = 1,2,\dots,3360 \\ 2, & for\ i = 3561, 3562, \dots, 5000 \end{cases}$$

Severely Unbalanced (10-90) Let

$$Y^*_{(i)} = \begin{cases} 1, & for\ i = 1,2,\dots,480 \\ 2, & for\ i = 681, 682, \dots, 5000 \end{cases}$$

We trimmed the middle 200 observations to put some threshold between the boundary of the two categories inducing the dichotomy. The population size for each distribution of the dichotomous category is 4800.

For the 3-category, the following are the settings for each distribution:

Balanced (33-33-34) Let

$$Y^*_{(i)} = \begin{cases} 1, & for\ i = 1,2,\dots,1500 \\ 2, & for\ i = 1751, 1752, \dots, 3250 \\ 3, & for\ i = 3501, 3502, \dots, 5000 \end{cases}$$

Moderately Unbalanced (60-20-20) Let

$$Y^*_{(i)} = \begin{cases} 1, & for\ i = 1,2,\dots,2700 \\ 2, & for\ i = 2951, 2952, \dots, 3850 \\ 3, & for\ i = 4101, 4102, \dots, 5000 \end{cases}$$

Severely Unbalanced (70-20-10) Let

$$Y^*_{(i)} = \begin{cases} 1, & for\ i = 1,2,\dots,3150 \\ 2, & for\ i = 3401, 3402, \dots, 4300 \\ 3, & for\ i = 4551, 4552, \dots, 5000 \end{cases}$$

We trimmed the middle 250 observations to put some threshold between the boundary of the 1st and 2nd categories and 250 trimmed observations between 2nd and 3rd category. The population size for each distribution of the dichotomous category is 4500.

Simulation scenarios are conducted for both PMD-SPC and GAS-PCA. Samples of dichotomous and 3-category response variables were simulated according to these scenarios. These scenarios are summarized in the table below.

Table 3. Summary of Simulation Scenarios

	Dichotomous Response	**3-category Response**
Transformation of Utility function values	Linear $Y^* = X_i\beta_i + \epsilon$ Exponential $Y^* = \exp(X_i\beta_i) + \epsilon$	Linear $Y^* = X_i\beta_i + \epsilon$ Exponential $Y^* = \exp(X_i\beta_i) + \epsilon$
Sample size (n) and number of predictors (p)	n(200) < p(2000), n(200)=p(200)	n(210) < p(2100), n(210)=p(210)
Distribution of the observations in the response category	(Cat 1, Cat 2) Balanced (50%, 50%) Moderately unbalanced (70%, 30%) Highly unbalanced (90%, 10%)	(Cat 1, Cat 2, Cat 3) Balanced (33%, 33%, 34%) Moderately unbalanced (60%, 20%, 20%) Highly unbalanced (70%, 20%, 10%)

2.5 Benchmark/ Baseline Test

To further evaluate the performance of the proposed test, a benchmark or baseline test was also implemented together with the proposed test using the same set of generated data. The baseline test that will be used Bonferroni Test for family wise error rate.

Some of the values that will also be considered are the following:

a. The number of times it was not removed by PMD-SPC and GAS-PCA.
b. The number of times each remaining predictor from the initial dimension reduction was considered as significant by the backfitting algorithm.
c. The number of times the Benjamini-Hochberg procedure rejected the significant predictors considered by the backfitting algorithm.

Power and size will be computed for each of the predictors simultaneously.

3. RESULTS AND DISCUSSION

The performance of the proposed test was evaluated by computing its size and power under different scenario settings. The results are organized into two sections based on SPCA method: PMD and GAS sections. For each SPCA subsections, results will be split into two: the truly significant predictors and the not truly significant predictors both covering the NHD and the HD cases.

The proposed method used two different transformations of the utility function: linear and exponential. The proposed method was not able to detect significant predictors when there is an exponential transformation. In all exponential transformation of the utility function, only the predictors from the L_1 were retained after the dimension reduction process. However, the backfitting algorithm only produced very few (min: 0; max: 4) significant predictors and were all eventually rejected by the BH procedure in the end. Thus, the paper will only report cases wherein the transformation of the utility function is linear.

The proposed method also was not able to detect any significant predictors coming from L_4 and L_5 groups. All significant predictors selected by the proposed method came from L_1 to L_3 group most of the time. The most consistent group with the highest number of predictors selected by PMD-SPC and GAS-PCA is L_1. In this light, the paper will report the individual power for each variable in the L_1 group.

The p-values computed came from the linear component of the ANOVA table. The p-values in the nonparametric part are always nonsignificant.

3.1 Estimated Size of the Test

In order to compute for the size, the not truly significant variables that were selected by the PMD-SPC and GAS-PCA are tested for significance. The variables in L_1 group will be reported since these variables are selected most of the time after dimension reduction. In the GAS-PCA, the case where the variance explained is 95% will only be reported since the case where the variance explained is 90% selected very few variables from the L_1 group (only 4 variables).

3.1.1 Estimated Size using PMD-SPC

In all cases using PMD-SPC, the proposed test is not always correctly sized. In fact, the baseline test is also not correctly sized at 0.05 level sometimes. The following tables provide a summary of the size for each variable.

Table 4. Estimated Size of the test for balanced, moderately unbalanced and severely unbalanced for dichotomous response for NHD and HD cases using PMD-SPC

	Balanced				**Moderately Unbalanced**				**Severely Unbalanced**			
	NHD		**HD**		**NHD**		**HD**		**NHD**		**HD**	
Va	**Prop**	**Base**	**Prop**	**Base**	**Prop**	**Base**	**Prop**	**Base**	**Prop**	**Base**	**Prop**	**Base**
X_9	0.042	0.041	0.025	0.010	0.099	0.078	0.102	0.073	0.412	0.32	0.34	0.276
X_{10}	0.044	0.06	0.031	0.012	0.084	0.060	0.092	0.067	0.342	0.256	0.361	0.287
X_{11}	0.053	0.049	0.039	0.025	0.128	0.094	0.112	0.064	0.431	0.325	0.421	0.282
X_{12}	0.050	0.054	0.047	0.016	0.101	0.084	0.115	0.079	0.375	0.287	0.363	0.277
X_{13}	0.047	0.039	0.021	0.015	0.127	0.102	0.106	0.067	0.378	0.306	0.341	0.252
X_{14}	0.062	0.052	0.036	0.016	0.124	0.076	0.089	0.051	0.370	0.267	0.381	0.312
X_{15}	0.055	0.067	0.033	0.022	0.118	0.092	0.104	0.065	0.391	0.304	0.394	0.297
X_{16}	0.046	0.054	0.042	0.028	0.110	0.092	0.103	0.065	0.392	0.324	0.378	0.282
X_{17}	0.056	0.047	0.073	0.032	0.142	0.111	0.143	0.098	0.400	0.290	0.365	0.284
X_{18}	0.052	0.043	0.029	0.016	0.125	0.080	0.135	0.087	0.409	0.319	0.353	0.296
X_{19}	0.073	0.067	0.051	0.026	0.136	0.124	0.155	0.095	0.426	0.290	0.453	0.348
X_{20}	0.047	0.039	0.088	0.043	0.124	0.102	0.133	0.095	0.422	0.316	0.392	0.317

Table 5. Estimated Size of the test for balanced, Moderately Unbalanced and Severely Unbalanced for 3-category response for NHD and HD cases using PMD-SPC

	Balanced				**Moderately Unbalanced**				**Severely Unbalanced**			
	NHD		**HD**		**NHD**		**HD**		**NHD**		**HD**	
Va	**Prop**	**Base**	**Prop**	**Base**	**Prop**	**Base**	**Prop**	**Base**	**Prop**	**Base**	**Prop**	**Base**
X_9	0.044	0.015	0.030	0.018	0.161	0.131	0.073	0.047	0.247	0.217	0.208	0.170
X_{10}	0.041	0.012	0.063	0.030	0.107	0.081	0.106	0.055	0.245	0.203	0.220	0.185
X_{11}	0.078	0.036	0.022	0.005	0.160	0.114	0.086	0.049	0.257	0.208	0.217	0.177
X_{12}	0.027	0.013	0.057	0.021	0.143	0.098	0.084	0.043	0.281	0.223	0.235	0.189
X_{13}	0.045	0.010	0.015	0.008	0.169	0.125	0.108	0.072	0.273	0.234	0.264	0.196
X_{14}	0.039	0.020	0.054	0.030	0.169	0.114	0.121	0.068	0.282	0.226	0.233	0.186
X_{15}	0.053	0.031	0.056	0.023	0.171	0.120	0.083	0.058	0.247	0.181	0.303	0.241
X_{16}	0.05	0.020	0.021	0.005	0.155	0.100	0.132	0.096	0.294	0.238	0.218	0.181
X_{17}	0.064	0.023	0.028	0.008	0.175	0.130	0.123	0.068	0.272	0.211	0.314	0.242
X_{18}	0.051	0.023	0.058	0.022	0.197	0.146	0.098	0.07	0.307	0.255	0.260	0.210
X_{19}	0.076	0.041	0.093	0.041	0.188	0.112	0.177	0.136	0.236	0.193	0.263	0.219
X_{20}	0.074	0.032	0.098	0.067	0.178	0.122	0.235	0.162	0.314	0.248	0.266	0.246
X_{21}	0.071	0.055	0.121	0.043	0.257	0.188	0.204	0.123	0.325	0.267	0.366	0.274

3.1.2 Estimated Size of Dichotomous response n=p using GAS-PCA

Based from the results, the proposed test is not always correctly sized. It can be seen that the in most cases, the baseline and the proposed method have the same size. Most of the cases also are incorrectly sized also. In the case of 3-category response where n=p, only 1 not truly significant variable was selected by the GAS-PCA at 95% variance explained. In all cases, the proposed test is incorrectly sized.

For the estimated power and size for dichotomous and 3-category response HD case using GAS-PCA**,** GAS-PCA failed to reduce the dimensionality of the predictors since HPCs available for use can only run a batch for 7 days but in that period, GAS-PCA failed to do so. Power and size cannot be computed for HD cases of GAS-PCA. This is recommended for further studies.

Table 6. Estimated Size of the test for balanced, Moderately Unbalanced and Severely Unbalanced for dichotomous response for n=p using GAS-PCA

	Balanced		**Moderately Unbalanced**		**Severely Unbalanced**	
Variable	**Proposed**	**Baseline**	**Proposed**	**Baseline**	**Proposed**	**Baseline**
X_9	0.098	0.073	0.179	0.119	0.407	0.264
X_{10}	0.125	0	0.071	0.071	0.368	0.316
X_{11}	0	0	0.167	0.167	0.556	0.444
X_{12}	0	0	0.167	0.167	0.667	0.667
X_{13}	0	0	0.167	0.167	0.556	0.556
X_{14}	0	0	0.333	0.167	0.444	0.444
X_{15}	0	0	0.333	0.333	0.444	0.444
X_{16}	0	0	0.167	0	0.889	0.556
X_{17}	0.167	0	0.167	0.167	0.667	0.556
X_{18}	0.167	0.167	0.167	0.167	0.444	0.333
X_{19}	0	0	0.167	0.167	0.444	0.444
X_{20}	0.333	0	0.333	0.333	0.333	0.222

Table 7. Estimated Size of the test for balanced, Moderately Unbalanced and Severely Unbalanced for 3-category response for n=p using GAS-PCA

	Balanced		**Moderately Unbalanced**		**Severely Unbalanced**	
Variable	**Proposed**	**Baseline**	**Proposed**	**Baseline**	**Proposed**	**Baseline**
X_9	0.152	0.043	0.143	0.06	0.238	0.142

3.2 Estimated Power of the Test

The proposed test is evaluated based on its power. To evaluate the power, different explained variances were considered. The distribution of the beta coefficients is also considered. The distribution that were taken into account were $\beta_i \sim U(8,12)$ and $\beta_i \sim U(20,30)$ The variance explained is at 40% and 60% for PMD-SPC and 90% and 95% for GAS-PCA. The power and size of each variable in L_1 group is reported since this group is the most consistent latent factor chosen by both PMD-SPC and GAS-PCA. The results for the L_2 and L_3 group can be seen in the full paper.

3.2.1 Estimated Power using PMD-SPC

For the estimated power using PMD-SPC, the test is powerful in detecting the significant predictors in all scenarios compared to the baseline test. It should be noted that as distribution

becomes more unbalanced, the power increases. Tables are provided for summary of the power for each variable.

Table 8. Estimated Power of the test for balanced, Moderately Unbalanced and Severely Unbalanced for dichotomous response for NHD and HD cases using PMD-SPC

	Balanced				Moderately Unbalanced				Severely Unbalanced			
	NHD		HD		NHD		HD		NHD		HD	
Va	Prop	Base	Prop	Base	Prop	Base	Prop	Base	Prop	Base	Prop	Base
X_1	0.644	0.554	0.700	0.605	0.591	0.433	0.637	0.549	0.803	0.760	0.718	0.655
X_2	0.435	0.281	0.488	0.315	0.433	0.318	0.445	0.273	0.703	0.640	0.580	0.435
X_3	0.286	0.144	0.276	0.130	0.307	0.200	0.363	0.230	0.594	0.514	0.428	0.318
X_4	0.217	0.123	0.208	0.110	0.250	0.170	0.268	0.160	0.509	0.416	0.315	0.235
X_5	0.169	0.074	0.160	0.060	0.232	0.156	0.228	0.108	0.479	0.409	0.310	0.210
X_6	0.133	0.066	0.148	0.078	0.198	0.141	0.173	0.105	0.457	0.367	0.290	0.213
X_7	0.135	0.071	0.108	0.055	0.186	0.121	0.205	0.118	0.428	0.320	0.243	0.178
X_8	0.123	0.060	0.118	0.070	0.163	0.115	0.173	0.108	0.443	0.341	0.233	0.183

Table 9. Estimated Power of the test for balanced, Moderately Unbalanced and Severely Unbalanced for 3-category response for NHD and HD cases using PMD-SPC

	Balanced				Moderately Unbalanced				Severely Unbalanced			
	NHD		HD		NHD		HD		NHD		HD	
Va	Prop	Base	Prop	Base	Prop	Base	Prop	Base	Prop	Base	Prop	Base
X_1	0.765	0.671	0.740	0.668	0.639	0.548	0.619	0.519	0.657	0.583	0.586	0.514
X_2	0.542	0.364	0.535	0.370	0.456	0.317	0.445	0.298	0.526	0.431	0.478	0.385
X_3	0.340	0.171	0.310	0.196	0.386	0.243	0.308	0.213	0.480	0.356	0.471	0.358
X_4	0.225	0.113	0.240	0.196	0.268	0.176	0.243	0.120	0.418	0.316	0.365	0.265
X_5	0.164	0.075	0.143	0.068	0.264	0.174	0.203	0.128	0.376	0.281	0.354	0.268
X_6	0.178	0.073	0.133	0.050	0.233	0.159	0.170	0.103	0.360	0.288	0.343	0.273
X_7	0.142	0.065	0.140	0.073	0.208	0.143	0.180	0.098	0.355	0.288	0.343	0.278
X_8	0.129	0.058	0.108	0.063	0.218	0.154	0.192	0.136	0.341	0.271	0.316	0.240

3.2.2 Estimated Power using GAS-PCA

For the dichotomous and 3-category response variable n=p, the test is powerful in detecting the significant predictors in all scenarios compared to the baseline test.

Table 10. Estimated Power of the test for balanced, Moderately Unbalanced and Severely Unbalanced for dichotomous and 3-category response for NHD case using GAS-PCA

	Balanced				Moderately Unbalanced				Severely Unbalanced			
	2-cat		3-cat		2-cat		3-cat		2-cat		3-cat	
Va	Prop	Base	Prop	Base	Prop	Base	Prop	Base	Prop	Base	Prop	Base
X_1	0.950	0.930	0.959	0.942	0.929	0.883	0.908	0.877	0.829	0.744	0.887	0.804
X_2	0.889	0.768	0.918	0.809	0.849	0.748	0.807	0.669	0.720	0.595	0.831	0.695
X_3	0.765	0.565	0.783	0.624	0.675	0.470	0.746	0.591	0.603	0.467	0.755	0.612
X_4	0.652	0.444	0.606	0.426	0.587	0.386	0.579	0.421	0.505	0.404	0.665	0.578
X_5	0.450	0.220	0.389	0.187	0.333	0.192	0.370	0.210	0.404	0.253	0.487	0.322
X_6	0.370	0.250	0.337	0.161	0.330	0.155	0.321	0.240	0.412	0.289	0.464	0.335
X_7	0.330	0.180	0.265	0.153	0.276	0.184	0.390	0.222	0.424	0.323	0.480	0.359
X_8	0.303	0.192	0.385	0.215	0.290	0.190	0.373	0.250	0.360	0.270	0.421	0.274

4. SUMMARY AND CONCLUSIONS

Variable selection is conducted for high dimensional discrete choice data. Previous studies have shown that by using PCA and SPCA, the discrete choice model of Abe (1999) could be extended for high dimensional data. The study resulted in high predictive ability on both PCA and SPCA. The proposed test focuses on the variable selection of high dimensional discrete choice data using PMD-SPC and GAS-PCA. Both methods utilize SPCA as a dimension reduction and variable selection tool. The selected variables are then subjected to test for significance and further tested for false discovery rate using the Benjamini-Hochberg Procedure. Power and size were computed and the average rejection of the BH procedure for the not truly significant predictors were reported.

The simulation showed that the proposed test is somehow correctly sized only in balanced cases. In all unbalanced cases, the proposed test is always incorrectly sized. Even though the test is incorrectly sized, the baseline test is also incorrectly sized and the proposed method showed a comparable size compared to the baseline test.

The result of the simulation shows that the proposed test is found to be more powerful than the baseline test across different parameters and number of variables involved.

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.

Aloulou, Foued (2018) The Application of Discrete Choice Models in Transport. Statistics-Growing Data Sets and Growing Demand for Statistics, Türkmen Göksel, IntechOpen, DOI: 10.5772/intechopen.74955.

Barrios, E. (2011) Bootstrap Methods. The Philippine Statistician Vol 60, pp. 129-132

Benjamini, Y., and Hochberg Y. (1995) Contolling False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of Royal Statistical Society. Series B, Vol. 57, No. 1, pp. 289-300.

Benjamini, Y., and Yekutieli, D. (2001) The Control of the False Discovery Rate in Multiple Testing Under Dependency. The Annals of Statistics. Series 2001, Vol. 29, No 4, pp 1165-1188.

Erichson, N. et al., (2018) Sparse Principal Component Analysis via Variable Projection https://arxiv.org/pdf/1804.00341.pdf (Retrieved: June 4, 2019)

Faraway, J. (2002) Practical Regression and Anova using R. https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf *(Retrieved: April 19, 2018)*

Giron, M. and Barrios, E. (2013) High Dimensional Nonparametric Discrete Choice Model. The Philippine Statistician, Vol. 62, No. 1, pp 59-75.

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

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

Jolliffe, I. T., (1972) Discarding variables in a principal component analysis. i: Artificial data. Journal of the Royal Statistical Society. Series C (Applied Statistics), 21(2):160 – 173

Kim, K. I. (2008). False discovery rate procedures for high-dimensional data Eindhoven: Technische Universiteit Eindhoven DOI: 10.6100/IR637929

Lansangan, J. (2013) Sparse Principal Component Regression. The Philippine Statistician, Vol. 62, No. 1, pp. 33-50.

Lansangan, J., Barrios, E. (2017) Simultaneous Dimension Reduction and Variable Selection in Modelling High Dimensional Data. Computational Statistics and Data Analytics, Volume 112, August 2017, 242-256

Leng, C. and Wang, H. (2008) On General Additive Sparse Principal Component Analysis. Journal of Computational and Graphical Statistics. Vol. 18, 201-205

Luce, R. and P. Suppes (1965). Preference, utility and subjective probability. In R. Luce, R. Bush, and E. Galanter (eds), Handbook of Mathematical Psychology, Vol. 3, Wiley, New York.

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

Newcombe, P., Connolly, S., Seaman, S., Richardson, S. Sharp., S. (2018) A two-step methid for Variable Selection in the Analysis of a Case Cohort Study. International Journal of Epidemiology, Vol. 47, Issue 2, pp 597-604.

Oliveira, J., Guimaraes, A., Fonseca, A., and da Rocha, J., (2015) A PCA and SPCA based procedure to variable selection in agriculture. Revista Brasileira de Computação Aplicada, Passo Fundo, Vol. 7, No. 1, pp. 30-41

Qi, Y., Xu, M., & Lafferty, J. (2014). Learning High-Dimensional Concave Utility Functions for Discrete Choice Models. http://www.cs.cmu.edu/~minx/docs/utility.pdf (retrieved June 3, 2019)

Shah, A. K. and Oppenheimer, D. M. (2008). Heuristics made easy: an effort- reduction framework. Psychological bulletin 134 207.

Supranes, M.V. and J.R. Lansangan (2017). Regression and Variable Selection via a Layered Elastic Net, The Philippine Statistician, Vol. 66, No. 2, pp. 15-31.

Torres, M. (2013) Sparse Nonparametric Discrete Choice Model for High Dimensional Data. http://nap.psa.gov.ph/ncs/12thncs/papers/INVITED/IPS-05%20Computational%20Statistics%20I/IPS-05_2%20Sparse%20Nonparametric%20Discrete%20Choice%20Model%20for%20High%20Dimensional%20Data.pdf (*Retrieved April 10, 2018)*

Train, K., (2001) Discrete Choice Methods with Simulation, New York, NY: Cambridge University Press

Wibowo, S., Kushari, B., Chalermpong, S., and Choocharukul, K. (2005) Performace of Bootsrap-Estimated Discrete Choice Models: A Case Study of Bangkok Transit System (BTS), Journal of the Eastern Asia Society for Transportation Studies, Vol. 6, pp. 1766-1774.

Witten, D., Hastie, T. and Tibshirani, R. (2009) Penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, Biostatistics 10(3): 515–534, doi:10.1093/biostatistics/kxp008

Zou, H., Hastie, T. and Tibshirani, R. (2006) Sparse principal component analysis, Journal of Computational and Graphical Statistics 15(2): 265-286.