INFERENCE ON HIGH DIMENSIONAL DISCRETE CHOICE MODEL

Danilo R. Si

Joseph Ryan G. Lansangan, Ph.D.

Erniel B. Barrios, Ph.D.

UP School of Statistics

Danilo R. Si

Ateneo de Manila Senior High School

14th NCS National Convention on Statistics
1-3 October 2019 | Crowne Plaza Manila Galleria
Organized by the Philippine Statistical System
Spearheaded by the Philippine Statistics Authority

Introduction

- Discrete choice models are presented as development and renovation of the classical choice theory. The model has been used extensively to discrete choice processes in different fields such as econometrics and transportation.
- Abe (1999) developed a methodology for discrete choice data using the generalized additive model (GAM).
- 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.
- 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.
- This study aims to conduct variable selection process on the explanatory variables using two methods of sparse principal component analysis. The

Methodology

- The methodology starts with the data generating process to create a high dimensional data for discrete choice model.

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

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, $x_1, x_2, \ldots, x_p$, measured on n subjects, with population covariance matrix S =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

Methodology

- The proposed estimation proposed procedure composed of the following:
  - Principal Component Analysis and Sparse Principal Component Analysis using
    - Penalized Matrix Decomposition Sparse Principal Component (PMD-SPC)
    - 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.

Simulation Study

- The data generation process involves five hidden factors that will generate 5 groups of strongly correlated predictors. 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:
- $L_1 \sim N(50, 40^2)$, $L_2 \sim N(50, 20^2)$, $L_3 \sim N(50, 10^2)$, $L_4 \sim N(50, 5^2)$, $L_5 \sim N(50, 5^2)$

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

- 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.

Simulation Study

- 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.
- 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.

Table1. Scenario Settings for Dichotomous Response Simulation Study

Case	Relative Contribution by Group				
	G 1	G 2	G 3	G 4	G 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				
	G 1	G 2	G 3	G 4	G 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)

Simulation Study

Balanced (50-50) Let

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

Moderately Unbalanced (70-30) Let

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

Severely Unbalanced (10-90) Let

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

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.

Balanced (33-33-34) Let

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

Moderately Unbalanced (60-20-20) Let

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

Severely Unbalanced (70-20-10) Let

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

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 3-category is 4500.

Simulation Study

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.

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

Results and Discussions

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.

Results and Discussions

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

Results and Discussions

- 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.
- In all cases using GAS-PCA, the proposed test is not always correctly sized. In most cases, the baseline and the proposed method have the same size. 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.

	Balanced			
	NHD		HD	
Va	Prop	Base	Prop	Base
X_9	0.042	0.041	0.025	0.010
X_{10}	0.044	0.06	0.031	0.012
X_{11}	0.053	0.049	0.039	0.025
X_{12}	0.050	0.054	0.047	0.016
X_{13}	0.047	0.039	0.021	0.015
X_{14}	0.062	0.052	0.036	0.016
X_{15}	0.055	0.067	0.033	0.022
X_{16}	0.046	0.054	0.042	0.028
X_{17}	0.056	0.047	0.073	0.032
X_{18}	0.052	0.043	0.029	0.016
X_{19}	0.073	0.067	0.051	0.026
X_{20}	0.047	0.039	0.088	0.043

Va	Prop	Base	Prop	Base
X_9	0.044	0.015	0.030	0.018
X_{10}	0.041	0.012	0.063	0.030
X_{11}	0.078	0.036	0.022	0.005
X_{12}	0.027	0.013	0.057	0.021
X_{13}	0.045	0.010	0.015	0.008
X_{14}	0.039	0.020	0.054	0.030
X_{15}	0.053	0.031	0.056	0.023
X_{16}	0.05	0.020	0.021	0.005
X_{17}	0.064	0.023	0.028	0.008
X_{18}	0.051	0.023	0.058	0.022
X_{19}	0.076	0.041	0.093	0.041
X_{20}	0.074	0.032	0.098	0.067
X_{21}	0.071	0.055	0.121	0.043

Results and Discussions

- 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.
- In all cases using GAS-PCA n=p, the test is powerful in detecting the significant predictors in all scenarios compared to the baseline test.

Summary and Conclusion

- 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.

14th NCS National Convention on Statistics
1-3 October 2019 | Crowne Plaza Manila Galleria
Organized by the Philippine Statistical System
Spearheaded by the Philippine Statistics Authority

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.

References

- 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

References

- 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.

INFERENCE ON HIGH DIMENSIONAL DISCRETE CHOICE MODEL

Danilo R. Si

Joseph Ryan G. Lansangan, Ph.D.

Erniel B. Barrios, Ph.D.

UP School of Statistics

Danilo R. Si

Mathematics Subject Area Coordinator

Ateneo de Manila Senior High School

dsi@Ateneo.edu

14th NCS National Convention on Statistics

1-3 October 2019 | Crowne Plaza Manila Galleria

Organized by the Philippine Statistical System

Spearheaded by the Philippine Statistics Authority