DISCRETE CHOICE GENERALIZED ADDITIVE MODELS USING HIGH DIMENSIONAL DATA

by

Maureen Dinna dela Cruz Giron

A thesis submitted in partial fulfillment of
the requirements for the degree of

MASTER OF SCIENCE (STATISTICS)

School of Statistics, University of the Philippines
Diliman, Quezon City

March 2010

ABSTRACT

The restriction of linearity in a multinomial logit model can impede the correct prediction of discrete choices. Specifying a nonparametric model can increase the chance of correctly predicting the outcome; however, as more predictors are introduced into the model, the variance increases, and if the number of predictors grows considerably, this increase in variance can be compensated only by an exponential increase in sample size. Exceedingly large sample sizes are rarely possible in practice due to inadequate time, manpower or financial resources. A generalized additive model (GAM) for high-dimensional data is proposed where the predictors undergo dimension reduction prior to modeling. Data are simulated with the dependent variable having two or three categories and the method is compared to the ordinary discrete choice model and the GAM using the original independent variables. For the dichotomous response, in the simple case where the sample size is sufficiently larger than the number of predictors, the proposed method is comparable to the ordinary discrete choice model when the principal components included in the model account for at least 80% of the variation in all the predictors. When the number of predictors substantially exceeds 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. For the three-category case, the model is more advantageous over the multinomial model in high dimensional cases ($p>>n$).