MULTILEVEL PRINCIPAL COMPONENT REGRESSION MODEL FOR HIGH DIMENSIONAL DATA: A SIMULATION STUDY

by

Ella Joyce S. Paragas, and Erniel B.Barrios, Ph.D.

ABSTRACT

Data mining has been rising rapidly. It happened after the booming of massive datasets in different field such as bioinformatics and e-commerce. The “large” data can be because of the number of variables, number of observations, or both (Kettenring, 2009). Modeling of high dimensional data is often confounded with multicollinearity and problem with interpretability of the fitted model. General Adaptive Sparse Principal Component Analysis (GAS-PCA) is used in reducing dimensionality that simultaneously induces sparsity. However, selection of few sparse components of the high dimensional predictors leads to specification bias. A random group level effect can help mitigate the bias in a model based on a few principal components. In this study, a two-level principal component regression model for high dimensional data was postulated. This study showed that GAS-PCA captured the structural dependencies of the data. It is showed that multilevel principal component regression model is best fitted to use when there are large number groups and when the variability of the group level effect is large.

Key words: *General adaptive sparse principal component, high dimensional data, multilevel model, principal component analysis, principal component regression*

1. Introduction

Data mining has been rising rapidly. It happened after the booming of massive datasets in different field such as bioinformatics and e-commerce. The “large” data can be because of the number of variables, number of observations, or both (Kettenring, 2009). Data are said to be high dimensional when the number of variables is greater than the number of observations.

One of the problems in high dimensional data is the curse of dimensionality. It refers to the complexity of making statistical inference as the data becomes multivariate (Banks, 2010). Multicollinearity among variables can also be a problem in this situation.

Principal component analysis (PCA) is used as reduction technique before performing other statistical techniques. PCA aims to lessen the dimensionality and solve the problem of multicollinearity of the data set with a large number of interrelated variables (Jolliffe, 2005).

Principal component regression (PCR) uses the principal components as the predictor variables in the regression model. PCA with sparsity constraint can be used to induce interpretability of the model. Leng and Wang (2009) proposed General Adaptive Sparse Principal Component Analysis (GAS-PCA). The method replaces the least- squares objective function in Sparse-PCA of Zou et. al. (2006) by a general objective function and uses adaptive lasso penalty in place of the lasso penalty in Sparse-PCA. Torres and Barrios (2013) employed GAS-PCA in a discrete choice model with high dimensional predictors. It turns out that the proposed procedure is better in terms of the interpretability of the predictors compared to the procedure using PCA.

Principal components regression resolves multicollinearity and curse of dimensionality but not the loss of information or specification bias (Umali and Barrios, 2014). This leads to the problem on predictive ability of the model. They addressed this problem by using nonparametric principal regression.

Multilevel model, which is also termed as hierarchical, nested or mixed model, has been popular in the biological and social sciences. This is due to the fact that humans belong to

different groups (Goldstein, 2011). As an example, students are nested in classrooms and classrooms are nested in schools. Thus when student performance is investigated, it is possible that other than study habits, classroom and school settings contribute to it.

In the Philippine setting, some reports on official statistics are at the provincial levels. However there is also a need to examine the region as a factor that can affect the statistics produced. This study then proposed to use the Multilevel model to account for the additional effect of higher level to the lower level statistics.

Suppose we want to model poverty incidence of all the provinces of the country, we can consider the aggregate data from the Philippine Statistics Authority on food security, population and employment which are related to poverty. The data are considered high dimensional if the number of variables is greater than the number of observations, which are provinces in this case. Principal Components Analysis can be applied to reduce dimension and then the extracted principal components will be used in the Principal Components Regression. However provinces belonging to the same region may have similarities. So we are interested in examining if the different regions can be potential sources of variability. Thus it is appropriate to use the multilevel model where provinces are the individual-level units and regions are the group-level units.

2. The Proposed Model

The study proposed the multilevel principal component regression model to mitigate the bias caused by using only the principal components in regression analysis.

In this study, only two levels in the multilevel were used. The higher level units are referred to as groups while the lower level units are the observations or units.

Given h group levels with n_i elements, $i = 1, \dots, h$, the postulated model is

$$Y_{ij} = \Psi_i + \sum_{d=1}^{D} \gamma Z_{ijd} + \varepsilon_{ij}$$

where:

Y_{ij} is the response variable of the j[th] unit within the i[th] group,
Ψ_i is the random effect in the i[th] group distributed as $N(\mu_i, \sigma_\Psi^2)$,
$\sum_{d=1}^{D} \gamma Z_{ijd}$ is sum of the 1[st] to the D[th] principal component of the j[th] unit within the i[th] group,
ε_{ij} is the random error distributed as $N(0, \sigma^2)$

Assumptions:

1. Y_{ij} could be discrete or continuous variable.
2. Z are orthogonal (uncorrelated) variables.
3. Ψ_i is the i[th] group random effect.
   This is the component of the model that addresses the effect of higher-level (or group level) of the observations. It is assumed that observations from the same group are homogeneous.
4. Groups are independent.
   The groups should be independent and mutually exclusive. Each observation should only belong to one group.
5. Group sizes may vary.
   The postulated model will be assumed to work for data with equal and unequal sizes of groups.

3. Estimation Procedure

Umali and Barrios (2014) proposed to mitigate the bias lost to selecting a few principal components only by allowing flexibility on the structure of the model, i.e. nonparametric link between the dependent variable and principal components was postulated.

In Multilevel Principal Component Regression Model, the parameters γ and Ψ of the model were estimated simultaneously using the Best Linear Unbiased Prediction (BLUP) since the model has an additive formulation. The Linear Mixed Effects (lme4) package in R was used since it is designed to fit a linear mixed model, a generalized linear mixed model or a nonlinear mixed model. We can fit a linear mixed effect models with fixed as well as random/nested effects to predict the response variable. The lme4 function uses Restricted Maximum Likelihood (REML) to estimate the variance components (which is preferred over the standard Maximum Likelihood).

4. Simulation Studies

A simulation study was conducted to evaluate the performance of the multilevel principal components regression model.

Table 1. Boundaries of Simulation Study

1) Number of groups	a) Small (*h=3*) b) Large (*h=10*)
2) Allocation per group	a) Equal b) Unequal
3) Group sizes	a) Equal number of groups *Small size (n_k=5)* *Large size (n_k=10)* b) Unequal number of groups *Small variation (randomly select from 2 to 10)* *Large variation (Randomly select from 4 to 20)*
4) Number of Structural Dependencies	a) Small structural dependencies (*s=2*) b) Large structural dependencies (*s=5*)
5) Number of PCs to Retain	a) d=1, d=2, d=3 (if *s=2*) b) d=4, d=5, d=6 (if *s=5*)
6) Group level effect	a) No variability (zero for all groups) b) Small Variability (increases by 500 for succeeding groups) c) Large Variability (increases by 2,500 for succeeding groups)
7) Misspecification *m* in the model	a) No misspecification (*m=1*) b) With misspecification (*m=5*)

The number of groups and the number of observations per group were considered in this study. It is of interest to know if the efficiency of the proposed model will be affected if there is an increase in the number of groups and number of observations. It has been examined whether the distribution (equal or unequal number of observations per group) has an impact to the adequacy of the model.

Different settings for generating high dimensional predictor variables were also explored. The number of predictor variables was fixed to 120 for all the settings. However, there are two of data generation (predictors), one with 2 structural dependencies and the other with 5 structural dependencies. This is because, in real life, variables are interrelated or do exhibit multicollinearity. Correlated predictors were generated in the following two scenarios:

a) Set 1: Small structural dependencies (s=2),
   $x_i = c_i x_1 + d_i + \varepsilon_i$ for $i = 2, \ldots, 60$

$x_i = c_i x_{61} + d_i + \varepsilon_i$ for $i = 62, \ldots, 120$

where:
c_i and d_i are constrants,
$x_1 \sim Normal(\mu, \sigma^2)$,
$x_{60} \sim Uniform(a, b)$ and
$\varepsilon_i \sim Normal(0,1)$.

b) Set 2: Large structural dependencies (s=5),
$x_i = c_i x_1 + d_i + \varepsilon_i$ for $i = 2, \ldots, 24$
$x_i = c_i x_{25} + d_i + \varepsilon_i$ for $i = 26, \ldots, 48$
$x_i = c_i x_{49} + d_i + \varepsilon_i$ for $i = 50, \ldots, 72$
$x_i = c_i x_{73} + d_i + \varepsilon_i$ for $i = 74, \ldots, 96$
$x_i = c_i x_{97} + d_i + \varepsilon_i$ for $i = 98, \ldots, 120$

where:
c_i and d_i are constrants,
$x_1, x_{49}, x_{73} \sim Normal(\mu_i, \sigma_i^2)$,
$x_{25}, x_{97} \sim Uniform(a_i, b_i)$
$\varepsilon_i \sim Normal(0,1)$

The dependent variable was computed as the sum of the predictor variable all the predictor variables, group level effect and error term which is distributed as $N(0, \sigma^2)$.

After the data generation, principal components (PC) were extracted. The number of PC extracted depends on the generated number of structural dependencies among the predictors. The first d PCs were obtained, as suggested in Jolliffe (2005). Then the PCs were used in the principal component regression.

Different group level effects were included in the simulation settings to assess the ability of higher level random effects to abate bias caused by using only the principal components in regression analysis.

For the misspecification error, a constant m was multiplied in the error term of Y where $\varepsilon \sim N(0, \sigma^2)$. The higher the value of m, the lower the predictive ability of the model.

There are 96 settings in total. For each setting, 100 sample replicates were generated.

5. Evaluation of the Model

The proposed procedure was evaluated through a simulation study. The selected principal components were used in the multilevel principal component regression model. Analysis using the proposed model was conducted for each replicate and the predicted response was obtained. The mean absolute percent error (MAPE) has been used to evaluate the model. The formula of MAPE is given below

$$MAPE = \frac{\sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|}{n} \times 100$$

To assess the goodness of fit, the rule of thumb for MAPE was utilized. The model is adequate when the MAPE is small, that is, atleast 20. The model will be rejected when MAPE is large.

Also, the nonzero loadings percentage (NZL) has been used in the evaluation of the principal components. NZL is the percent of nonzero entries over the total number of entries of the PCA loadings. Low NZL is a good indicator that the PCA loadings are sparse. On the other hand, an NZL close to 100 suggests a non-sparse PCA loadings.

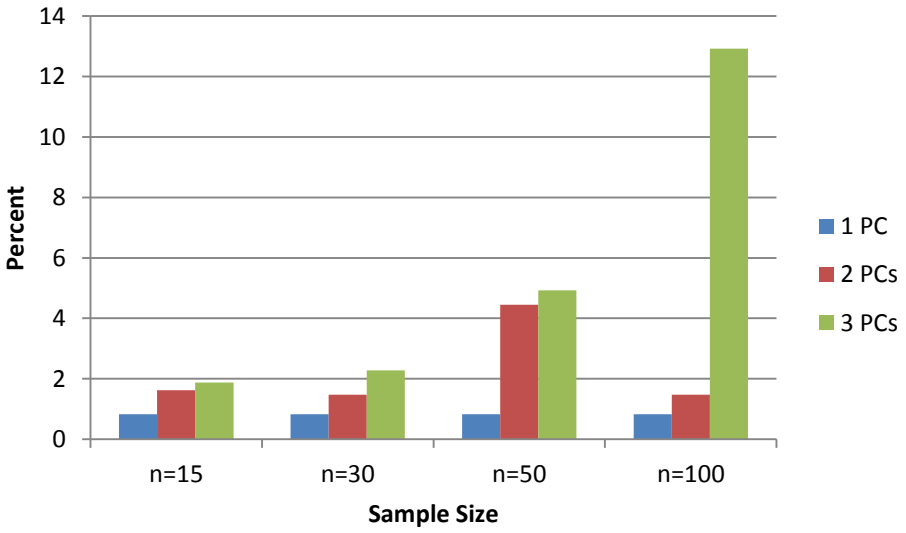
The fitted model from the multilevel principal component regression model was compared to the nonparametric principal component regression model of Umali and Barrios (2014). The mean MAPE for each setting was computed to determine the performance of the two models.

6. Discussion

General Adaptive Sparse Principal Component Analysis (GAS-PCA) proposed by Torres and Barrios (2013) was used to reduce the dimension of the data set. The number of principal components (PCs) extracted was varied depending on the structural dependencies among the generated predictor variables.

Sparsity and specification bias are two main considerations when deciding on how many principal components to retain. For Set 1 the loadings are sparse on the average, with atmost 5% nonzero loadings (NZL), except for the case where *n=100* and the selection of 3 PCs (See Figure 1).

Figure 1. Average Percent Nonzero Loadings for Data Set 1



However, the average percent explained variance (PEV) by the first PC has the largest contribution of atmost 29%. The 2nd PC also has additional PEV of almost 10%. The third PC has a very small contribution, 5% at the most. It can be deduced that the GAS-PCA captured the 2 structural dependencies on the generated predictors.

Figure 2. Average Percent Explained Variances for Data Set 1

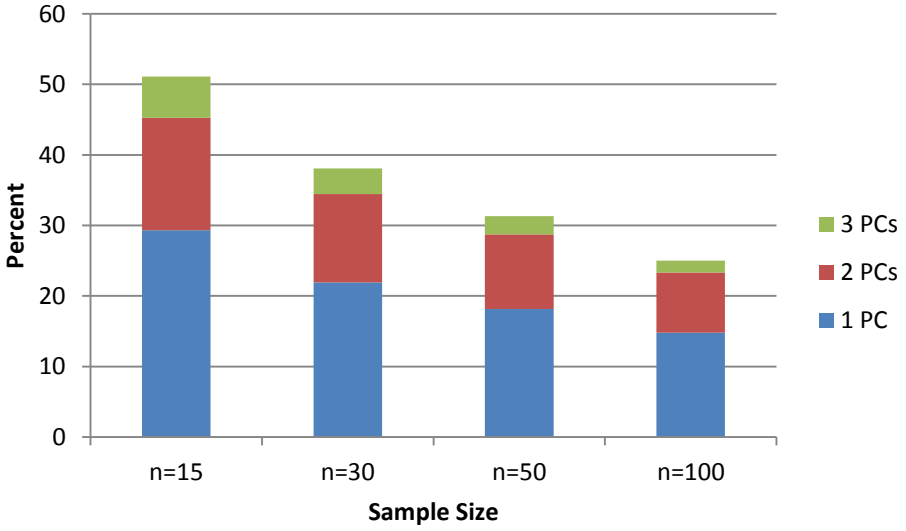


Figure 3. Average Percent Nonzero Loadings for Data Set 2

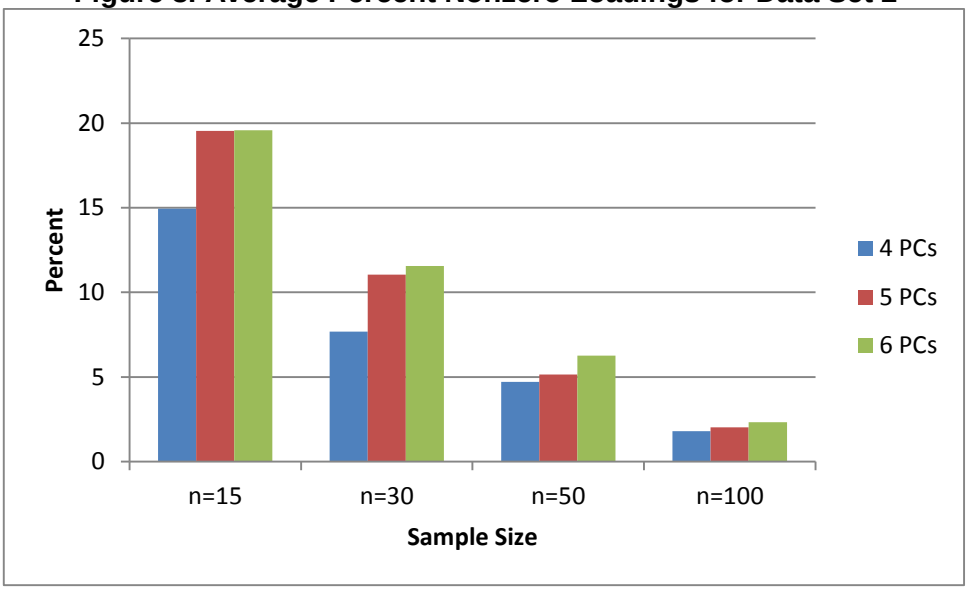
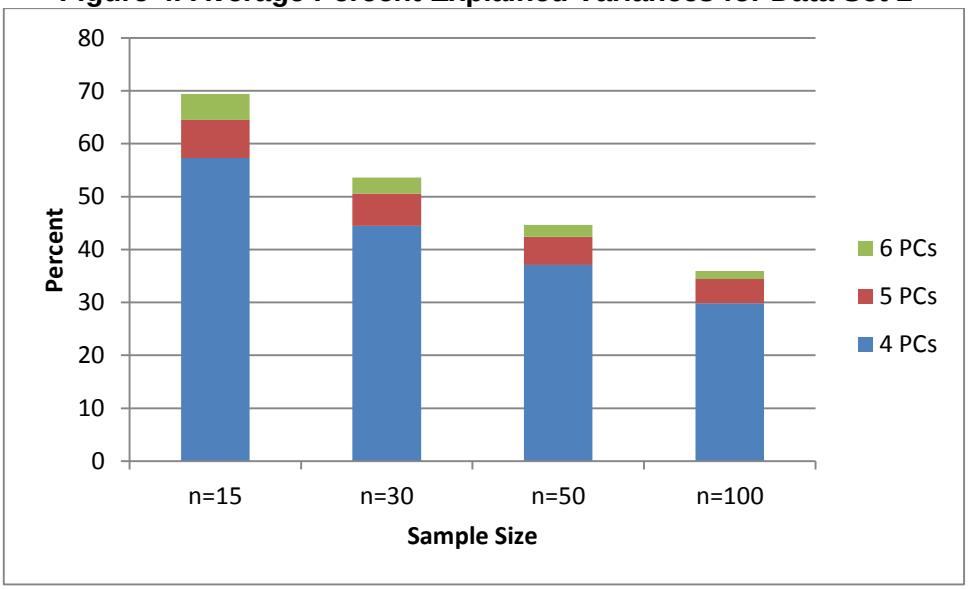
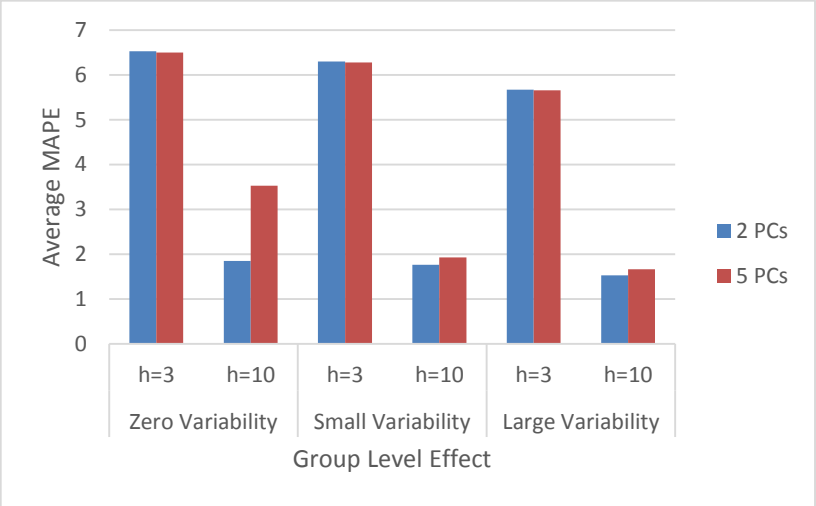


Figure 4. Average Percent Explained Variances for Data Set 2



It can be noticed from Figure 3 that as the sample size of Data Set 2 increases the GAS-PCA produce more sparse loadings. Moreover, as the sample size increases the average percentage variance explained by the principal components decreases (See Figure 4). The 5 dependency structure has been also captured by the GAS-PCA.

Figure 5. Average MAPE for Varying Number of Groups



In data sets with hierarchical structures, individuals can be categorized in few or many groups. Figure 5 shows that as the variability of the group level effect increases, the average MAPE decreases.

Table 3. Average MAPE for Varying Distribution of Observations per Group

Group Level Effect	Distribution	2 PCs		5 PCs	
		h=3	*h=10*	*h=3*	*h=10*
Zero Variability	**Equal**				
	5 Units	8.81	2.29	8.56	8.44
	10 Units	4.19	1.32	4.27	1.60
	Unequal				
	Small	8.85	2.48	8.76	2.49
	Large	4.28	1.32	4.40	1.60
Small Variability	**Equal**				
	5 Units	8.50	2.18	8.28	2.26
	10 Units	4.11	1.29	4.20	1.56
	Unequal				
	Small	8.45	2.34	8.35	2.36
	Large	4.16	1.28	4.28	1.54
Large Variability	**Equal**				
	5 Units	7.56	1.85	7.38	1.93
	10 Units	4.12	1.19	4.28	1.40
	Unequal				
	Small	7.22	1.93	7.10	1.96
	Large	3.76	1.16	3.87	1.37

The average MAPE values show that the performance of the multilevel principal component regression model is more affected by the number of observations per group than the allocation of observations per group. As the number of units per group increases, the MAPE decreases on the average (Table 2).

The misspecification error accounts for all possible sources of variation which is not explained in the model.

Table 3. Average MAPE for Varying Levels of Misspecification

Group Level Effect	Level of Misspecification	2 PCs	5 PCs
Large Variability	*m=1*	1.19	1.23
	m=5	6.01	6.10
Small Variability	*m=1*	1.29	1.33
	m=5	6.79	6.87
Zero Variability	*m=1*	1.34	1.38
	m=5	7.05	8.65

The estimated model may have lower predictive ability if it is misspecified. The results in Table 3 shows that the average MAPE for the model without misspecification is almost five times lower than that of with misspecification. This is true for both data sets.

7. CONCLUSIONS AND RECOMMENDATION

The multilevel principal component regression model was postulated in this study. General Adaptive Sparse Principal Component Analysis (GAS-PCA) was used as a dimension reduction method. The simulations show that the proposed model have addressed the issues of "curse of dimensionality" and multicollinearity among high dimensional predictors. In addition, the problems on interpretability and specification bias due to extraction of few principal components have been resolved by the model.

The multilevel principal component regression (MLPCR) model performs better when there is large number of groups. Also, the performance of the model improves as the number of units per group increases. However, the proposed model is robust to the manner of allocation of units per group.

The multilevel principal component regression model can be used when the researcher is interested in modelling high dimensional data and the observations are nested in different groups. This model can assure that the problems on dimensionality, interpretability and specification bias are treated.

References

Banks, D. L. (2010). Statistical data mining. Wiley Interdisciplinary Reviews: Computational Statistics, 2(1), 9-25.

Goldstein, H. (2011). Multilevel statistical models (Vol. 922). John Wiley & Sons.

Jolliffe, I. (2005). Principal component analysis. John Wiley & Sons, Ltd.

Kettenring, J. R. (2009). Massive datasets. Wiley Interdisciplinary Reviews: Computational Statistics, 1(1), 25-32.

Leng, C., & Wang, H. (2009). On general adaptive sparse principal component analysis. Journal of Computational and Graphical Statistics, 18 , 201-215.

Torres, M. G. T., & Barrios, E. B. (2013). Sparse Nonparametric Discrete Choice Model for High Dimensional Data. 12th National Convention on Statistics (NCS). EDSA Shangri-La Hotel, Mandaluyong City. October 1-2, 2013.

Umali, J., & Barrios, E. B. (2014). Nonparametric Principal Components Regression. Communications in Statistics-Simulation and Computation, 43:7, 1797-1810.

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