Partial Least Squares-Based Classification and Selection of Predictive Variables of Crimes against Properties in Nigeria

K. O. Olorede[1]; W. B. Yahya[2]; A. O. Garuba[3]; A. W. Banjoko[4]; K. A. Dauda[5]

[1,3,5]Department of Mathematical Sciences,
Kwara State University,
Malete, Nigeria.
e-mail: kabir.olorede@kwasu.edu.ng[1]; Anabelantonnia@gmail.com[3]

[2,4]Department of Statistics,
University of Ilorin,
Ilorin, Nigeria.
e-mail: wbyahya@unilorin.edu.ng[2]

***Abstract*--- In this study, the state-of-the-art Partial Least Squares (PLS) based models (PLS-Discriminant analysis (PLS-DA), Sparse PLS-DA (SPLS-DA) and Sparse Generalized PLS (SGPLS)) were employed to model and classify the rate of crimes (low or high) committed against properties across the 36 states in Nigeria and the Federal Capital Territory (FCT). The core variables that are predictive of this crime type in Nigeria were identified using the LASSO penalty method via the PLS. Data on occurrences of cases of offences against property obtained from the data base of Nigerian Police Force were utilized in this study. The missing values due to non-occurrence or non-reportage of crime cases were imputed, using the techniques of multivariate imputation by chained equation. The complete data set were partitioned into training and test sets using 80:20 holdout scheme. The 80% training set was used to build the PLS-based models that were in turn used to predict the overall crime rates of Nigerian cities in the 20% held out test data over 200 Monte-Carlo cross-validation runs. All the PLS-based models yielded good classification of unseen test samples into either of two qualitative classes of high and low crime rates with average Correct Classification Rate (CCR) of 94%. Other performance metrics including sensitivity, specificity, positive and negative predictive values, balance accuracy and diagnostic odds ratio were estimated to further examine their classification efficiencies. The SGPLS identified fewer (just 3 out of 12) core relevant crime variables that are predictive of the overall crime rates in Nigerian states with highest CCR than the SPLS which selected 9 such variables to achieved about the same feat.**

Keywords: *Sparse Partial Least Squares, Partial Least Squares, Dimension Reduction, Correct Classification Rate, LASSO, Training Set, Test Set.*

I. Introduction

Because we are certainly living in a time when moving is second nature to us and we move because living in one place for the rest of our lives is not a sentence we wish to serve, it is important for people to be able to correctly determine safety level of a city, based on the actual figure of crimes committed and identify the core factors that are responsible for such crime at any given time.

This study seeks to determine which of the three partial least squares (PLS) based models which include the *Partial Least Squares Discriminant Analysis* (PLS-DA) [1-3,11], the recently proposed *Sparse Partial Least Squares Discriminant Analysis* (SPLS-DA) and *Sparse Generalized Partial Least Squares* (SGPLS) [5,6] classification methods is best at identifying and selecting the core variables (factors) that are predictive of crimes against properties in Nigeria as well as classifying the 36 states in Nigerian including the Federal Capital Territory (FCT) according to the rate of crimes (low or high) committed against properties in such sates based on the identified factors.

II. Materials and Methods

A. *Data Description and Preparation*

This work utilizes a low-dimensional $(n > p)$univariate (with 1 response variable) binary response (Low Crime Rate (Low)$= 1$, High Crime Rate (High) $= 0$) dataset with $p = 12$ crime-related variables (relating to offences against properties) as predictors of crime rates (safety level) in the $n = 37$ states in Nigerian including FCT was collected from the Nigeria Police Force (NPF), Police Head Quarters,

Ilorin, Kwara State, Nigeria for the year 2013. The class labels High = 0 and Low = 1 implies that study target is low crime rate since high crime rate in a given state of Nigeria is an indication of low safety level in such a state and vice versa. Table 1 further summarizes the data.

Table 1: Data Summary

Description	Response Category		Features
	High	Low	
Class Distribution	30	7	12
Class Label	0	1	
Sample Fraction	30/37	7/37	
Sample Distribution	Unbalanced		

With the assumption of missing completely at random (MCAR), we carefully imputed missing values for the data set using predictive mean matching methods in the multivariate imputation by chained equation algorithm developed in the mice library [5] of the R 3.4.2 statistical software (R Core Team, 2017).

To classify a state into either of the two classes of high or low crime rate, the ground (overall) median number of cases of all crimes committed against properties in the data was computed. This was used as the threshold median value against which the estimated median value of each state of Nigeria was compared. A state with median number of cases of crimes against properties above the ground median number of cases in the country is classified as having high crime rate (low safety level) with its dummy variable coded 0, while a state with median number of cases of such crimes below the ground median is classified as having low crime rate (high safety level) with its dummy coded 1. The summary of the observed crime rate status of all the 36 states in Nigeria including FCT based on the above classification is presented in Table 1.

The data set were partitioned into training and test sets using 80:20 holdout scheme. The 80% (30 samples) training set was used to train the PLS-based models using the `'plsgenomics'` [3] and `'spls'` [7] libraries of the R software. The trained PLS-based classification models were in turn used to predict the overall crime rates of Nigerian cities in the 20% (7 samples) held out test data over 200 Monte Carlo cross-validation runs.

B. *Brief Overview of the Standard PLS Classifier*

After centering the response $(Y_{n\times q})$ and the predictor matrix $(X_{n\times p})$, PLS regression assumes latent components $(T_{n\times K})$ underlying both Y and X, using the PLS model given by

$$\boldsymbol{Y} = \boldsymbol{TQ}^T + \boldsymbol{F} \quad (1)$$

and

$$\boldsymbol{X} = \boldsymbol{TP}^T + \boldsymbol{E} \quad (2)$$

where $\boldsymbol{P}_{p\times K}$ and $\boldsymbol{Q}_{q\times k}$ are coefficients $(loadings)$ and $\boldsymbol{E}_{n\times P}$ and $\boldsymbol{F}_{n\times q}$ are errors. The latent components T are defined as $\boldsymbol{T} = \boldsymbol{XW}$, where $\boldsymbol{W}_{p\times k}$ are k direction vectors $(1 \le k \le min\{n,p\})$. The main PLS machinery involves finding direction vectors. The k^{th} direction vector $\widehat{W}_k$ is the solution of the optimization problem in equation (3).

Objective function: $\max_W \boldsymbol{W}^T\boldsymbol{MW}$ (3)

subject to: $\boldsymbol{W}^T\boldsymbol{W} = 1$ and $\boldsymbol{W}^T\boldsymbol{S}_{xx}\widehat{\boldsymbol{W}}_l = 0, l = 1,\cdots,k-1$

where $\boldsymbol{M} = \boldsymbol{X}^T\boldsymbol{YY}^T\boldsymbol{X}$ and $\boldsymbol{S}_{xx}$ represents the sample covariance matrix of the predictors (Frank and Friedman[8]). For our univariate PLS, the objective function in equation 3 can be interpreted [8] as:

$$\max_W \boldsymbol{cor}^2(Y, Xw)var(Xw) (4).$$

However, SPLS [5] incorporates variable selection into the standard PLS by solving the following minimization problem, instead of the original PLS formulation in (1) through (3). The objective function:

$$\min_{w,k} -k\boldsymbol{w}^T\boldsymbol{Mw} + (1-k)(c-\boldsymbol{w})^T\boldsymbol{M}(c-\boldsymbol{w}) + \lambda_1\|c\|_1 + \lambda_2\|c\|_2 \quad (5)$$

Subject to $\boldsymbol{w}^T\boldsymbol{w} = 1$, where $\boldsymbol{M} = \boldsymbol{X}^T\boldsymbol{YY}^T\boldsymbol{X}$.

The formulation (5) promotes exact zero property by imposing L_1 penalty onto a surrogate of direction vector (c) instead of the original direction vector $(\boldsymbol{w})$, while keeping $\boldsymbol{w}$ and c close to each other. Here, L_2 penalty takes care of the potential singularity of the $\boldsymbol{M}$ matrix.

For the univariate PLS, the solution of the formulation in (5) results in the soft threshold direction vector of the form:

$$\hat{c} = (|Z| - \lambda_1/2)_+ sign(z),$$

where $\boldsymbol{Z} = \boldsymbol{X}^T\boldsymbol{Y}/\|\boldsymbol{X}^T\boldsymbol{Y}\|$ *and* $(x)_+ = max(0,x)$. Chun and Keles [5] recast this soft thresholding as

$$\hat{c} = \left(|Z| - \eta \max_{i\le j\le p}|Z_j|\right)_+ sign(z),$$

where $0 \le \eta \le 1$ and justify setting $0 < k < 0.5$ and $\lambda_2 = \infty$. Thus, there are two key tuning parameters η and k in this formulation. Controlling η instead of the direction vector specific parameters λ_k, $k = 1,\cdots,K$, avoids combinatorial tuning of the set of sparsity parameters and provides a bounded range for the sparsity parameter, i.e., $0 \le \eta \le 1$. whenever $\eta = 0$, the SPLS reduces to PLS[6].

III. ANALYSIS

A. *Parameter Tuning*

Following standard literature procedures, we tuned optimal parameters for the standard two-stage PLS-DA which incorporates only dimension reduction for classification of

qualitative response and its sparse variants (SPLS and SGPLS) which incorporate simultaneous dimension reduction and variable selection for classification of qualitative response using 5 and 10-fold cross-validations. With the optimal parameters, we trained and tested prediction models for binary crime rates in Nigerian states using the PLS models.

B. *Model Assessment Criteria*

All the three methods employed in this work were trained on the training set and tested for classification and prediction performances on the test set. These methods were assessed for the classification of Nigerian states into binary classes of low crime safety and high crime safety statuses, using Correct Classification Rate (CCR), Sensitivity (SEN), Specificity (SPEC), Precision (PREC), Positive Predictive Value (PPV), Negative Predictive Value (NPV), Balance Accuracy (BA) and Diagnostic Odds Ratio (DOR) in a binary confusion matrix (see a sketch as provided by Table 2) over 200 cross-validation runs.

Table 2: Binary Confusion Matrix

Model Prediction	Actual Crime Status	
	High	Low
High	TP	FP
Low	FN	TN

TP: True Positive; **TN:** True Negative; **FP:** False Positive and **FN:** False Negative

The performance metrics used to assess performances of the classification methods are computed as presented in the following seven equations:

$$\boldsymbol{CCR} = \frac{TP + TN}{TN + TP + FN + FP} \quad (6)$$

$$\boldsymbol{SEN} = \frac{TP}{TP + FN} \quad (7)$$

$$\boldsymbol{SPEC} = \frac{TN}{TN + FP} \quad (8)$$

$$\boldsymbol{PPV} = \frac{TP}{TN + TN} \quad (9)$$

$$\boldsymbol{NPV} = \frac{TN}{TN + TP} \quad (10)$$

$$\boldsymbol{BA} = (SEN + SPEC)/2 \quad (11)$$

$$\boldsymbol{DOR} = \frac{TP \times TN}{FP * FN} \quad (12)$$

IV. RESULTS

We tuned parameters using 5 and 10 fold cross-validations for the PLS-DA, SPLS-DA and SGPLS models on the entire data set via $pls.lda.cv()$, $cv.spls()$ and $cv.sgpls()$ functions in the 'plsgenomics' and 'spls' libraries of the R software.

Table 3: Optimal Parameter and Tuning Complexity

Models	Tuning Parameters	Opt. Parameters	CV-MSEP	Tune Time
SPLS	Eta: seq(0.1,0.9,0.1)	0.9	0.025	6.83 Sec.
	K: seq(1,5,1)	3.0		
SGPLS	Eta: seq(0.1,0.9,0.1)	0.9	0.05	162.25 Sec.
	K: seq(1,5,1)	1.0		
PLS	K: seq(1,5,1)	9.0		0.77 Sec.

K: Number of PLS Components; **Eta**(η): SPLS Soft Thresholding Parameter; **Opt.:** Optimal
CV-MSEP: Crossvalidation Mean Square Error of Prediction

It is evident in Table 3 that PLS-DA utilized highest number of optimal parameters (components) with least computational complexity.

Table 4: Classification Accuracy

Models	Training Result		Test Result	
	CCR (%)	Miss	CCR (%)	Miss
SPLS	97.23	1	94.10	0
SGPLS	94.16	2	93.25	0
PLS	99.44	0	92.54	0

Miss: Number Misclassified

In terms of classification performance (Table 4), the three PLS-based models compete favourably in terms of test sample classification accuracy (CCR ≅ 94% for SPLS and CCR ≅ 93% for SGPLS & PLS). The performances of the both the PLS-DA and SPLS methods on the training data obviously showed an evidence of over-fitting with CCR ≅ 99% achieved by PLS-DA and CCR ≅ 97% by SPLS.

Table 5: Class-Specific Classification Accuracy

Models	Crime Status	Training	Test
		CCR (%)	CCR (%)
SPLS	High	99.50	96.95
	Low	96.94	94.02
SGPLS	High	99.50	98.80
	Low	93.68	92.91
PLS	High	99.50	84.83
	Low	99.43	93.79

Per-class CCR in Table 5 further supports the classification results in Table 4 with the PLS-DA evidently competed favourably with its sparse variants (SGPLS and SPLS). Among the 30 states with high crime rates (from Table 1), the PLS-DA was able to correctly classify about 85% of

them in the test data while both the SPLS and SGPLS correctly classified about 97% and 99% of such states respectively as evident from Table 5.

In terms of ability to capture target class (low crime rate) in the held out sample, SGPLS is best with sensitivity of 99.50%. Next to it is the SPLS-DA (SENS = 99.15%).

Table 6: Model Sensitivity, Specificity and Balance Accuracy.

Models	Training Result			Test Result		
	SEN (%)	SPEC (%)	BA (%)	SEN (%)	SPEC (%)	BA (%)
SPLS	83.69	99.50	91.60	99.15	62.69	80.69
SGPLS	61.00	99.50	80.25	99.50	55.51	77.37
PLS	99.10	99.50	99.30	97.66	60.51	79.05

The SPLS-DA is most specific (62.69%) in capturing high city crime rates. In overall, both PLS-DA and SPLS-DA are most balance in their true positive and true negative detections with BA of 79.05% and 80.69% respectively.

Table 7: Model Predictive Values and Effectiveness

Models	Training Result			Test Result		
	PPV (%)	NPV (%)	DOR (%)	PPV (%)	NPV (%)	DOR
SPLS	99.50	96.94	17.20	98.99	96.95	5.51
SGPLS	99.50	93.68	03.07	99.98	98.80	4.12
PLS	99.50	99.43	38.67	98.91	88.00	1.57

In terms of discriminatory effectiveness of the models (DOR): ratio of the odds of crime being classified as high in Nigerian states if the state has high crime rate, relative to the odds of the state being classified as having high crime rate if the state does not have high crime rate, the SPLS and SGPLS are the most effective among the three PLS-based models with DOR = 5.51 and DOR = 4.12 respectively.

Based on unseen test set classification, SGPLS was the most predictive (PPV = 99.98% and NPV = 98.80%) of binary response group of the crime rates types. Next to it was SPLS-DA (90.74% PPV and 96.95% NPV). This may be traceable to their feature selection via LASSO-based shrinkage penalty [10] since some of the original features may be noisy.

Table 8: Computational Complexity and Feature Selection

PLS Models	Training Time	Number of Features Selected
SPLS	0.03 Sec.	07
SGPLS	0.02 Sec.	03
PLS	0.02 Sec.	12

Over 200 iterations (table 8), all 3 PLS models utilized less than 5 seconds training time on a 4GB RAM 64 bits windows operating system. SGPLS selected only 3 out of 12 crime variables as core predictive variables of offences against property in Nigerian cities. SPLS-DA selected additional 4 and PLS-DA utilized all variables.

Table 9: Selected Features and Coefficients

SPLS-DA		SGPLS	
Crime Variable	Coefficient	Crime Variable	Coefficient
Theft	0.4167	Theft	0.1631
Receiving Stolen Ppt.	1.3546	Receiving Stolen Ppt.	0.1705
Other Offences	0.8643	Other Offenses	0.1535
Armed Robbery	0.3528	...	...
House-Breaking	-0.3544	...	...
False Pretense	-0.6828	...	...
Unlawful Possession	0.2856	...	...

Based on results in Table 9, most relevant predictive variables for rate of crimes against property in Nigerian states as selected by both SGPLS and SPLS-DA include: Theft, Receiving Stolen Properties, and Other Offences. Another four variables including *House-Breaking, False Pretence, Armed Robbery,* and *Unlawful Possessions* may be informative if carefully studied as they were also detected as predictive of the crime rate against properties by the SPLS-DA method.

V. Conclusion

This work has demonstrated partial least squares dimension reduction and classification methods as efficient methods for modelling crime data and selecting core predictive variables for judging safety level of Nigerian cities.

From the foregoing, it may be reasonable to conclude that all PLS-based methods are efficient but good choice of them needs to be centred on whether only classification, feature selection, or both are intended.

Most relevant predictive variables for rate of crimes against property in Nigerian cities include: Theft, Receiving Stolen Properties, and Other Offences.

Acknowledgment

The authors appreciate Major Abubakar Hassan of Nigeria Army for his efforts at getting the data sets used for this work from the Nigeria Police Force headquarters, Ilorin.

References

[1] Boulesteix, A.-L. (2004). PLS dimension reduction for classification with microarray data, StatisticalApplications in Genetics and Molecular Biology, 3, Article 33.

[2] Boulesteix, A.-L. and Strimmer, K. (2006). Partial least squares: a versatile tool for the analysis ofhigh-

dimensional genomic data. *Brief. Bioinformatics*, Vol. 7, 32-44.

[4] Boulesteix A-L., Durif G., Lambert-Lacroix S., Peyre J. and Strimmer K. (2017). plsgenomics: PLS Analyses for Genomics. R package version 1.5-1.https://CRAN.R-project.org/package=plsgenomics

[3] Buuren S.F., Groothuis-Oudshoorn K. (2011). mice:Multivariate Imputation by ChainedEquations in R. *Journal of Statistical Software*, 45(3), 1-67. URL: http://www.jstatsoft.org/v45/i03/.

[5] Chung, D., and Keles, S., (2010a): Sparse Partial Least Squares Classification for High Dimensional Data. *Statistical Applications in Genetics and Molecular Biology*. Volume 9, Issue 1, 1544-6115

[6] Chun, H. and Keles, S. (2010b): Sparse partial least squares for simultaneous dimension reduction and variable selection, Journal of Royal Statistical Society, Series B, (http://www.stat.wisc.edu/~keles/Papers/SPLS_Nov07.pdf).

[7] Chung D., Chun H. and Keles S. (2013). `spls`: Sparse Partial Least Squares (SPLS) Regression andClassification. R package version 2.2-1. https://CRAN.R-project.org/package=spls

[8] Frank, I. E. and Friedman, J.H. (1993), A Statistical View of Some chemometrics Regression Tools, *Technometrics*, 35, 109-135

[9] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[10] Tibshirani R., (1996): Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*. Series B (Methodological), Vol. 58, No. 1 (1996), pp. 267-288

[11] Wold, H. (1966). *Estimation of Principal Components and Related Models by Iterative Least Squares*. New York: Academic Press