IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 15, 2019, accepted October 8, 2019, date of publication October 21, 2019, date of current version November 1, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2948782

Filter-Based Factor Selection Methods in Partial Least Squares Regression

TAHIR MEHMOOD[1], MARYAM SADIQ[2,3], AND MUHAMMAD ASLAM[3]

[1]School of Natural Sciences (SNS), National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan
[2]Department of Statistics, University of Azad Jammu and Kashmir, Muzaffarabad 13100, Pakistan
[3]Department of Mathematics and Statistics, Riphah International University, Islamabad 45210, Pakistan

Corresponding author: Maryam Sadiq (hussainulahmad@gmail.com)

ABSTRACT Factor discovery of high-dimensional data is a crucial problem and extremely challenging from a scientific viewpoint with enormous applications in research studies. In this study, the main focus is to introduce the improved subset factor selection method and hence, 9 subset selection methods for partial least squares regression (PLSR) based on filter factor subset selection approach are proposed. Existing and proposed methods are compared in terms of accuracy, sensitivity, F1 score and number of selected factors over the simulated data set. Further, these methods are practiced on a real data set of nutritional status of children obtained from Pakistan Demographic and Health Survey (PDHS) by addressing performance using a Monte Carlo algorithm. The optimal method is implemented to assess the important factors of nutritional status of children. Dispersion importance (DIMP) factor selection index for PLSR is observed to be a more efficient method regarding accuracy and number of selected factors. The recommended factors contain key information for the nutritional status of children and could be useful in related research.

INDEX TERMS Factor selection, filter, partial least squares, regression.

I. INTRODUCTION

It is challenging to make sense out of the high-dimensional data, particularly in the case of multicollinearity. Robust, efficient and scaleable analysis of collinear data is the need of the hour as most fields like public health, chemometrics, machine learning etc. deal with noisy and massive data. Consider the general regression equation of the form; $y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon_n$, the problem is to estimate an unknown subset of influential explanatory factors $(p_1, p_2, \ldots, p_j)$ on response y from n explanatory factors. Many new methods/techniques have been developed to extract subset of influential explanatory factors for better understanding of specific mechanisms and to attempt solve particular public health problems for linear regression. While predictive models could be based on numerous factors, several reasons motivate the search for subset selection of influential factors. First, limiting the number of factors often help to reduce over-fitting when we deal high dimensional data with many factors (p)/few samples(n) and thus, can lead to better predictions in terms of statistical aspect. Second, from a health viewpoint, selecting and inspecting the influential factors may shed light on health processes and suggest novel targets for prevention.

Child malnutrition is one of the key public health issue in developing countries like Pakistan. The time period between ab-lactation and the age of five is nutritionally considered as the most valuable interval of maturation regarding physical growth, mental power and immune system [1]. Weight-for-age, a composite index of height-for-age and weight-for-height, is considered as a primary indicator of nutritional status of children. Weight-for-age identifies acute (wasting) and chronic (stunting) malnutrition. According to PDHS 2012-13, 45 % of children under age 5 years are estimated as stunted, 11 % are wasted, and 30 % are underweight in Pakistan [2]. Dietary patterns, socio-economic status, maternal factors, demographic characteristics and environmental health conditions are reported as significant risk factors of nutritional status among children [3], [4]. Since nutritional status greatly varies among regional disparities, investigating its causative factors within this context is essential to prioritize development interventions to mitigate malnutrition.

In recent decades, partial least squares regression (PLSR) has been widely applicable for analyzing high dimensional, collinear multivariate data due to its versatility. PLSR being

The associate editor coordinating the review of this manuscript and approving it for publication was Yuhao Liu.

a supervised method is specifically proven to administrate prediction perspective. Although, the main focus of PLSR is to determine the subspace of relevant predictor factors and it has no execution of selecting influential factors in its basic algorithm, but several factor selection methods in PLSR have been proposed so far [5]. Partial least squares or projection to latent structures is a method based on projection which neglects directions in the factor space spanned by irrelevant, noisy factors. PLS estimator has the build-in property of up and down weighting of factors and hence, selection of influential factors may be unnecessary [6]. In case of very large p and very small n, PLSR estimators have asymptotic inconsistency for univariate outcome [7] and too many factors may cause large variation for prediction purpose [8]. These flaws highlight the necessity to specify the correct size of the relevant subspace in the p-dimensional factor space for a large number of factors [9] and motivate to discover improved factor selection procedures [10]. Selection of influential factors may upgrade the accuracy of model however some effective redundant factors may also be discarded simultaneously. Uniformity of selected factor is also an essential aspect [11] since using few factors for prediction also indicate that we are assigning large contribution of each factor in finally selected model [12]. For improved understanding and interpretation of the model for the target response, identification of important factors is essential.

In PLSR, factor selection (FS) methods can be classified into three main categories namely filter methods, wrapper methods, and embedded methods. Filter methods purely operate for identification of subset of influential factors using PLSR output. Wrapper methods are iterative procedures which re-fit the PLSR for the factors identified by filter methods. Embedded methods incorporate factor selection as a part of PLSR. We only focus here filter methods which can be generalized for other categories. Loading weight (LW) vector (wa), the measure of covariance between the response (y) and predictors (X), is the most commonly and easily accessible filter measure of importance of factors in PLSR model. The factors with loading weights below a certain threshold may be discarded [13]. Another filter measure, variable importance in projection (VIP) measures the significance of each factor by considering loading weight and sum of square explained by corresponding component. The range of threshold is 0 to ∞ for this measure [14]. In addition, selectivity ratio (SR), being an alternative filer methods, measures the ratio between explained variance and residual variance of a factor on response obtained from PLSR model. The cutoff value for SR is based on F-test statistic [15]. Another important factor selection measure is significance multivariate correlation (SMC) defined as the ratio between mean square regression and mean square residual [16].

The main objective of this study was to introduce improved subset factor selection methods for high dimensional data set to enhance the understanding of regression model in the presence of mullticollinearity and identification problem without dropping efficiency. For this purpose, we introduced filter-based FS measures using partial least squares (PLS) regression to determine the significant factors of high dimensional data sets. The accuracy of existing and proposed factor selection methods have been compared over the simulated data with continuous response factor. Moreover, real data set of nutritional status of children is considered to determine its influential factors. Nutritional status Although the importance of adequate nutrition is acknowledged in Pakistan but not enough investigations are held to explore the determinants of child nutritional status. All these FS methods are based on factor importance measures for linear regression. The proposed measures are compared with existing methods regarding performance and selected factors.

II. MATERIALS AND METHODS

A. DATA SIMULATION

In order to make comparison of factor selection methods, data was simulated for continuous response variable (y) and 500 explanatory factors (X). Different correlation structures $R = [0.5, 0.5, 0.3, 0.5, 0, 0, 0, 0, 0, 0]$ were used to introduced important and non-important factors implying 200 significant and 300 non-significant factors. For 100 iterations, simulated data was split into training and test sets and derive the quantiles $Q = (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)$ of this data by means of filter methods to determine the set of threshold. This set of quantiles is used as threshold for the corresponding method. Then, 4 existing and 9 proposed filter factor selection methods were executed for a corresponding set of threshold with 100 iterations for stable estimates and comparison.

B. REAL DATASET

This study used internationally comparable, nationally representative and globally authorized Demographic and Health Survey (DHS) real data set from Pakistan, collected during 2012-13 by the support of United States Agency for International Development and ICF International. The DHS collects high-quality data on nutritional status of mothers and children, maternal and child health, immunization, fertility levels and preferences, infant mortality levels, contraceptive use, awareness about diseases and maternal and neonatal morbidity and mortality. The unit of analysis for this study is children aged 0-60 months born to ever married women aged 15-49 years. Child's nutritional status is the continuous response variable for this analysis. Weight-for-age Z-scores (WAZ) are used as an indicator of child's nutritional status [17]. The sample consists of 96 children belonging to urban area of Punjab region and 139 categorical explanatory variables are included. This study was exempted from ethical considerations as authorized secondary data is used.

C. PARTIAL LEAST SQUARES REGRESSION (PLSR)

PLS modelling is considered here due to multicollinearity and dimensionality problem. The simplest version of PLS named orthogonal score PLS is used due to its broad

pertinence regarding factor selection methods. In PLS, the matrix of explanatory variables $X_{(n,p)}$ is assumed to be linearly associated with the response $y_{(n,1)}$ through the model $y = \alpha + X\beta + \epsilon$, where α and β are the unknown regression parameters and ϵ is the error term. After centering and scaling the variables into $X_0 = X - 1\bar{x}'$ and $y_0 = y - 1\bar{y}$, and some A (where $A \leq p$) is considered to represents number of components for prediction. Then, for a=1,…,A, the general algorithm executes as;

1. Computes the loading weights by $w_a = X'_{a-1}y_{a-1}$.
2. Quantify the score vector by $t_a = X_{a-1}w_a$.
3. Evaluate X-loadings (P_a) and Y-loadings (q_a) by $p_a = X'_{a-1}\frac{t_a}{t'_a t_a}$ and $q_a = y'_{a-1}\frac{t_a}{t'_a t_a}$ respectively.
4. Deflate X_{a-1} and y_{a-1} by $X_a = X_{a-1} - t_a p'_a$ and $y_a = y_{a-1} - t_a q_a$.
5. Repeat the algorithm, if $a < A$.

Assume that W, T, P and q are the matrices/vectors to respectively compile the loading weights, scores, X-loadings and Y-loadings computed at each iteration of the algorithm then the regression coefficient estimators of PLSR model are established by $\hat{\beta} = W(P'W)^{-1}q$ and $\hat{\alpha} = \bar{y} - \bar{X}\hat{B}$.

D. FACTOR SELECTION (FS) IN PLSR

E. REFERENCE METHODS

We considered widely used five existing PLSR FS methods as reference namely standard PLS (PLS), loading weights(LW), variable importance in Projection (VIP), significance multivariate correlation (SMC) and selectivity ratio (SR).

F. PROPOSED METHODS

We proposed 9 filter-based FS methods for PLSR over simulated and real data sets to compare them with reference methods. All the proposed FS methods are based on FS algorithm for linear regression. The proposed methods are outlined as;

1) PRATT'S FS INDEX (PFS)

Pratt in 1987 [18] proposed relative importance of a predictor factor for linear regression. Pratt's factor importance index (PFII) for PLSR represented by R^2 is defined as

$$R^2 = w_j\rho_j, \quad j = 1, 2, \ldots, p \tag{1}$$

where ρ_j is the correlation between y and x and w_j are the loading weights of PLS model.

2) THOMAS FS INDEX (TFS)

Thomas et al. (1998) [19] interpreted factor importance measure as the sample estimates of Pratt's measure, normalized by R^2. The Thomas factor importance index (TFII) symbolized as d_j for PLSR is proposed as

$$d_j = \hat{w}_j\hat{\rho}_j \Big/ R^2, \quad j = 1, 2, \ldots, p \tag{2}$$

where R_2 is the proportion of sample variance explained.

3) STANDARDIZED LOADING WEIGHTS (SLW)

Following standardized regression coefficient [21], [22], standardized loading weights (SLW) presented by w'_j are proposed for PLSR as

$$w'_j = w_j\sqrt{\frac{\sum X_j^2}{\sum Y^2}}, \quad j = 1, 2, \ldots, p \tag{3}$$

4) MEAN VALUE DECOMPOSITION (MVD)

Considering the general linear regression model, Holgersson et al.(2014) determined the contribution of explanatory variable to the mean value of response by mean value decomposition (MVD) [23] and is adopted for PLSR as

$$MVD = \hat{w}_j\bar{X}_j \Big/ \bar{Y}, \quad j = 1, 2, \ldots, p \tag{4}$$

5) LEVEL IMPORTANCE (LIMP)

In economics, level importance being a popular measure refers to the contribution of the predictors in the context of linear regression [24]. Limp for PLSR is established as

$$Limp = w_j\mu_j, \quad i = 1, 2, \ldots, p \tag{5}$$

where μ_i is the mean value of predictor variables.

6) FACTOR CONTRIBUTION (FC) INDEX

Contribution index is a method of driver or importance analysis, used to calculate the extent to which an independent variable explains variation in the dependent variable [25]. Factor contribution (FC) index for PLSR represented by C_j is

$$C_j = \sum_{i}^{p}(w_jX_{ij})^2 \Big/ \sum_{j'}^{n}\sum_{i}^{p}(w_{j\prime}X_{ij\prime})^2 \tag{6}$$

where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots p$.

7) DISPERSION IMPORTANCE (DIMP)

In regression analysis, dispersion importance measures the variance attributed to predictor variable (x) [26] For PLSR, Dimp is modified as

$$Dimp = w_x\sigma_x \Big/ \sigma_y \tag{7}$$

8) STANDARDIZED LOADING WEIGHT IN ECONOMIC IMPORTANCE (ECOLW)

For simple regression model, standardized regression coefficient in economic importance is used to calculate economic importance of explanatory variables [27]. Following standardized regression coefficient in economic importance, EcoLW marked as w^* is presented to measure contribution of explanatory variables in PLSR as

$$w^* = \frac{w_j\sigma_j}{\sqrt{e^2 + \sum_{j=1}^{p}(w_j\sigma_j)^2}} \tag{8}$$

where e^2 is the mean square error.

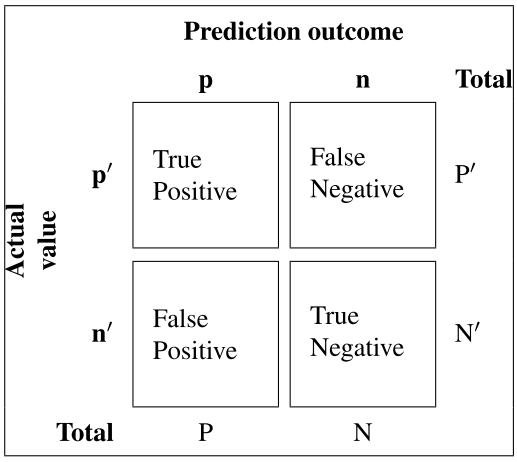


FIGURE 1. Confusion matrix.

9) ABSOLUTE STANDARDIZED LOADING WEIGHT IN ECONOMIC IMPORTANCE (AECOLW)

The absolute standardized regression coefficient in economic importance measures the percentage contribution of the predictors to deviations in y [27]. AEcoLW identified as α_i for PLSR is introduced as

$$\alpha_i = \frac{|w_j\sigma_j|}{|e| + \sum_{j=1}^{p}|w_j\sigma_j|} \tag{9}$$

G. CROSS-VALIDATION

To validate the model performance, we used a cross validation approach to randomly split the data into training and test sets for simulated and real data set. The data is divided into 10 subsets having equal sizes. Test data is constituted on 3 subsets and remaining 7 subsets are included in training data. Hence 70% samples of each data set are included in the training set and the rest 30% are in test set.

H. MODELS EVALUATION

A matrix of explanatory variables for a continuous response over simulated and real data is taken as input for all methods. For simulated data set with defined (true) correlated (important) factors, predicted number of influential factors by each method is obtained and compared through confusion matrix presented in Fig1 for a binary classifier.

Where true positives(TP) are the factors predicted as important by the certain method and they are actually defined significant in the data set,true negatives (TN) are the factors predicted as non-important by a certain method and they are actually defined non-significant in the data set,false positives (FP) (also known as Type I error) are the factors predicted as important by a certain method and they are actually defined non-significant in the data set and false negatives (FN) (also known as Type II error) are the factors predicted as non-important by a certain method and they are actually defined significant in the data set. Then the accuracy measure is used to compute the proportion of correct classifications and is defined as

$$Accuracy = (TP + TN)\Big/(P + N) \tag{10}$$

Sensitivity measures the recognition rate as

$$Sensitivity = TP\Big/P \tag{11}$$

F1 score is the harmonic mean of precision and sensitivity and is defined as

$$F1score = \frac{2(Precision * Sensitivity)}{Precision + Sensitivity} \tag{12}$$

Accuracy,F1 score and sensitivity based on confusion matrix are the measures considered for comparison of existing as well as proposed methods over simulated data analysis.

For real data set, after data pre-processing, each factor selection method is carried out over the data and then PLSR is executed for the selected influential factors obtained by the respective method for 100 iterations. For evaluation purpose, Akaike information criterion (AIC) and Bayesian information criterion (BIC) are also recorded for each PLS model at each iteration. Since PLSR is a factor selection method itself, hence it is used as a standard (reference) for comparison. Based on PLS, 4 existing and 9 proposed factor selection methods were considered and compared. AIC and BIC are used due to their wide applicability for model selection. AIC is used as an estimator of the relative quality of models for a given set of data and hence, provides a means for model selection. AIC is mathematically defined as

$$AIC = 2k - 2ln\hat{L} \tag{13}$$

where k and $\hat{L}$ are the number of estimated parameters and maximized value of the likelihood function for the model respectively. The efficient model is the one having the minimum AIC value among a candidate models for the data. BIC, closely related to the AIC, is also model selection criterion based on likelihood function among a finite set of models, defined as

$$BIC = ln(n)k - 2ln\hat{L} \tag{14}$$

where the additional quantity n denotes the number of observations. The model with minimum BIC value is preferred to select. Improved filter-based FS methods were proposed for performing the factor selection task in PLSR and comparing the performance with existing methods. The following improved FS methods are proposed.

III. RESULTS

A. SIMULATION BASED RESULTS

For simulation study, 150 samples and 500 variables were selected with different correlation structures for continuous response. Since the main purpose of this study was to investigate the performance of factor selection methods in the presence of multicollinearity, hence we considered simulated data with correlated factors having correlation matrix R=c(0.5,0.5,0.3,0.5,0,0,0,0,0,0). Among total 500 factors,

TABLE 1. Anova results demonstrating the significance of factor selection methods for simulated data by analyzing the variation in accuracy is presented.

Factor Selection	Estimate	Standard error	P-value
LW	0.352	0.003	< 0.001
VIP	0.361	0.003	< 0.001
SMC	0.326	0.003	< 0.001
SR	0.322	0.003	< 0.001
PFS	0.170	0.003	< 0.001
TFS	0.171	0.003	< 0.001
SLW	0.345	0.003	< 0.001
MVD	0.213	0.003	< 0.001
$Limp$	0.131	0.003	< 0.001
$Dimp$	0.343	0.003	< 0.001
FC	0.129	0.003	< 0.001
$EcoLW$	0.316	0.003	< 0.001
$AEcoLW$	0.339	0.003	< 0.001

200 were associated with response due to defined correlation structure. The constructed data is then split into test and training sets with sizes 100 and 50 respectively to train and evaluate the existing and proposed PLS factor subset selection methods.

For stable estimates and comparison of factor selection methods, 100 iterations were executed and selected influential factors by each method were compared with actually important factor using confusion matrix technique. Accuracy, sensitivity, F1 test score and number of selected factors were observed and recorded for each method. To make a more formal test, we conducted analysis of variance with results presented in Table1 to assess the significance of factor selection methods in explaining the variation in accuracy for simulated data set. For all factor selection methods, p-value is computed for consideration of significant factor selection methods. The analysis supports that all the indexes included in the study are significantly different from standard PLS. Further, the distribution of accuracy and number of selected variables based on confusion matrix for each filter method are presented in FigureIII-A. We found that VIP and LW performs better among existing methods while three proposed methods including SLW, DIMP and AEcoLW compete these existing ones in terms of accuracy showing that these proposed methods might be the alternatives of the existing methods. But more interestingly, SLW, DIMP and AEcoLW are found to select fewer factors compared to all other methods and exhibiting to approach the number of actually important factors. Although LW and VIP showed nearly similar accuracy as SLW, DIMP and AEcoLW but higher number of selected factor negated their exercise compared to proposed methods. To strengthen the recommendation of these proposed measures, we computed the sensitivity and F1-score for all methods which are shown in Fig3. The graphical representation based on sensitivity showed that these three proposed methods with higher accuracy and appropriate number of selected factors have dominating sensitivity rate exhibiting highest proportion of identifying the actually important factors. Based on precision, F1-score also advocate the employment of these proposed methods. On the basis of accuracy, sensitivity, F1-score and number of selected factors, the proposed methods

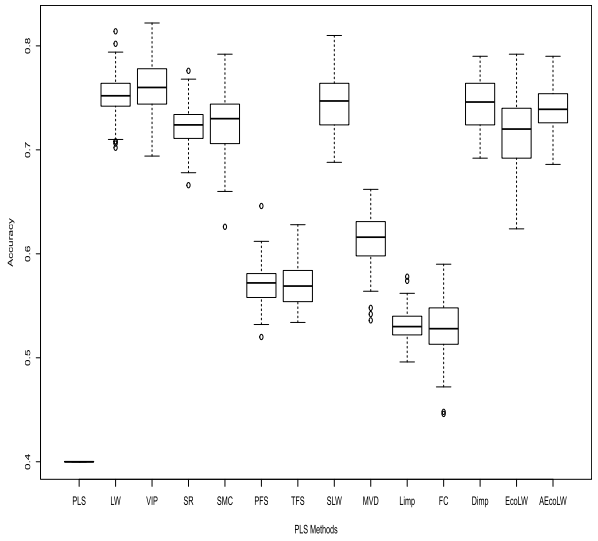


FIGURE 2. The accuracy of FS methods for simulated data set including loading weights(LW), variable importance in projection (VIP), selectivity ratio (SR), significance multivariate correlation (SMC), pratt's FS (PFS) index, thomas FS (TFS) index, standardized loading weight (SLW), mean value decomposition (MVD), level importance (Limp), factor contribution (FC), dispersion importance (Dimp), standardized loading weight in economic importance (EcoLW) and absolute standardized loading weight in economic importance (AEcoLW) is presented in upper panel, the lower panel illustrated the number of influential factors being selected by these FS measures.

are highly recommended to improve the model performance in PLSR as the proposed measures enhance the understanding of regression model without losing accuracy.

B. APPLICATION

The real data set initially contains 96 samples (children) with 157 factors. Then PLSR model is fitted and PLS scores were plotted to identify and remove the outliers and leverages from the data and to allocate clusters. The upper panel of Fig3 showed PLS scores obtained from component 1 and component 2 representing that the closest points are most similar, and those far apart are dissimilar and considered as outliers. The plot in lower panel of Fig3 showed PLS scores corresponding to PLS loadings and is used to gain process understanding. The points close together represent strongly correlated factors showing multicollinearity, preferring use of PLSR instead of multiple linear regression. After removing outliers and leverage observations, the k-mean clustering of PLSR score was carried out to obtain independent and uncorrelated samples.

Following k-mean clustering, the resulting data set contains 137 factors measured over 50 samples (children). For each individual model, to minimize the AIC and BIC, 70% of the samples in the training set and 30% in the test set are randomly selected for cross-validation. Following the principle of the procedure, 100 PLS models were built with Monte Carlo cross-validation method to determine the accuracy (in terms of AIC and BIC) and optimum number of variables for each FS measure. We executed 13 FS measures in PLSR namely loading weights(LW), variable

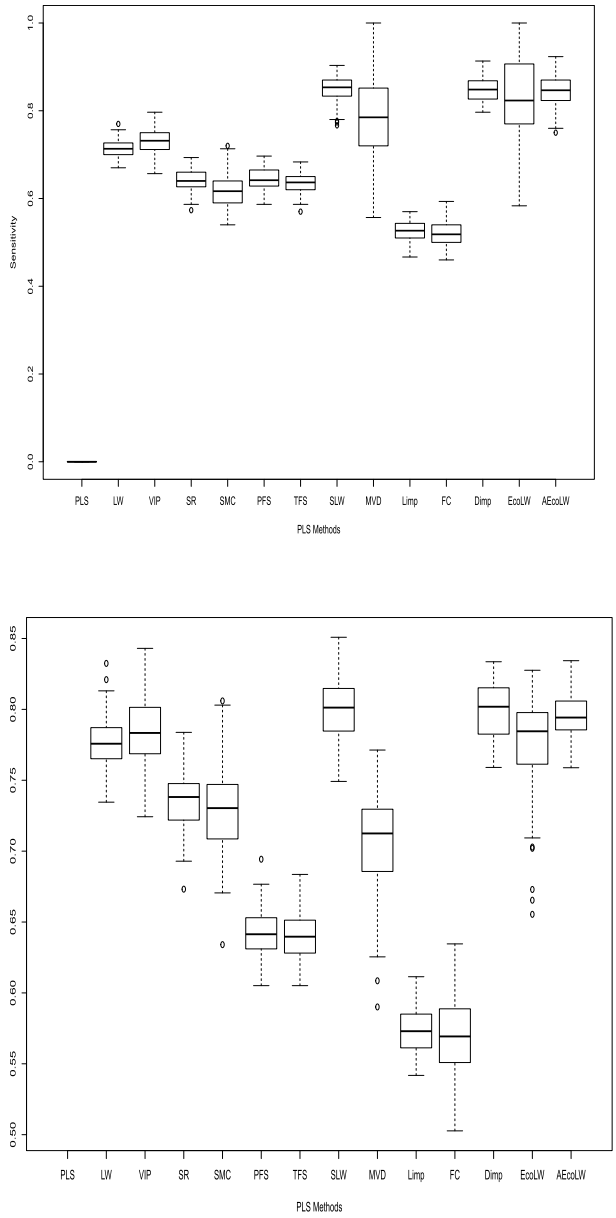


FIGURE 3. The sensitivity of FS methods for simulated data set is presented in upper panel, the lower panel illustrated the F1-score measured by these FS measures.

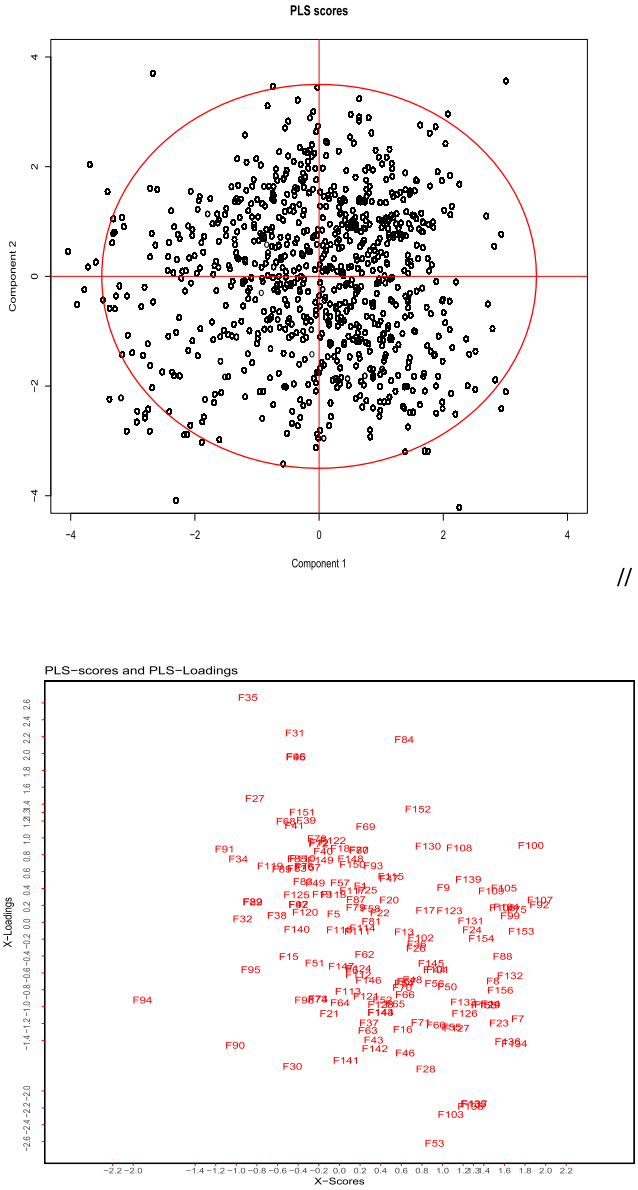
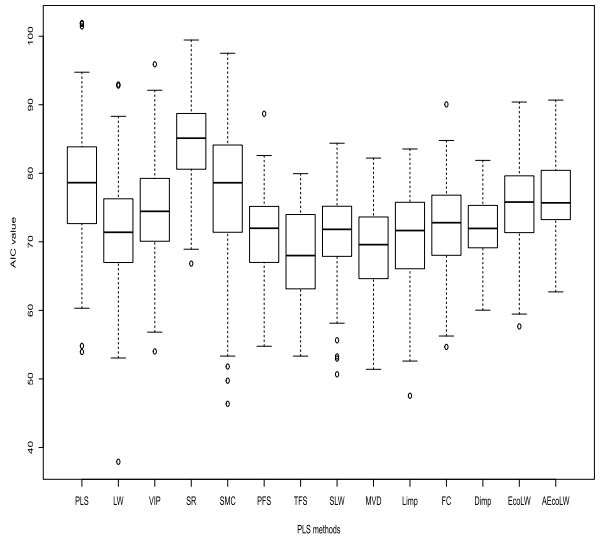


FIGURE 4. The PLS scores from component 1 and component 2 were plotted in upper panel. Children lying out of red circle were considered outliers. Score-loading correspondence plot with loadings from loading matrix and scores from score matrix is presented in lower panel.

importance in projection (VIP), significance multivariate correlation (SMC), selectivity ratio (SR), Pratt's FS index (PFS), Thomas FS index (TFS), standardized loading weights (SLW), mean value decomposition (MVD), level importance (LIMP), dispersion importance (DIMP), factor contribution (FC) index, standardized loading weight in economic importance (EcoLW), absolute standardized Loading weight in economic importance (AEcoLW). The standard PLS algorithm without any FS method is considered for comparison purpose. Contrary to addressing a single value as threshold, a set of quantiles for given data set is considered as threshold values for simulated and real data sets for proposed FS methods which can be further applied to other research studies. The performance comparison over AIC for standard partial least squares (PLS), PLS with loading weights(LW); variable importance in projection (VIP); significance multivariate correlation (SMC); pratt's FS index (PFS); thomas FS index (TFS); standardized loading weight (SLW); mean value decomposition (MVD); level importance (Limp); factor contribution (FC); dispersion importance (Dimp); standardized loading weight in economic importance (EcoLW) and absolute standardized loading weight in economic importance (AEcoLW) based factor subset selection methods are presented in upper panel of Fig 5 indicating higher performance of all FS methods compared to standard PLS for real data set.

All the proposed methods showed relatively higher accu-



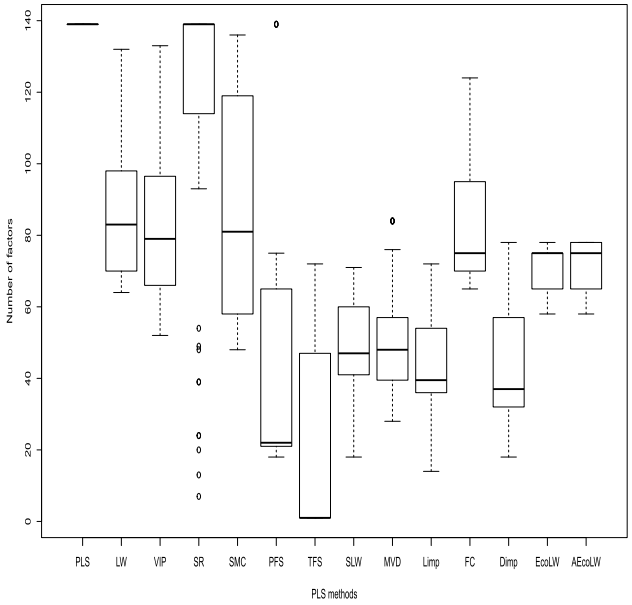


FIGURE 5. **The accuracy performance based on AIC over real data set for standard and proposed PLS methods is presented in upper panel, the lower panel illustrated the number of influential factors for real data being selected by these FS measures.**

racy (low AIC) in predicting nutritional status of children while one existing methods also exhibited comparative efficiency. Among existing FS methods, LW showed highest predictive ability in terms of accuracy while PFS, DIMP, LIMP and SLW competed this existing method as shown in Fig5 reflecting these proposed methods to be alternatives of this existing method showing parallel efficiency. Interestingly, TFS and MVD showed lowest AIC reflecting highest accuracy compared to all other existing as well as proposed methods. Among these two efficient methods, TFS featured highest performance in terms of accuracy. Other four proposed methods including DIMP and LIMP showed minimal negligible difference of AIC value with TFS and

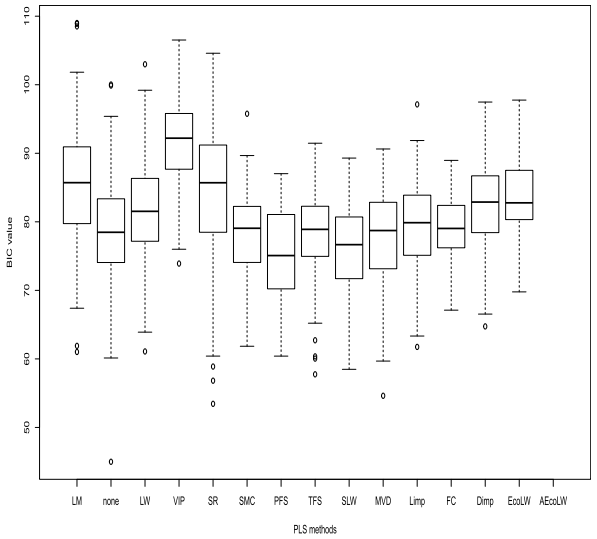


FIGURE 6. **The performance based on BIC for standard and proposed PLS methods over real data set is presented.**

MVD and hence, these four proposed methods are considered as competing each other.The performance of all FS methods based on BIC is presented in Fig 6 for real data set. Nearly consistent performance as AIC is noticed by BIC results showing higher efficiency of TFS, MVD, LIMP and DIMP among all other methods. This evidence further supported the improved accuracy, higher performance and increased efficiency of the proposed methods. Along with efficiency, number of selected factors is also considered for final model selection. To identify a set of important factors among all potential explanatory factors, further real data analysis is presented in lower panel of Fig 5. All the existing methods showed higher number of selected factors with lessen efficiency. Among proposed methods, MVD, LIMP and DIMP showed better accuracy with prudent number of factors. The TFS indicated unacceptable results identifying the set of selected factors having only one factor on average which dropped the further consideration of this measure for real data set. We observed that on average, SLW and MVD choose 47 and 48 out of 137 factors respectively. LIMP showed that 40 factors might be sufficient to obtain maximum accuracy. Dimp is observed to show median value of 37 factors to select. Since magnified accuracy and uniformity of selected factors are essential aspects of a regression model, hence, Dimp is proposed as the FS method with higher efficiency and fewer selected factors for prediction purpose over the given data set.

Cross-validation with 10 components is being used to select a model with the appropriate number of components that has good predictive ability. In this setting, we notice that proposed FS methods lessen the complexity of model by selecting reduced number of factors. Although the model with smallest number of factors that explain a sufficient amount of variability in the predictors and the responses is recommended to select, increased accuracy is also strongly advocated to prefer. Hence, with cross-validation, we selected the model with higher average accuracy and appropriate num-

TABLE 2. Coefficients of finally fitted Dimp method to select influential factors of nutritional status of children are presented.

Selected Factor	Estimate
Household has own transport facility	0.091
Number of daughters	-0.083
Currently pregnant	0.305
Number of living children	-0.247
Fortified food given to child	0.031
Soup/broth given to 196 child yesterday	-0.119
Drank milk from bottle yesterday	0.014
Having bed net	-0.121
Father's education	0.555
Father's occupation	0.464
Father's age	0.108
Preceding birth interval	0.263
Desired pregnancy	-0.317
Timing of first antenatal visit	0.276
Size of child at birth	-0.417
Number of times ate solid, semi-solid or soft food yesterday	-0.224
Weighed during pregnancy	0.272
Ultrasound examination during pregnancy	-0.097
Urine sample taken during pregnancy	0.069
Blood sample taken during pregnancy	0.600
Used iron medicine during pregnancy	0.028
Given nothing to child in first 3 days of birth	-0.242
Antenatal care provided by government hospital	0.251
Antenatal care provided by private doctor	0.025
Baby postnatal check within 2 months	-0.154
Person who performed postnatal checkup	-0.034
Place of first check up of new born	0.142
Received POLIO 0	0.096
Received vitamin A in last 6 months	0.161
Mother work cash after marriage	0.362
Hepatitis awareness	0.013
Mother's age at the time of child birth	-0.347

ber of selected factors.

Dimp is the finally fitted FS method to select influential factors of nutritional status of children and the results are presented in Table2. Dimp select 32 influential factors out of 137 with 9 optimum components and threshold is determine to be 0.003 for the given data set. The results showed that transport facility (car/truck), number of daughters, current pregnancy, number of living children, fortified food given to child, soup/broth given to child yesterday, drank milk from bottle yesterday, having bed net, father's education; occupation and age, preceding birth interval, desired pregnancy, timing of first antenatal visit, size of child at birth, number of times ate solid, semi-solid or soft food yesterday, weighed; ultrasound examination; urine sample taken; blood sample taken and used iron medicine during pregnancy, given nothing to child in first 3 days of birth, antenatal care provided by government hospital, antenatal care provided by private doctor, baby postnatal check within 2 months, person who performed postnatal checkup, place of first check up of new born, received Polio 0, received vitamin A in last 6 months, mother work cash after marriage, hepatitis awareness, mother's age at the time of child birth are associated influential factors with nutritional status of children in Pakistan.

IV. DISCUSSION

We first noticed that among highly recommended existing methods, LW and VIP exhibits higher efficiency while SR showed lowest performance for simulated as well as real data set. LW is being widely used [28]–[30], [32] due to its simplicity and high efficiency. All the existing and proposed filter-based FS methods are observed to be more efficient than PLSR regarding accuracy. Hence, it is highly recommended to apply FS measures to improve efficiency and performance of regression model in presence of multicollinearity. Interestingly proposed methods including standardized loading weights (SLW), dispersion importance (Dimp) and absolute standardized Loading weight in economic importance (AEcoLW) showed nearly equal accuracy rate as existing methods for simulated data. Real data analysis showed that all proposed methods increased the efficiency of regression model in terms of accuracy based on AIC and BIC. Four proposed methods including standardized loading weights (SLW), mean value decomposition (MVD),level importance (Limp) and dispersion importance (Dimp) showed higher performance among all FS methods for real data set. Although existing methods, no doubt, have their own merits but selection of important influential factors is another major concern in regression models. Theoretically, LW only considers the standard loading weights while Dimp also takes into account the amount of variance that is attributed to each predictor variable. This variation highlights the features of Dimp for FS. Dimp is recommended as most appropriate FS method regarding the explanatory features of regression models in the behavioral sciences [26]. On the other hand, LIMP incorporates mean value with loading weights to determine the contribution of predictors while MVD calculates influence of causal factors by mean value decomposition (MVD) [23]. SLW stabilized the loading weights and hence showed equivalent accuracy as LW but with fewer selected factors. AEcoLW considers loading weights along with deviations in response and improve the selection criteria of important factors for better understanding and interpretation of regression model in addition with equal accuracy. Although the simulated and real data sets support the implementation of proposed methods but it highlights the danger of blindly trusting a FS method, which in this case gives improved performance.

Determination of threshold value always remain under discussion in literature as this quantity is the essential requirement to classify selected factors. The FS is regulated by determination of appropriate threshold but detection of optimal threshold is debatable. For PLSR, Hotelling's T^2 statistic coupled with Jacknife testing is recommended as a threshold value for LW [31], [32]. Another general criteria suggested for LW threshold is the range between 0 to 1 [11]. The cutoff values for LW determined by the present study lay in the previously suggested range. No specific criteria is proposed and tested for threshold determination in literature for other FS methods. Percentage contribution of the predictor is

considered in some cases [25], [27] while for others, larger value is recommended as importance of factor [5], [19]. This study proposed and tested a set of quantiles as threshold values for all FS measures including existing and proposed FS methods which can be further applied to other research studies. The influential factors of nutritional status of children selected by Dimp measure were consistent with various previous studies [33]–[45].

V. CONCLUSION

Proposed factor selection methods are shown to be a better choice regarding model performance and number of selected factors in the context of PLS. It indicates that these filter methods produce models with superior interpretation potential. Using PLS with DIMP, the factors identified as the important predictors of nutritional staus of children commensurated with other studies. So, PLS regression algorithms along with proposed factor selection filter methods have the potential as a multivariate technique in scientific research to treat high-dimensional correlated data more efficiently.

REFERENCES

[1] A. A. Ramalho, S. A. Mantovani, B. M. Delfino, T. M. Pereira, A. C. Martins, H. Oliart-Guzmán, A. M. Brãna, Fernando LCC Branco, R. G. Campos, A. S. Guimarães, T. S Araújo, C. S. Oliveira, C. T. Codeço, P. T. Muniz, and M. da Silva-Nunes, "Nutritional status of children under 5 years of age in the brazilian western amazon before and after the interoceanic highway paving: A population-based study," *BMC Public Health*, vol. 13, no. 1, 2013, Art. no. 1098.

[2] Pakistan Demographic. (2015). *Health Survey 2012–13. Islamabad and Calverton, MA: National Institute of Population Studies and ICF International; 2013*. [Online]. Available: https://dhsprogram.com/data

[3] B. Batiro, T. Demissie, Y. Halala, and A. A. Anjulo, "Determinants of stunting among children aged 6-59 months at Kindo Didaye Woreda, Wolaita zone, Southern Ethiopia: Unmatched case control study," *PLoS One*, vol. 12, no. 12, 2017, Art. no. e0189106.

[4] S. B. Geberselassie, S. M. Abebe, Y. A. Melsew, S. M. Mutuku, and M. M. Wassie, "Prevalence of stunting and its associated factors among children 6-59 months of age in Libo-Kemekem district, Northwest Ethiopia; a community based cross sectional study," *PLoS One*, vol. 13, no. 5, 2018, Art. no. e0195361.

[5] T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø, "A review of variable selection methods in partial least squares regression," *Chemometrics Intell. Lab. Syst.*, vol. 118, pp. 62–69, Aug. 2012.

[6] H. Martens and T. Naes, *Multivariate Calibration*. Chichester, U.K.: Wiley, 1989.

[7] H. Chun and S. Keleş, "Sparse partial least squares regression for simultaneous dimension reduction and variable selection," *J. Roy. Stat. Soc., B (Stat. Methodol.)*, vol. 72, no. 1, pp. 3–25, 2010.

[8] A. Höskuldsson, "Variable and subset selection in PLS regression," *Chemometrics Intell. Lab. Syst.*, vol. 55, nos. 1–2, pp. 23–38, 2001.

[9] I. S. Helland, "Some theoretical aspects of partial least squares regression," *Chemometrics Intell. Lab. Syst.*, vol. 58, no. 2, pp. 97–107, 2001.

[10] G. Heinze, C. Wallisch, and D. Dunkler, "Variable selection—A review and recommendations for the practicing statistician," *Biometrical J.*, vol. 60, no. 3, pp. 431–449, 2018.

[11] T. Mehmood, H. Martens, S. Sæbø, J. Warringer, and L. Snipen, "A partial least squares based algorithm for parsimonious variable selection," *Algorithms Mol. Biol.*, vol. 6, no. 1, 2011, Art. no. 27.

[12] L. Nørgaard, A. Saudland, J. Wagner, J. P. Nielsen, L. Munck, and S. B. Engelsen, "Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy," *Appl. Spectrosc.*, vol. 54, no. 3, pp. 413–419, 2000.

[13] I. E. Frank, "Intermediate least squares regression method," *Chemometrics Intell. Lab. Syst.*, vol. 1, no. 3, pp. 233–242, 1987.

[14] S. Wold, M. Sjöström, and L. Eriksson, "Partial least squares projections to latent structures (PLS) in chemistry," *Encyclopedia Comput. Chem.*, vol. 3, Apr. 2002. doi: 10.1002/0470845015.cpa012.

[15] T. Rajalahti, R. Arneberg, F. S. Berven, K.-M. Myhr, R. J. Ulvik, and O. M. Kvalheim, "Biomarker discovery in mass spectral profiles by means of selectivity ratio plot," *Chemometrics Intell. Lab. Syst.*, vol. 95, no. 1, pp. 35–48, 2009.

[16] T. N. Tran, N. Lee Afanador, L. M. C. Buydens, and L. Blanchet, "Interpretation of variable importance in partial least squares with significance multivariate correlation (sMC)," *Chemometrics Intell. Lab. Syst.*, vol. 138, pp. 153–160, Nov. 2014.

[17] M. U. Mushtaq, S. Gull, K. Mushtaq, H. M. Abdullah, U. Khurshid, U. Shahid, M. A. Shad, and J. Akram, "Height, weight and BMI percentiles and nutritional status relative to the international growth references among Pakistani school-aged children," *BMC Pediatrics*, vol. 12, no. 1, p. 31, 2012.

[18] J. W. Pratt, "Dividing the indivisible: Using simple symmetry to partition variance explained," in *Proc. 2nd Int. Tampere Conf. Statist.*, 1987, pp. 245–260.

[19] D. R. Thomas, E. Hughes, and B. D. Zumbo, "On variable importance in linear regression," *Social Indicators Res.*, vol. 45, nos. 1–3, pp. 253–275, 1998.

[20] J. H. Zar, *Biostatistical Analysis*. London, U.K.: Pearson, 1999.

[21] K. Murray and M. M. Conner, "Methods to quantify variable importance: Implications for the analysis of noisy ecological data," *Ecology*, vol. 90, no. 2, pp. 348–355, 2009.

[22] J. H. Zar, *Biostatistical Analysis*. London, U.K.: Pearson, 1999.

[23] H. E. T. Holgersson, T. Norman, and S. Tavassoli, "In the quest for economic significance: Assessing variable importance through mean value decomposition," *Appl. Econ. Lett.*, vol. 21, no. 8, pp. 545–549, 2014.

[24] W. Kruskal and R. Majors, "Concepts of relative importance in recent scientific literature," *Amer. Statistician*, vol. 43, no. 1, pp. 2–6, 1989.

[25] *Driver (Importance) Analysis*. Accessed: Jan. 2019. [Online]. Available: https://wiki.q-researchsoftware.com/wiki/driver(importance)analysis

[26] D. R. Thomas, P. Zhu, and Y. J. Decady, "Point estimates and confidence intervals for variable importance in multiple linear regression," *J. Educ. Behav. Statist.*, vol. 32, no. 1, pp. 61–91, 2007.

[27] O. Sterck, "Beyond the stars: A new method for assessing the economic importance of variables in regressions," Centre Study Afr. Econ., Univ. Oxford, Oxford, U.K., Tech. Rep. 2016-31, 2016.

[28] L. Gidskehaug, E. Anderssen, A. Flatberg, and B. K. Alsberg, "A framework for significance analysis of gene expression data using dimension reduction methods," *BMC Bioinformatics*, vol. 8, no. 1, 2007, Art. no. 346.

[29] D. Jouan-Rimbaud, B. Walczak, D. L. Massart, I. R. Last, and K. A. Prebble, "Comparison of multivariate methods based on latent vectors and methods based on wavelength selection for the analysis of near-infrared spectroscopic data," *Analytica Chim. Acta*, vol. 304, no. 3, pp. 285–295, 1995.

[30] F. Liu, Y. He, and L. Wang, "Determination of effective wavelengths for discrimination of fruit vinegars using near infrared spectroscopy and multivariate analysis," *Analytica Chim. acta*, vol. 615, no. 1, pp. 10–17, 2008.

[31] M. Xiong, J. Zhao, and E. Boerwinkle, "Generalized T^2 test for genome association studies," *Amer. J. Hum. Genet.*, vol. 70, no. 5, pp. 1257–1268, 2002.

[32] H. Martens and M. Martens, *Multivariate Analysis of Quality: An Introduction*. Bristol, U.K.: IOP, 2001.

[33] N. Shaukat, M. Iqbal, and M. A. Khan, "Detrimental effects of intimate partner violence on the nutritional status of children: Insights from PDHS 2012-2013," *Int. J. Community Med. Public Health*, vol. 5, no. 5, pp. 1742–1749, 2018.

[34] J. Alom, M. A. Islam, M. A. Quddus, T. H. Miah, H. Tofazzal, and M. A. Islam, "Socioeconomic factors influencing nutritional status of under-five children of agrarian families in Bangladesh: A multilevel analysis," *Bangladesh J. Agricult. Econ.*, vol. 32, nos. 1–2, pp. 63–74, 2009.

[35] G. Brhane and N. Regassa, "Nutritional status of children under five years of age in shire indaselassie, North Ethiopia: Examining the prevalence and risk factors," *Kontakt*, vol. 16, no. 3, pp. e161–e170, 2014.

[36] M. V. Capanzana, D. V. Aguila, G. M. P. Gironella, and K. V. Montecillo, "Nutritional status of children ages 0–5 and 5–10 years old in households headed by fisherfolks in the Philippines," *Arch. Public Health*, vol. 76, no. 1, 2018, Art. no. 24.
[37] K. G. Dewey and R. J. Cohen, "Does birth spacing affect maternal or child nutritional status? A systematic literature review," *Maternal Child Nutrition*, vol. 3, no. 3, pp. 151–173, 2007.
[38] C. H. D. Fall, H. S. Sachdev, C. Osmond, M. C. Restrepo-Mendez, C. Victora, R. Martorell, A. D. Stein, S. Sinha, N. Tandon, L. Adair, I. Bas, S. Norris, and L. M. Richter, and COHORTS investigators, "Association between maternal age at childbirth and child and adult outcomes in the offspring: A prospective study in five low-income and middle-income countries (COHORTS collaboration)," *Lancet Global Health*, vol. 3, no. 7, pp. e366–e377, 2015.
[39] S. F. Hasnain and S. K. Hashmi, "Consanguinity among the risk factors for underweight in children under five: A study from rural Sindh," *J. Ayub Med. College Abbottabad*, vol. 21, no. 3, pp. 111–116, 2009.
[40] O. R. Katoch and A. Sharma, "Socioeconomic factors, living conditions and child undernutrition among school going children in rural areas of district Doda, Jammu & Kashmir, India: A preliminary study," *Indian J. Nutrition*, vol. 3, no. 1, p. 123, 2016.
[41] J. Lambert, C. Agostoni, I. Elmadfa, K. Hulshof, E. Krause, B. Livingstone, P. Socha, D. Pannemans, and S. Samartín, "Dietary intake and nutritional status of children and adolescents in Europe," *Brit. J. Nutrition*, vol. 92, no. S2, pp. S147–S211, 2004.
[42] J. Liang, Z. Zhang, W. Yang, M. Dai, L. Lin, Y. Chen, J. Ma, and J. Jing, "Association between cesarean section and weight status in Chinese children and adolescents: A national survey," *Int. J. Environ. Res. Public Health*, vol. 14, no. 12, p. 1609, 2017.
[43] W. Girma and T. Genebo, "Determinants of nutritional status of women and children in Ethiopia," ORC Macro, Calverton, MD, USA, Tech. Rep., 2002.
[44] R. Pongou, M. Ezzati, and J. A. Salomon, "Household and community socioeconomic and environmental determinants of child nutritional status in Cameroon," *BMC Public Health*, vol. 6, no. 1, 2006, Art. no. 98.
[45] L. Samiak and T. I. Emeto, "Vaccination and nutritional status of children in Karawari, East Sepik Province, Papua New Guinea," *PLoS One*, vol. 12, no. 11, 2017, Art. no. e0187796.

TAHIR MEHMOOD received the Ph.D. degree in statistics from Norwegian University of Life Science (NMBU), Norway, in 2012. He is currently an Associate Professor with the School of Natural Science (SNS), National University of Sciences and Technology (NUST), Islamabad, Pakistan. His research interests include multivariate statistics, statistical learning, classification, clustering, variable selection, and the application of these methods/algorithm covers chemometrics, enviormetrics, public health and in related areas.

MARYAM SADIQ received the M.Phil. degree in statistics from Arid Agriculture University, Islamabad, Pakistan, in 2014. She is currently pursuing the Ph.D. degree in statistics with Riphah International University, Islamabad. She is also an Assistant Professor with the Department of Statistics, University of Azad Jammu and Kashmir, Muzaffarabad, Pakistan. Her research interests include multivariate statistical modeling techniques, high dimensional data, factor selection, and regression analysis techniques.

MUHAMMAD ASLAM received the Ph.D. degree in statistics from the University of Wales, U.K., in 1996, and the M.Sc. degree in statistics from the University of the Punjab, Pakistan. In 1980, he started his career as a Lecturer of statistics with the University of Balochistan, Pakistan. He served Quaid-i-Azam University (QAU), Pakistan, for about 25 years, where he was the Head of the Department of Statistics, for six years. He has 38 years teaching and research experience at graduate level. He has supervised nine Ph.D. degree scholars and 112 M.Phil. degree students. He has 156 research publications in international repute journals. He is currently a Professor of Statistics with Riphah International University, Pakistan. His research interests include Bayesian inference and mathematical statistics.