Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 5561752, 8 pages
https://doi.org/10.1155/2021/5561752

Research Article

Regularized Feature Selection in Categorical PLS for Multicollinear Data

Tahir Mehmood

School of Natural Sciences (SNS), National University of Sciences and Technology (NUST), Islamabad, Pakistan

Correspondence should be addressed to Tahir Mehmood; tahime@gmail.com

Received 21 February 2021; Revised 10 April 2021; Accepted 7 May 2021; Published 17 May 2021

Academic Editor: Xiaoheng Chang

Article presents the algorithm which models the categorical multicollinear data by providing the balance in model accuracy on test data and number of selected features in the model. In all scientific fields, multicollinear data is being generated, where obviously some variables are noise and some are influential reference to response variable. Features and response appeared to be categorical in mathematical and statistical modeling of public health data. These datasets usually appeared to collinear, where partial least squares (PLS) is the potential method, which is not feature selection at its default level and deals with quantitative features. Recently, categorical PLS (Cat-PLS) is introduced. We have implemented the regularized feature selection in Cat-PLS where filter-based feature selection and categorical mean through Cramer's V, Phi coefficient, Tschuprow's T coefficient, Contingency Coefficient, and Yule's Q and Yule's Y are used. Monte carlo simulation with 100 runs indicates Cramer $V * \text{VIP}$ is the better choice in terms of better model performance, number of feature selection, and interpretations for modeling the stillbirths, which is taken as the case study. The framework can be used in related areas to explore and model the related data structures.

1. Introduction

Sciences are experiencing the multicollinear datasets where the task is to establish the meaningful relation for better understanding and interpretation of real life process [1–4]. Like other sciences, in public health, the selection of influential features X_{nxp} is of the researcher's interest [2, 5, 6] which explains the variation in response or outcome y_{nx1}, where n is sample size and p is number of features. Here, data usually comprises of correlated features having categorical nature [7]. Logistic regression-type models are the famous candidates for modeling the categorical response [8]. In presence of multicollinearity, logistic regression estimates' variance gets so large; hence, logistic regression does not work optimally if the features are correlated. Ridged logistic regression controls variance of the estimates but is unable to find the influential features [9] and is not designed for categorical X. Alternative to the logistic-type regression model is Partial Least Square (PLS) regression, which is a statistical learning approach specifically designed to model the correlated X [10]. Developments in PLS algorithm is going on with passage of time, and a latest contribution is the up gradation of PLS algorithm for categorical X features [7], called it Categorical PLS (CPLS). In PLS, loading weights have a pivotal role in model building which reflects the mutual correlation of a respective X feature with response y. The PLS loading weights is somehow closer to Pearson coefficient of correlation which is being replaced with Cramer's V, Phi coefficient, Tschuprow's T coefficient, Contingency Coefficient, and Yule's Q and Yule's Y correlation measure in CPLS. Hence, CPLS is a potential candidate for modeling the categorical response y with multicollinear X. In PLS, several methods have been proposed for influential feature selection and are reviewed in [11]. Feature selection in PLS can be grouped into three broader categories which are filter, embedded, and wrapper. Filter feature selection is a two-step procedure where, at the first stage, PLS is fitted and, at the second stage, filter measure is computed. Features above a threshold are marked as influential. In embedded feature selection, filter selection is embedded in iterative computational loop of PLS. In wrapper feature selection, an external loop is considered

over filter selection, whereas in each loop, selection is carried out, the model is updated, and performance is evaluated. All three types of feature selection methods have their own advantages and disadvantages. For instance, filter methods are very fast but may have low performance. Embedded and wrapper methods are relatively time expensive but are expected to perform better compared to filter methods. Regularized elimination procedure for feature selection in PLS [12] is a potential candidate from wrapper selection methods. Regularized elimination procedure in PLS selects a model next to optimal, given that the selected model's performance is not significantly different from the optimal model, whereas the selected model has fewer features compared to the optimal model. In this article, we have proposed the modification in regularized elimination procedure over categorical PLS instead of standard PLS.

We have implemented the proposed regularized elimination in categorical PLS over the still birth, which is crucial issue in developing countries. Although considerable progression has been observed in the last 25 years [13], but still the issue needs considerable attention [14]. Several surveys cover the issues regarding stillbirth in Pakistan, but there is need to determine the causes of stillbirth [15]. The most comprehensive and reliable source to have data related to still births and related features is Pakistan Demographic and Health Survey (PDHS) conducted by the National Institute of Population Studies (NIPS) and is technically assisted and funded by USAID. Although the case study taken here is from public health, but the proposed method is applicable over the categorical multicollinear data. Possibilities include, engineering, robotics, gaming, chemometrics, and bioinformatics.

2. Methods

2.1. Data Set. The data set was obtained from the Pakistan Demographic and Health Survey (PDHS) (https://www.nips.org.pk/PDHS-Data-Set.htm) from 2017–18, which was designed to provide population and health indicators at the national and regional levels. The sample design contained specific indicators for each of the five provinces (Punjab, Sindh, Khyber Pakhtunkhwa (KPK), Balochistan, and Gilgit Baltistan) of Pakistan. According to WHO definition of late fetal deaths for international standards, the sample of stillbirth for this study was restricted to birth of 28 or more weeks of gestation. Women with incomplete information were excluded from the sample, and then, 752 women who experienced stillbirths and 1504 women who had live births were included in the analysis. The sampled women included in the case group were mothers of newborn babies without signs of life after at least 28 weeks of pregnancy, while women included in the control group had live births. The response variable y of this study was the occurrence (labeled as 1) or nonoccurrence (labeled as 0) of stillbirths among women of child bearing age (15–49 years). Maternal features such as socio, economic, and other health features related to still births were taken as explanatory features, i.e., X matrix.

2.2. Categorical Partial Least Squares (CPLS). Categorical Partial Least Squares (CPLS) [7] models the categorical data set which is the upgradation of standard PLS [10]. The algorithm for CPLS starts with centered features' data $\mathbf{X}_0 = \mathbf{X} - \mathbf{1}\overline{\mathbf{x}}'$ and response $\mathbf{y}_0 = \mathbf{y} - 1\overline{y}$. CPLS is an iterative procedure based on C iterations called components. In each CPLS component, $c = 1, 2, \ldots, C$ loading weights, score vectors, X matrix, y loadings, and deflated X and y are computed as

(1) Loading weights can be defined through Cramer's V w_{CV} [16], Phi coefficient w_{PC} [16], Tschuprow'T coefficient w_{TC}[17], Contingency Coefficient w_{CC} [18], and Yule's Q w_{YQ} and Yule's Y w_{YY} [19] as

$$w_{\mathrm{CV}} = \sqrt{\frac{\chi^2/n}{\min(r-1, c-1)}}, \tag{1}$$

where χ^2 is derived from Pearson's chi-squared test, n is the total number of observations, and r and c denote number categories, respectively, in response and in respective feature:

$$w_{\mathrm{PC}} = \frac{\chi^2}{n}, \tag{2}$$

which is referred as mean square contingency coefficient:

$$w_{\mathrm{TC}} = \sqrt{\frac{\phi^2}{\sqrt{(r-1)(c-1)}}}, \tag{3}$$

which is the refined form of Phi loading weights, where r and c denote the number of categories, respectively, in response and in respective feature. ϕ is the mean square contingency defined as

$$\phi = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(\pi_{ij} - \sum_{j=1}^{c}\pi_{ij}\sum_{i=1}^{r}\pi_{ij}\right)^2}{\sum_{j=1}^{c}\pi_{ij}\sum_{i=1}^{r}\pi_{ij}}, \tag{4}$$

which is the proportion of the sample in the $(i, j)^{\mathrm{th}}$ cell of the $r \times c$ contingency table:

$$w_{\mathrm{CC}} = \sqrt{\frac{\chi^2}{N + \chi^2}}, \tag{5}$$

which measures the strength of association between categorical features:

$$\begin{aligned} w_{YQ} &= \frac{OR - 1}{OR + 1}, \\ w_{YY} &= \frac{\sqrt{OR} - 1}{\sqrt{OR} + 1}, \end{aligned} \tag{6}$$

determines the strength of relationship between feature and the response based on odds' ratio (OR). Normalizing the loading weights, $\mathbf{w}_k \leftarrow \mathbf{w}_c / \|\mathbf{w}_c\|$.

(2) Compute the score vector $\mathbf{t}_c$ by $\mathbf{t}_c = \mathbf{X}_{c-1}\mathbf{w}_c$.

(3) Computing the X-loading $\mathbf{p}_c$ through regressing $\mathbf{X}_{c-1}$ on the score vector, $\mathbf{p}_c = \mathbf{X}'_{c-1}\,(\mathbf{t}_c/\mathbf{t}'_c\mathbf{t}_c)$.

Similarly, compute the Y-loading q_k through

$$q_k = \mathbf{y}_{c-1}{}' \frac{\mathbf{t}_c}{\mathbf{t}'_c\mathbf{t}_c}. \tag{7}$$

(4) Deflate $\mathbf{X}_{c-1}$ and $\mathbf{y}_{c-1}$ by subtracting the involvement of $\mathbf{t}_c$:

$$\begin{aligned} \mathbf{X}_c &= \mathbf{X}_{c-1} - \mathbf{t}_c\mathbf{p}'_c, \\ \mathbf{y}_c &= \mathbf{y}_{c-1} - \mathbf{t}_c q_c. \end{aligned} \tag{8}$$

(5) If $c < C$, come back to 1.

From each component-computed loading weights, score vector, loadings, and deflated data are stored in respective matrices/vectors $\mathbf{W}$, $\mathbf{T}$, $\mathbf{P}$, and $\mathbf{q}$.

2.3. Regularized Elimination in CPLS. Regularized elimination is the wrapper feature selection method. The current version is modification and simplification of regularize elimination in PLS [12]. Here, we need to attach the filter measures with CPLS in a wrapper function. We have considered the following filter measures, which reflects the level of importance of each explanatory feature for response:

(i) Loading weights (LW): PLS LW reflect the covariance of feature j with response; hence, the importance of feature j at CPLS component c can be measured by $r_j = |\mathbf{w}_{c,j}/\max \mathbf{w}_{\mathbf{c}}|$. Features having $|\mathrm{LW}| < u$ for some user defined fixed threshold can be eliminated from the model.

(ii) Regression coefficients (RC): RC is an established and well-known measure for feature selection defined as $\mathrm{RC} = \mathbf{W}(\mathbf{P}'\mathbf{W})^{-1}\mathbf{q}$. Features having $|\mathrm{RC}| < u$ for some user-defined fixed threshold can be eliminated from the model.

(iii) Variable importance on projections (VIP): VIP for the feature j is defined according to [20] as

$$\mathrm{VIP}_j = \sqrt{\frac{p\sum_{c=1}^{c^*}\left[\left(\mathbf{p}_{2c}^2\mathbf{t}'_c\mathbf{t}_c\right)\left(\mathbf{w}_{cj}/\|\mathbf{w}_c\|\right)^2\right]}{\sum_{c=1}^{c^*}\left(\mathbf{p}_{2c}^2\mathbf{t}'_c\mathbf{t}_c\right)}}, \tag{9}$$

where $c = 1, 2, \ldots, c^*$, $\mathbf{w}_{cj}$ is the loading weight for feature j using c components, and $\mathbf{t}_c$, $\mathbf{w}_c$, and $\mathbf{p}_{2c}$ are, respectively, CPLS scores, loading weights, and y-loadings, respectively, corresponding to the c^{th} component. Feature j can be eliminated if $\mathrm{VIP}_j < u$ for some user-defined threshold $u \in [0, \infty)$.

(iv) Selectivity ratio (SR): SR is based on the target projection approach [21] which is postprojection of the predictor features onto the fitted response vector from the estimated model. For each feature j, r_j can be computed as

$$\mathrm{SR}_j = \frac{\mathrm{Var}_{\exp,j}}{\mathrm{Var}_{\mathrm{res},j}}, \tag{10}$$

where $\mathrm{Var}_{\exp,j}$ is the explained variance and $\mathrm{Var}_{\mathrm{res},j}$ is the residual variance for feature j from the target projection model. Feature j can be eliminated if $\mathrm{SR}_j < u$ for some user-defined threshold $u \in [0, \infty)$

(v) Significance multivariate correlation (SMC): SMC [22] is defined as the ratio of mean square of regression compared to mean square residuals:

$$\mathrm{SMC}_i = \frac{\mathrm{MS}_{i,PLS_{\mathrm{regression}}}}{\mathrm{MS}_{i,PLS_{\mathrm{residuals}}}} = \frac{\left\|\widehat{y}\widehat{\beta}'_i/\|\beta'_i\|^2\right\|^2}{\left\|x_i - \left(\widehat{y}\widehat{\beta}'_i/\|\beta'_i\|^2\right)\right\|^2/(n-2)}, \tag{11}$$

where feature j can be eliminated if $\mathrm{SMC}_j < u$ for some user-defined threshold $u \in [0, \infty)$.

Once the filter measures are defined, the elimination procedure for removing the 'worst' features from the CPLS model is presented here. Let $\mathbf{M}_0 = \mathbf{X}$ and F_j be any filter measure from LW, RC, VIP, SR, or SMC.

(1) For iteration g, run $\mathbf{y}$ and $\mathbf{Z}_g$ through cross-validated CPLS and performance P_g is computed. The matrix $\mathbf{Z}_g$ has p_g columns, and for used filter measure, we get p_g criterion values which are sorted as $s_{(1)}, \ldots, s_{(p_g)}$.

(2) There will M criterion values below the threshold u, i.e., number of noninfluential features. Let $N = \lceil fM \rceil$ for some fraction $f \in \langle 0, 1]$. Eliminate the features corresponding to the N most extreme criterion values.

(3) If there are still more than one feature left, let $\mathbf{Z}_{g+1}$ contain these features, and return to (1).

The fraction f determines the part of the elimination algorithm, where small f will eliminate few features from each iteration. With each iteration, number of influencing features in $\mathbf{Z}_g$ decreases, but the performance may increases or decrease. The increase in performance is because of removal of noise features and decrease in performance is because of relevant features. After the optimal iteration g^* with performance $P^* = P_{g^*} = \max_g P_g$, there is reduction in the number features against the modest drop in performance. Hence, by eliminating beyond g^*, one could have a much simpler model with small loss of performance. To conduct this regularization, McNemar test can be used. The prediction for the optimal model is compared with the models next to the optimal model beyond g^* iteration. If the prediction difference is not significant and this happens over several iterations beyond g^*, then the selected model is the one having the least number of features.

2.4. Model Fitting and Validation. The regularized elimination in CPLS has several parameters to tune, for instance, elimination fraction f, number of CPLS components c, and

threshold used for filter measure u. Elimination fraction f affects the number of iteration in regularized elimination in CPLS; hence, it mostly affects over the computational time, so we can take $f = 0.1$, which means, in each iteration, we are eliminating only 10 % of extreme criterion features. u and c can effect the model performance; hence, they need to tune. We have considered $c = 1, 2, \ldots, 20$ and distribution-based 4 levels, i.e., quantiles for u. For this, we first computed the respective filter measure for all features in the model; then, the 4 quantiles of the filter measure were used as different levels of u. For fitted model's evaluation and parameter tuning, we have adopted the cross-validation procedure. For this, full data set was divided into training (70 %) and test (30 %) data. Using training data, the CPLS model was fitted against all possible combination of u and c and performance on test data was computed. We have used accuracy as performance measure, i.e., how good CPLS predicts the response on test data. Since split of data into test and training is random, to minimize the effect of this randomness, we have used Monte Carlo simulation with 100 runs, where, in each step, the CPLS model was fitted and evaluated as per above description.

2.5. Computations. All methods are implemented in the *R* computing environment (http://www.r-project.org/) and codes are available from corresponding author upon request.

3. Results and Discussion

The data set contains a total of 2256 births with 752 still-births, and we observed several outliers in the data. Moreover, for modeling, we assume the samples should be independent of each other. So, in order to remove outliers and to ensure samples are independent, we have used k-mean clustering over still birth and alive birth samples separately. We found data of 141 independent samples with 94 alive births and 47 still births with 34 features covering maternal features, placental deficiency, fetal growth limitations, fetal growth features, and congenital features related to still births which were taken as explanatory features, i.e., X−matrix. In CPLS, there are six categorical measure-based loading weights, i.e., Cramer's V w_{CV}, Phi coefficient w_{PC}, Tschuprow'T coefficient w_{TC}, Contingency Coefficient w_{CC}, and Yule's Q w_{YQ} and Yule's Y w_{YY}. Each CPLS was fitted within regularized elimination which utilizes five filter measures that is loading weights (LW), regression coefficients (RC), variable importance on projections (VIP), selectivity ratio (SR), and significance multivariate correlation (SMC). Hence, there are $6 \times 5 = 30$ regularized elimination in CPLS models to fit and to compare. For this, 100 Monte Carlo simulations were executed. The response y and explanatory feature matrix X were divided into training (70%) and test (30%) data sets in each Monte Carlo simulation. Each of 30 regularized eliminations in the CPLS model was fitted over the training data, while test data was used to tune the model parameters and to measure the model performance that is accuracy. Hence, each model was fitted and evaluated 100 times.

Since regularized elimination in CPLS selects the model after the optimal model having nonsignificant difference in response prediction. Hence, we have two models named as the optimal model and selected model for each of 30 regularized eliminations over each Monte Carlo run in CPLS. In regularized elimination, we have used McNemar test with $p - value = 0.05$. Since we have used several filter measures and categorical measures in regularized categorical PLS, so we have used Kruskal–Wallis test to study their significance over the accuracy on test data. It appears the accuracy on test data is significantly varying with filter measures ($p -$ value ≤ 0.01) and is also significantly varying with categorical measures ($p -$ value ≤ 0.01). Figure 1 presents the distribution and comparison of accuracy from both models. This indicates the CPLS based on Contingency Coefficient w_{CC} with filter measures SMC, i.e., ContCoef $*$ SMC and with filter measure SR, i.e., ContCoef $*$ SR have low accuracy in optimal model and consequently in the selected model. Left-hand panel of Figure 1 is a magnified view of the upper right part of the right-hand panel. This indicates CPLS based on Tschuprow'T coefficient w_{TC} with filter measure LW, i.e., TschuprowT $*$ LW, Phi coefficient w_{PC} with filter measure LW, i.e., Phi $*$ LW and with filter measure VIP, i.e., Phi $*$ VIP are performing with best accuracy over optimal model but having relatively low accuracy over the selected model. Yule's Q w_{YQ} with filter measure LW, i.e., Yule $Q*$ LW and Cramer's V w_{CV} with filter measure VIP, i.e., Cramer $V*$ VIP have reasonably good performance, which is dully supported by Wilcoxon rank sum test with continuity correction ($p -$ value $= 0.047$).

When choosing a model for feature selection, the stability of the model is an important aspect to consider. Figure 2 presents the standard deviations of accuracy for all fitted models. The variation is smaller for Yule $Y*$ LW, Yule $Q*$ SMC, Yule $Q*$ SR, and Cramer $V*$ VIP. In concert with accuracy analysis and stability analysis expressed from Figures 1 and 2, respectively, we can conclude that Cramer $V*$ VIP performs good average accuracy both on the optimal model and the selected model and, at the same time, has higher stability in accuracy on the selected model since this has shown lower standard deviation of accuracy.

When it comes to the parsimonious model, sample size together with accuracy of the fitted model is important. Smaller number of the features in the fitted model presents the model is better for interpretation and understanding the real-life phenomena. The distribution of numbers of features used in the selected and optimal model is presented in Figure 3. All selected models have relativity low numbers of features compared to the optimal model, which is the expected pattern from regularized elimination in CPLS algorithm. Since filter measures mainly contribute in feature selection for CPLS, hence, distribution of feature selection measures for CPLS algorithms under consideration is important. It appears SR and then VIP choose lowest number of features for all types of CPLS except for Yule' Y and Yule' Q. Moreover, for Yule' Y and Yule' Q, RC and VIP are choosing lowest number of features. For further investigation, the number of PLS components presenting the complexity of the fitted model are

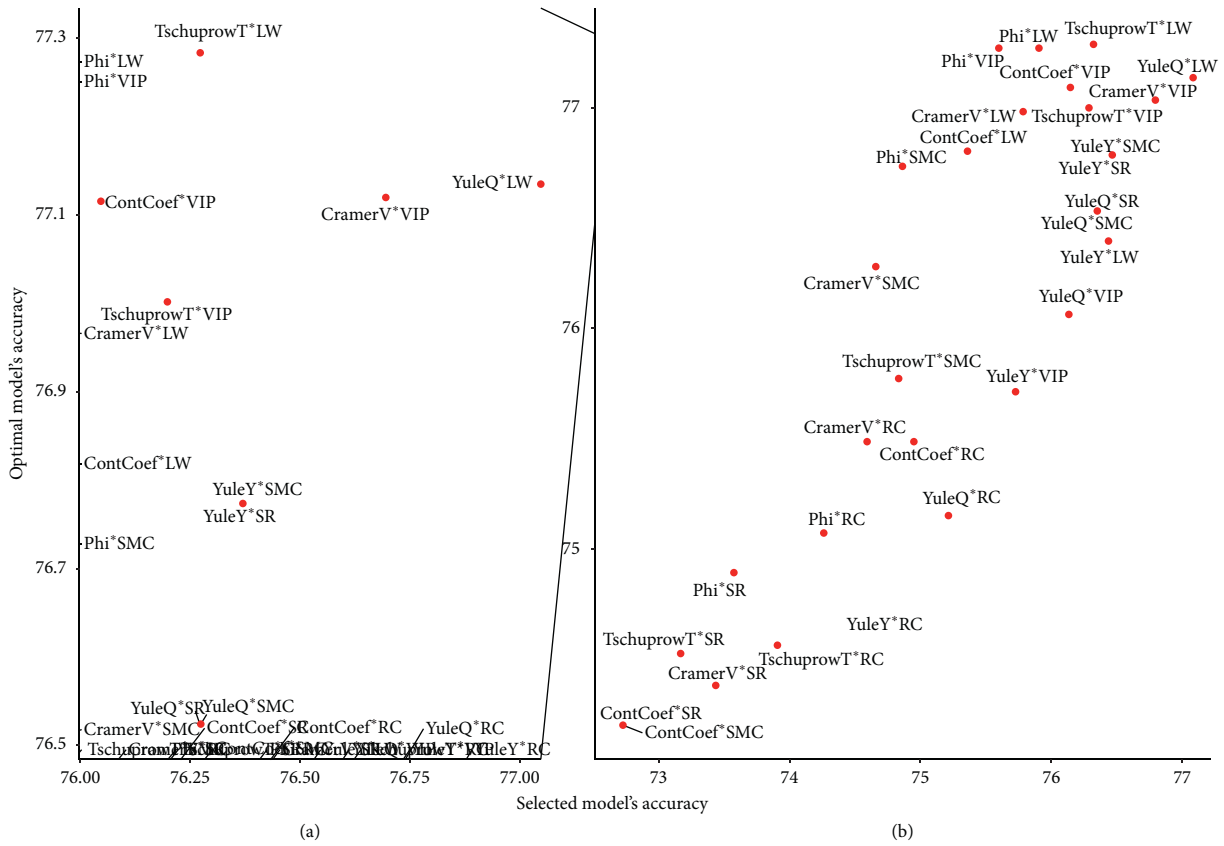

FIGURE 1: Average accuracy of optimal model versus the average accuracy of the selected model is presented here. Left-hand panel is a magnified view of the upper right part of the right-hand panel.

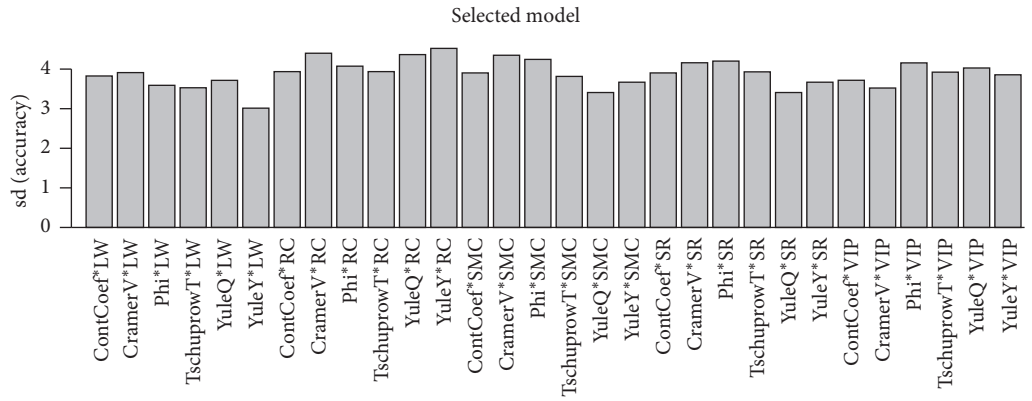

FIGURE 2: The standard deviations of accuracy of all fitted model is presented.

exhibited in Figure 4. Left panel shows the distribution of number of components consumed for evaluated filter measures indicating RC, SMC, and SR has lower level of complexity compared to LW and VIP. The right panel presents the distribution of the number of components considered in categorical choice of PLS (CPLS), indicating ContCoef, Cramer V, and Phi-based CPLS are relatively less complex compared to other CPLS methods.

It appears Cramer V $*$ VIP has smaller number of features in the selected model as well as having good and consistent accuracy, hence can be used as a potential candidate for modeling the occurrence or nonoccurrence of stillbirths among women of child bearing age. Influential features obtained by fitting the Cramer V $*$ VIP are presented in Table 1. The count and percentage of these features over the occurrence and nonoccurrence of still birth together with their odds' ratio (OR) and significance is also presented. As the small subset of refined cases is considered here, hence the presented effects are not for quantification to be used for trend.

Results indicate that it is 1.51 times more likely to have still births in Baluchistan province compared to Punjab

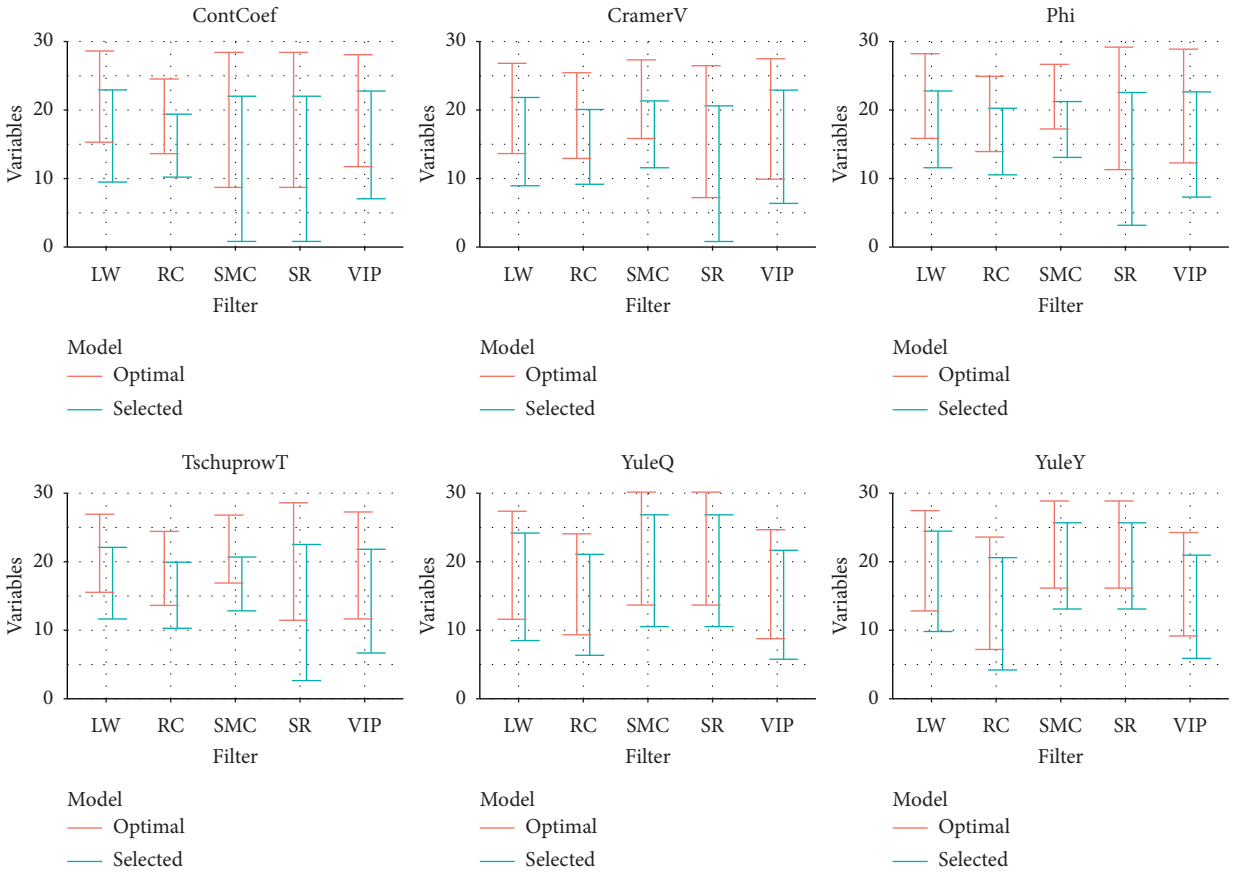

Figure 3: The distribution of numbers of features used in the selected and optimal model is presented.

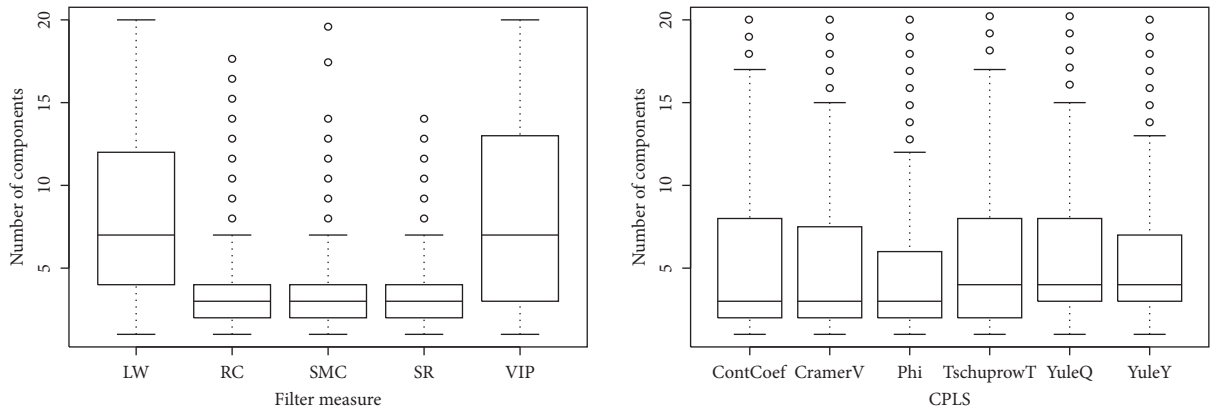

Figure 4: The distribution of number of components over the filter measures and CPLS obtained from 100 Monte Carlo simulation runs is presented here.

province. Punjab is more developed province compared to Baluchistan province. The health and related facilities are much better in Punjab compared to Baluchistan [23] and the same trend is reflected in the results. Similarly, it is 2.5 times more likely to have still births in rural areas compared to urban areas.

This is again more likely since cities or big towns are expected to have better health facilities. It is 0.685 times less likely to have still births if mother's age increases from ≤ 19 to $27 - 33$. These results support the findings reported in [24]. With nurses and traditional attendance, the chances of still birth decreases by 0.979 times and 0.901 times, respectively. Increase in antenatal care (ANC) visits upto 3 counts, the chances of still birth decreases by 0.441 times. It is 1.089 times more likely to have still births for women having pregnancy complication. The use of iron tablets during pregnancy decreases the risk of still birth by 0.916 times. If a mother is socially dependent for medical assistance, then the chances of stillbirth get increased by 2.53 times. Primary, secondary, and higher level of husband's

Table 1: Influential features obtained from Cramer $V *$ VIP which affects the occurrence or nonoccurrence of stillbirths among women are presented with counts, %, p values, and odds' ratios (OR).

Feature		Cases					
		Born alive		Stillbirth			
	Levels	Count	%	Count	%	p values	OR
Region	Punjab	36	66.7	18	33.3	Reference	
	Sindh	24	64.9	13	35.1	0.516	1.08
	KPK	22	75.9	7	24.1	0.270	0.64
	Balochistan	12	57.1	9	42.9	0.004	1.51
Residence place	Urban	46	78.0	13	22.0	Reference	
	Rural	48	58.5	34	41.5	0.012	2.50
Mother's age	≤ 19	12	63.2	7	36.8	Reference	
	20–26	47	64.4	26	35.6	0.561	0.94
	27–33	35	71.4	14	28.6	0.050	0.685
Nurse assistance	No	23	56.1	18	43.9	Reference	
	Yes	71	71.0	29	29.0	0.015	0.979
Traditional assistance	No	64	66.0	33	34.0	Reference	
	Yes	30	68.2	14	31.8	0.017	0.901
ANC timing	1st	8	53.3	7	46.7	Reference	
	2nd	29	61.7	18	38.3	0.840	0.7094
	3rd	57	72.2	22	27.8	0.045	0.441
Pregnancy complications	No	50	67.6	24	32.4	Reference	
	Yes	44	65.7	23	34.3	0.057	1.089
Iron tablets	No	54	65.9	28	34.1	Reference	
	Yes	40	67.8	19	32.2	.044	0.916
Dependent medical help	No	51	76.1	16	23.9	Reference	
	Yes	39	58.1	31	41.9	.005	2.533
Husband education	Illiterate	14	50.0	14	50.0	Reference	
	Primary	14	63.6	8	36.4	.099	0.571
	Secondary	37	69.8	16	30.2	.004	0.432
	Higher	29	76.3	9	23.7	.028	0.310
Working women	No	12	44.4	15	55.6	Reference	
	Yes	82	71.9	32	28.1	.004	0.312
Pregnancy order	1–3	59	78.7	16	21.3	Reference	
	4–6	27	60.0	18	40.0	.002	2.46
	≥7	8	38.1	13	61.9	.004	5.99

education decreases the risk of still birth by 0.57, 0.43, and 0.31 times, respectively, compared to illiterate husband. It is observed that working women are 0.312 times less likely to have stillbirths. Compared to 1–3 pregnancy order, a woman having 4–6 and ≥7 pregnancy order are 2.46 and 5.99 times more likely to have stillbirths. Most importantly, it is reported that education reflects the socioeconomic position and improved socioeconomic status generated by higher education and better working status of women and ends with healthier mother and child [25]. Notably, the proposed method is applicable over the categorical multicollinear data only; if data conditions vary, the performance of the proposed method may vary.

4. Conclusion

A comprehensive comparison of filter-based feature selection and categorical PLS loading weights in the frame work of regularized elimination in PLS is conducted. Monte Carlo-based simulation with 100 runs indicated that Cramer $V *$ VIP is the better choice for modeling the occurrence or nonoccurrence of stillbirths in terms of improved model performance, number of feature selection, and interpretation. Influential features which affect the occurrence of still birth covers the maternal socio, economic, and health facilitation-related features. The proposed method is applicable over the categorical multicollinear data only; if data conditions vary, the performance of the proposed method may vary. The framework can be used in related areas to explore and model health-related issues.

Data Availability

The data used to support the findings of the study are available from the corresponding author upon request.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] P. Bühlmann and S. Van De Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer Science & Business Media, Berlin, Germany, 2011.

[2] T. Mehmood, M. Sadiq, and M. Aslam, "Filter-based factor selection methods in partial least squares regression," *IEEE Access*, vol. 7, pp. 153499–153508, 2019.

[3] D. N. Reshef, Y. A. Reshef, H. K. Finucane et al., "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, 2011.

[4] T. Smitha and V. Sundaram, "Comparative study of data mining algorithms for high dimensional data analysis," *International Journal of Advances in Engineering & Technology*, vol. 4, p. 173, 2012.

[5] A. Ghaffar, S. Pongponich, N. Ghaffar, and T. Mehmood, "Factors associated with utilization of antenatal care services in balochistan province of Pakistan: an analysis of the multiple indicator cluster survey (mics) 2010," *Pakistan Journal of Medical Sciences*, vol. 31, pp. 1447–52, 2015.

[6] S. Pongpanich, A. Ghaffar, N. Ghaffar, and T. Mehmood, "Skilled birth attendance in balochistan, Pakistan," *Asian Biomedicine*, vol. 10, pp. 25–34, 2016.

[7] M. Sadiq, T. Mehmood, and M. Aslam, "Identifying the factors associated with cesarean section modeled with categorical correlation coefficients in partial least squares," *PLoS One*, vol. 14, 2019.

[8] R. E. Wright, *Logistic Regression, Reading and Understanding Multivariate Statistics*, American Psychological Association USA, Washington, DC, USA, 1995.

[9] D. Inan and B. E. Erdogan, "Liu-type logistic estimator," *Communications in Statistics - Simulation and Computation*, vol. 42, no. 7, pp. 1578–1586, 2013.

[10] H. Martens and T. Næs, "Multivariate calibration," *Chemometrics*, pp. 147–156, 1984.

[11] T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø, "A review of variable selection methods in partial least squares regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 118, pp. 62–69, 2012.

[12] T. Mehmood, H. Martens, S. Sæbø, J. Warringer, and L. Snipen, "A partial least squares based algorithm for parsimonious variable selection," *Algorithms for Molecular Biology*, vol. 6, p. 27, 2011.

[13] D. You, P. Bastian, J. Wu, and T. Wardlaw, *Levels and trends in child mortality: report 2013*, The World Bank, Washington, DC, USA, 2013.

[14] N. J. Kassebaum, A. Bertozzi-Villa, M. S. Coggeshall et al., "Global, regional, and national levels and causes of maternal mortality during 1990–2013: a systematic analysis for the global burden of disease study 2013," *The Lancet*, vol. 384, pp. 980–1004, 2014.

[15] M. Z. Zakar, R. Zakar, M. Mustafa, A. Jalil, and F. Fischer, "Underreporting of stillbirths in Pakistan: perspectives of the parents, community and healthcare providers," *BMC Pregnancy and Childbirth*, vol. 18, p. 302, 2018.

[16] H. Cramér, *Mathematical Methods Of Statistics (PMS-9)*, Vol. Vol. 9, Princeton university press, Princeton, NJ, USA, 2016.

[17] A. A. Tschuprow and M. Kantorowitsch, "Principles of the mathematical theory of correlation," Technical Report, William Hodge, Cambridge, UK, 1939.

[18] M. Friendly and S. Institute, *Visualizing Categorical Data*, Sas Institute, Cary, NC, USA, 2000.

[19] G. U. Yule, "On the methods of measuring association between two attributes," *Journal of the Royal Statistical Society*, vol. 75, no. 6, pp. 579–652, 1912.

[20] L. Eriksson, E. Johansson, N. Kettaneh-Wold, and S. Wold, *Multi-and Megavariate Data Analysis*, Umetrics Umeå, Umea, Sweden, 2001.

[21] O. M. Kvalheim and T. V. Karstang, "Interpretation of latent-variable regression models," *Chemometrics and Intelligent Laboratory Systems*, vol. 7, no. 1-2, pp. 39–51, 1989.

[22] T. N. Tran, N. L. Afanador, L. M. C. Buydens, and L. Blanchet, "Interpretation of variable importance in partial least squares with significance multivariate correlation (smc)," *Chemometrics and Intelligent Laboratory Systems*, vol. 138, pp. 153–160, 2014.

[23] A. Kols, Z. Gorar, M. Sharjeel et al., "Provincial differences in levels, trends, and determinants of childhood immunization in Pakistan," *Eastern Mediterranean Health Journal*, vol. 24, no. 4, pp. 333–344, 2018.

[24] J. Gardosi, V. Madurasinghe, M. Williams, A. Malik, and A. Francis, "Maternal and fetal risk factors for stillbirth: population based study," *Bmj*, vol. 346, no. 3, p. f108, 2013.

[25] I. K. Sørbye, C. Stoltenberg, J. Sundby, A. K. Daltveit, and S. Vangen, "Stillbirth and infant death among generations of pakistani immigrant descent: a population-based study," *Acta obstetricia et gynecologica Scandinavica*, vol. 93, pp. 168–174, 2014.