Research Article

Robust Group Identification and Variable Selection in Regression

Ali Alkenani and Tahir R. Dikheel

Department of Statistics, College of Administration and Economics, University of Al-Qadisiyah, Al Diwaniyah, Iraq

Correspondence should be addressed to Ali Alkenani; ali.alkenani@qu.edu.iq

Received 16 September 2017; Accepted 3 December 2017; Published 20 December 2017

Academic Editor: Aera Thavaneswaran

Copyright © 2017 Ali Alkenani and Tahir R. Dikheel. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The elimination of insignificant predictors and the combination of predictors with indistinguishable coefficients are the two issues raised in searching for the true model. Pairwise Absolute Clustering and Sparsity (PACS) achieves both goals. Unfortunately, PACS is sensitive to outliers due to its dependency on the least-squares loss function which is known to be very sensitive to unusual data. In this article, the sensitivity of PACS to outliers has been studied. Robust versions of PACS (RPACS) have been proposed by replacing the least squares and nonrobust weights in PACS with MM-estimation and robust weights depending on robust correlations instead of person correlation, respectively. A simulation study and two real data applications have been used to assess the effectiveness of the proposed methods.

1. Introduction

The latest developments in data aggregation have generated huge number of variables. The large amounts of data pose a challenge to most of the standard statistical methods. In many regression problems, the number of variables is huge. Moreover, many of these variables are irrelevant. Variable selection (VS) is the process of selecting significant variables for use in model construction. It is an important step in the statistical analysis. Statistical procedures for VS are characterized by improving the model's prediction, providing interpretable models while retaining computational efficiency. VS techniques, such as stepwise selection and best subset regression, may suffer from instability [1]. To tackle the instability problem, regularization methods have been used to carry out VS. They have become increasingly popular, as they supply a tool with which the VS is carried out during the process of estimating the coefficients in the model, for example, LASSO [2], SCAD [3], elastic-net [4], fused LASSO [5], adaptive LASSO [6], group LASSO [7], OSCAR [8], adaptive elastic-net [9], and MCP [10].

Searching for the correct model raises two matters: the exclusion of insignificant predictors and the combination of

predictors with indistinguishable coefficients (IC) [11]. The above approaches can remove insignificant predictors but be unsuccessful to merge predictors with IC. Pairwise Absolute Clustering and Sparsity (PACS, [11]) achieves both goals. Moreover, PACS is an oracle method for simultaneous group identification and VS.

Unfortunately, PACS is sensitive to outliers due to its dependency on the least-squares loss function which is known as very sensitive to unusual data. In this article, the sensitivity of PACS to outliers has been studied. Robust versions of PACS (RPACS) have been proposed by replacing the least squares and nonrobust weights in PACS with MM-estimation and robust weights depending on robust correlations instead of person correlation, respectively. RPACS can completely estimate the parameters of regression and select the significant predictors simultaneously, while being robust to the existence of possible outliers.

The rest of this article proceeds as follows. In Section 2, PACS has been briefly reviewed. The robust extension of PACS is detailed in Section 3. Simulation studies under different settings are presented in Section 4. In Section 5, the proposed robust PACS has been applied to two real datasets. Finally, a discussion concludes in Section 6.

2. A Brief Review of PACS

Under the linear regression model setup with standardized predictors x_{ij} and centered response values y_i , $i = 1, 2, \dots, N$ and $j = 1, 2, \dots, p$. Sharma et al. [11] proposed an oracle method PACS for simultaneous group identification and VS. PACS has less computational cost than OSCAR approach. In PACS, the equality of coefficients is attained by adding penalty to the pairwise differences and pairwise sums of coefficients. The PACS estimates are the minimizers of the following:

$$\sum_{i=1}^N \left(y_i - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \left\{ \sum_{j=1}^p \omega_j |\beta_j| + \sum_{1 \leq j < k \leq p} \omega_{jk(-)} |\beta_k - \beta_j| + \sum_{1 \leq j < k \leq p} \omega_{jk(+)} |\beta_k + \beta_j| \right\}, \quad (1)$$

where $\lambda \geq 0$ is the regularization parameter and ω is the nonnegative weights.

The penalty in (1) consists of $\lambda \{\sum_{j=1}^p \omega_j |\beta_j|\}$ that encourages sparseness, $\lambda \{\sum_{1 \leq j < k \leq p} \omega_{jk(-)} |\beta_k - \beta_j|\}$, and $\lambda \{\sum_{1 \leq j < k \leq p} \omega_{jk(+)} |\beta_k + \beta_j|\}$ that encourages equality of coefficients. The second term of the penalty encourages the same sign coefficients to be set as equal, while the third term encourages opposite sign coefficients to be set as equal in magnitude.

Choosing of appropriate adaptive weights is very important for PACS to be an oracle procedure. Consequently, Sharma et al. [11] suggested adaptive PACS that incorporate correlations into the weights which are given as follows:

$$\begin{aligned} \omega_j &= |\tilde{\beta}_j|^{-1}, \\ \omega_{jk(-)} &= (1 - r_{jk})^{-1} |\tilde{\beta}_k - \tilde{\beta}_j|^{-1}, \\ \omega_{jk(+)} &= (1 + r_{jk})^{-1} |\tilde{\beta}_k + \tilde{\beta}_j|^{-1} \end{aligned} \quad (2)$$

for $1 \leq j < k \leq p$,

where $\tilde{\beta}$ is $\sqrt{n}$ consistent estimator of β , such as the ordinary least squares (OLS) estimates or other shrinkage estimates like ridge regression estimates and r_{jk} is Pearson's correlation between the (j, k) th pair of predictors.

Sharma et al. [11] suggest using ridge estimates as initial estimates for β 's to obtain weights perform well in studies with collinear predictors.

3. Robust PACS

3.1. Methodology of Robust PACS. The satisfactory performance of PACS under normal errors has been demonstrated in [11]. However, the high sensitivity to outliers is the main drawback of PACS where a single outlier can change the good performance of PACS estimate completely.

Note that, in (1), the least-squares criterion is used between the predictors and the response. Also, the weighted

penalty contains weights which depend on Pearson's correlation in their calculations. However, the least-squares criterion and Pearson's correlation are not robust to outliers. To achieve the robustness in estimation and select the informative predictors robustly, the authors propose replacing the least-squares criterion with MM-estimation [12] where the MM-estimators are efficient and have high breakdown points. Moreover, the nonrobust weights replaced with robust weights depend on robust correlations such as the fast consistent high breakdown (FCH) [13], reweighted multivariate normal (RMVN) [13], Spearman's correlation (SP), and Kendall's correlation (KN). The RPACS estimates minimizing the following:

$$\begin{aligned} \sum_{i=1}^N \rho_1 \left(\frac{R_i(\beta)}{S_n} \right) + \lambda \left\{ \sum_{j=1}^p \text{Row}_j |\beta_j| + \sum_{1 \leq j < k \leq p} \text{Row}_{jk(-)} |\beta_k - \beta_j| + \sum_{1 \leq j < k \leq p} \text{Row}_{jk(+)} |\beta_k + \beta_j| \right\}, \end{aligned} \quad (3)$$

where $\lambda \geq 0$ is the regularization parameter and Row is the robust version of the nonnegative weights which are describes in (2). $R_i(\beta) = y_i - \sum_{j=1}^p x_{ij} \beta_j$, S_n is M-estimate of scale of the residuals, and it is defined as a solution of

$$\frac{1}{N} \sum_{i=1}^N \rho_0 \left(\frac{R_i}{S_n} \right) = K, \quad (4)$$

where K is a constant and ρ_0 function satisfies the following conditions:

- (1) ρ_0 is symmetric and continuously differentiable, and $\rho_0(0) = 0$.
- (2) There exist $a > 0$ such that ρ_0 is strictly increasing on $[0, a]$ and constant on $[a, \infty)$.
- (3) $K/\rho_0(a) = 1/2$.

The MM estimator in the first part of (3) is defined as an M-estimator of β using a redescending score function, $\psi(u) = \partial \rho_1(u)/\partial u$, and S_n obtained from (4). It is a solution to

$$\sum_{i=1}^N x_{ij} \psi \left(\frac{R_i(\beta)}{S_n} \right) = 0 \quad j = 1, 2, \dots, p, \quad (5)$$

where ρ_1 is another bounded ρ function such that $\rho_1 \leq \rho_0$.

3.2. Choosing the Robust Weights. The process of choosing the suitable weights is very important in order to obtain an oracle procedure [11]. The weights, which are described in (2), depend on Pearson's correlation in their calculations. From a practical point of view, it is well known that Pearson's correlation is not resistant to outliers and thus choosing weights in (2) based on this correlation will cause uncertain and deceptive results. Consequently, in order to get robust

weights, there is a need to estimate the correlation by using robust approaches. There are two types of robust versions for Pearson's correlation. The first type consists of those that are robust to the outliers, without interest in the general structure of the data, whereas the second type gives attention to the general structure of the data when dealing with outliers [14]. KN and MCD (minimum covariance determinant) are examples for the first and second types, respectively. Olive and Hawkins [13] proposed FCH and RMVN methods as practical consistent, outlier resistant estimators for multivariate location and dispersion. Alkenani and Yu [15] employed FCH and RMVN estimators instead of Pearson's correlation in the canonical correlation analysis (CCA) to obtain robust CCA. The authors showed that these estimators have good performance under different settings of outliers.

In this article, the FCH, RMVN, SP, and KN correlations have been employed instead of Pearson's correlation in order to obtain robust weights as follows:

$$\begin{aligned} Row_j &= |\check{\beta}_j|^{-1}, \\ Row_{jk(-)} &= (1 - Ror_{jk})^{-1} |\check{\beta}_k - \check{\beta}_j|^{-1}, \\ Row_{jk(+)} &= (1 + Ror_{jk})^{-1} |\check{\beta}_k + \check{\beta}_j|^{-1} \end{aligned} \quad (6)$$

for $1 \leq j < k \leq p$,

where Ror is a robust version of Pearson's correlation such as FCH, RMVN, SP, and KN correlations. $\check{\beta}$ is a robust initial estimate for β and we suggest using robust ridge estimates as initial estimates for $\check{\beta}$'s.

4. Simulation Study

In this section, five examples have been used to assess our proposed method RPACS by comparing it with PACS which is suggested in [11]. A regression model has been generated as follows:

$$y = X\beta + \epsilon \quad \epsilon \sim N(0, \sigma^2 I). \quad (7)$$

In all examples, predictors are standard normal. The distributions of the error term ϵ and the predictors are contaminated by two types of distributions, t distribution with 5 degrees of freedom ($t_{(5)}$) and Cauchy distribution with mean equal to 0 and variance equal to 1 (Cauchy (0, 1)). Also, different contamination ratios (5%, 10%, 15%, 20%, and 25%) were used. The performance of the methods is compared by using model error (ME) criterion for prediction accuracy which is defined by $(\hat{\beta} - \beta)'V(\hat{\beta} - \beta)$ where V represents the population covariance matrix of X . The sample sizes were 50 and 100 and the simulated model was replicated 1000 times.

Example 1. In this example, we choose the true parameters for the model of study as $\beta = (2, 2, 2, 0, 0, 0, 0, 0)^T$, $X \in \mathbb{R}^8$. The first three predictors are highly correlated with correlation equal to 0.7 and their coefficients are equal in magnitude, while the rest are uncorrelated.

Example 2. In this example, the true coefficients have been assumed as $\beta = (0.5, 1, 2, 0, 0, 0, 0, 0)^T$, $X \in \mathbb{R}^8$. The first three predictors are highly correlated with correlation equal to 0.7 and their coefficients differ in magnitude, while the rest are uncorrelated.

Example 3. In this example, the true parameters are $\beta = (1, 1, 1, 0.5, 1, 2, 0, 0, 0, 0)^T$, $X \in \mathbb{R}^{10}$. The first three predictors are highly correlated with correlation equal to 0.7 and their coefficients are equal in magnitude, while the second three predictors have lower correlation equal to 0.3 and different magnitudes. The rest of predictors are uncorrelated.

Example 4. In this example, true parameters are $\beta = (1, 1, 1, 0.5, 1, 2, 0, 0, 0, 0)^T$, $X \in \mathbb{R}^{10}$. The first three predictors are correlated with correlation equal to 0.3 and their coefficients are equal in magnitude, while the second three predictors have correlation equal to 0.7 and different magnitudes. The rest of predictors are uncorrelated.

Example 5. In this example, the true parameters are assumed as $\beta = (2, 2, 2, 1, 1, 0, 0, 0, 0, 0)^T$, $X \in \mathbb{R}^{10}$. The first three predictors are highly correlated with pairwise correlation equal to 0.7 and the second two predictors have pairwise correlation of 0.7, while the rest are uncorrelated. It can be observed that the groups of three and two highly correlated predictors have coefficients which are equal in magnitude.

To avoid repetition, the observations about the results in Tables 1–5 have been summarized as follows.

From Tables 1, 2, 3, 4, and 5, when there is no contamination data, PACS has good performance compared with our proposed methods. It is clear, when the contamination ratio of $t_{(5)}$ or Cauchy (0, 1) goes up the performance of PACS goes down while RPACS with all the robust weights has a stable performance, and the preference is for RPACS.RMVN and RPACS.RFCH, respectively, for all the samples sizes. The variations in ME values for the RPACS estimates with all the robust weights are close under different setting of contamination and sample sizes, and they are less than the variations of PACS estimates.

5. Analysis of Real Data

In this section, the RPACS methods with all the robust weights and PACS method have been applied in real data. The NCAA sports data from Mangold et al. [16] and the pollution data from McDonald and Schwing [17] have been studied.

The response variable was centered and the predictors were standardized. To verify RPACS, the two data sets have been analyzed by including outliers in the response variable and the predictors. The two data sets have been contaminated with (5%, 10%, 15%, and 20%) data from multivariate t distribution with three degrees of freedom.

To evaluate the estimation accuracy of the RPACS methods, the correlation between the estimated parameters according to the different methods under consideration and the estimated parameters from PACS without outliers, denoted as $\text{Corr}(\hat{\beta}, \hat{\beta}_{\text{PACS},0})$, was presented. Also, the effective

TABLE 1: ME results of Example 1.

Dist.	n	Outliers%	PACS	RPACS.KN	RPACS.SP	RPACS.FCH	RPACS.RMVN
$t_{(5)}$	50	0	0.02304	0.02964	0.03083	0.02979	0.02902
		5	0.20135	0.08124	0.08135	0.05575	0.04655
		10	0.25043	0.14048	0.14543	0.06579	0.05664
		15	0.30788	0.17578	0.18152	0.07225	0.06153
		20	0.34708	0.19266	0.21286	0.08195	0.06939
	100	25	0.40692	0.21584	0.22533	0.10242	0.08238
		0	0.02004	0.02644	0.02863	0.02772	0.02700
		5	0.19100	0.07100	0.08030	0.05111	0.04025
		10	0.23012	0.13011	0.14002	0.06116	0.05013
		15	0.28715	0.15523	0.17137	0.06899	0.05902
		20	0.32520	0.18670	0.19234	0.07115	0.06005
		25	0.36692	0.20522	0.21404	0.09032	0.07784
Cauchy (0, 1)	50	5	0.18112	0.07004	0.07237	0.04390	0.03581
		10	0.23263	0.12001	0.12273	0.05472	0.04454
		15	0.28368	0.15274	0.16138	0.06237	0.05079
		20	0.33511	0.17162	0.18556	0.07381	0.05848
		25	0.38488	0.19330	0.20405	0.09342	0.07211
	100	5	0.17214	0.06111	0.07335	0.04277	0.03581
		10	0.22263	0.11001	0.11273	0.04672	0.03854
		15	0.27368	0.14274	0.15138	0.05237	0.04079
		20	0.31511	0.16162	0.17556	0.06381	0.04848
		25	0.35488	0.18330	0.19405	0.08342	0.06211

TABLE 2: ME results of Example 2.

Dist.	N	Outliers%	PACS	RPACS.KN	RPACS.SP	RPACS.FCH	RPACS.RMVN
$t_{(5)}$	50	0	0.11372	0.12032	0.12151	0.12047	0.11970
		5	0.29201	0.17191	0.17203	0.14644	0.13725
		10	0.34113	0.23117	0.23611	0.15646	0.14730
		15	0.39857	0.26647	0.27221	0.16294	0.15222
		20	0.43778	0.28336	0.30355	0.17263	0.16006
	100	25	0.49761	0.30653	0.31602	0.19312	0.17308
		0	0.10354	0.11022	0.11131	0.10407	0.10050
		5	0.28171	0.16170	0.17100	0.14180	0.13094
		10	0.32082	0.22080	0.23072	0.15185	0.14082
		15	0.37783	0.24591	0.26205	0.15967	0.14970
		20	0.41560	0.27700	0.28300	0.16185	0.15071
		25	0.45762	0.29592	0.30473	0.18101	0.16854
Cauchy (0, 1)	50	5	0.27182	0.16072	0.16306	0.13460	0.12650
		10	0.32333	0.21071	0.21342	0.14541	0.13523
		15	0.37434	0.24340	0.25204	0.15303	0.14145
		20	0.42581	0.26232	0.27626	0.16451	0.14918
		25	0.47558	0.284	0.29475	0.18412	0.16281
	100	5	0.26282	0.15181	0.16405	0.13345	0.12651
		10	0.31331	0.20071	0.20343	0.13742	0.12923
		15	0.36435	0.23341	0.24205	0.14304	0.13149
		20	0.40581	0.25232	0.26625	0.15451	0.13916
		25	0.44557	0.27400	0.28473	0.17412	0.15281

TABLE 3: ME results of Example 3.

Dist.	N	Outliers%	PACS	RPACS.KN	RPACS.SP	RPACS.FCH	RPACS.RMVN
$t_{(5)}$	50	0	0.14172	0.14831	0.14950	0.14844	0.14743
		5	0.32001	0.19991	0.20003	0.17441	0.16522
		10	0.36913	0.25915	0.26411	0.18444	0.17530
		15	0.42653	0.29443	0.30021	0.19094	0.18022
		20	0.46576	0.31135	0.33154	0.20063	0.18806
	100	25	0.52561	0.33453	0.34402	0.22112	0.20107
		0	0.13042	0.13501	0.13645	0.13344	0.13255
		5	0.30971	0.18971	0.19901	0.16982	0.15894
		10	0.34882	0.24883	0.25872	0.17985	0.16882
		15	0.40582	0.27391	0.29003	0.18765	0.17774
		20	0.44365	0.30501	0.31103	0.18983	0.17871
		25	0.48562	0.32392	0.33271	0.20901	0.19650
Cauchy (0, 1)	50	5	0.29982	0.18872	0.19106	0.1626	0.1545
		10	0.35133	0.23871	0.24142	0.17341	0.16323
		15	0.40234	0.2714	0.28004	0.18103	0.16945
		20	0.45381	0.29032	0.30426	0.19251	0.17718
		25	0.50358	0.312	0.32275	0.21212	0.19081
	100	5	0.32001	0.19991	0.20003	0.17445	0.16525
		10	0.36913	0.25917	0.26411	0.18441	0.17536
		15	0.42655	0.29444	0.30021	0.19093	0.18022
		20	0.46575	0.31134	0.33153	0.20063	0.18804
		25	0.525610	0.33453	0.34401	0.22112	0.20106

TABLE 4: ME results of Example 4.

Dist.	N	Outliers%	PACS	RPACS.KN	RPACS.SP	RPACS.FCH	RPACS.RMVN
$t_{(5)}$	50	0	0.15251	0.15910	0.16035	0.15921	0.15823
		5	0.33081	0.21070	0.21082	0.18520	0.17601
		10	0.37991	0.26993	0.27491	0.19523	0.18612
		15	0.43732	0.30521	0.31101	0.20175	0.19102
		20	0.47653	0.32216	0.34233	0.21143	0.19887
	100	25	0.53641	0.34531	0.35482	0.23192	0.21185
		0	0.13342	0.13901	0.14125	0.13814	0.13713
		5	0.32051	0.20051	0.20981	0.18062	0.16973
		10	0.35962	0.25965	0.26952	0.19067	0.17962
		15	0.41662	0.28471	0.30083	0.19847	0.18853
		20	0.45446	0.31581	0.32183	0.20066	0.18951
		25	0.49642	0.33472	0.34351	0.21981	0.20757
Cauchy (0, 1)	50	5	0.31062	0.19952	0.20188	0.1734	0.16538
		10	0.36216	0.24951	0.25222	0.18421	0.17404
		15	0.41316	0.2822	0.29087	0.19184	0.18025
		20	0.46461	0.30112	0.31507	0.20331	0.18798
		25	0.51438	0.32284	0.33357	0.22294	0.20161
	100	5	0.33083	0.21071	0.21083	0.18525	0.17606
		10	0.37993	0.26995	0.27491	0.19521	0.18613
		15	0.43733	0.30522	0.31101	0.20175	0.19102
		20	0.47653	0.32217	0.34233	0.21143	0.19886
		25	0.53641	0.34533	0.354814	0.23192	0.21188

TABLE 5: ME results of Example 5.

Dist.	N	Outliers%	PACS	RPACS.KN	RPACS.SP	RPACS.FCH	RPACS.RMVN
$t_{(5)}$	50	0	0.06031	0.06695	0.06815	0.06701	0.06602
		5	0.23861	0.11851	0.11862	0.09305	0.08381
		10	0.28771	0.177735	0.182712	0.10303	0.09392
		15	0.34512	0.21301	0.21881	0.10955	0.09886
		20	0.38433	0.22996	0.25015	0.11923	0.10667
	100	25	0.44424	0.25315	0.26262	0.13972	0.11965
		0	0.04125	0.04684	0.04908	0.04597	0.04496
		5	0.22837	0.108313	0.11765	0.08846	0.07755
		10	0.26744	0.16745	0.17733	0.09846	0.08743
		15	0.32445	0.19256	0.20865	0.10627	0.09636
		20	0.36228	0.22365	0.22966	0.10844	0.09733
		25	0.40425	0.24257	0.25131	0.12761	0.11537
Cauchy (0, 1)	50	0	0.06031	0.06695	0.06815	0.06701	0.06602
		5	0.21845	0.10737	0.10963	0.08125	0.07316
		10	0.26997	0.15734	0.16006	0.09206	0.08183
		15	0.32095	0.19007	0.19865	0.09963	0.08806
		20	0.37244	0.20896	0.22289	0.11115	0.09579
	100	25	0.42217	0.23067	0.24135	0.13073	0.10948
		0	0.04125	0.04684	0.04908	0.04597	0.04496
		5	0.23865	0.11854	0.11865	0.09308	0.08389
		10	0.28775	0.17779	0.18274	0.10303	0.09397
		15	0.34513	0.21304	0.21885	0.10958	0.09885
		20	0.38435	0.22998	0.25015	0.11926	0.10667
		25	0.44423	0.25314	0.26261	0.13977	0.11967

model size after accounting for equality of absolute coefficient estimates has been reported.

5.1. NCAA Sports Data. The NCAA sport data is taken from a study of the effects of sociodemographic indicators and the sports programs on graduation rates. The dataset is available from the website (<http://www4.stat.ncsu.edu/~boos/var.select/ncaa.html>). The data size is $n = 94$ and $p = 19$ predictors. The response variable is the average of 6 year graduation rate for 1996–1999. The predictors are students in top 10% HS (X1), ACT COMPOSITE 25TH (X2), on living campus (X3), first-time undergraduates (X4), Total Enrollment/1000 (X5), courses taught by TAs (X6), composite of basketball ranking (X7), in-state tuition/1000 (X8), room and board/1000 (X9), avg BB home attendance (X10), Full Professor Salary (X11), student to faculty ratio (X12), white (X13), assistant professor salary (X14), population of city where located (X15), faculty with PHD (X16), acceptance rate (X17), receiving loans (X18), and out of state (X19).

5.2. Pollution Data (PD). The PD is taken from a study of the effects of different air pollution indicators and sociodemographic factors on mortality. The dataset is available from the website (<http://www4.stat.ncsu.edu/~boos/var.select/pollution.html>). The data contains $n = 60$ observations and $p = 15$ predictors. The response is the total Age

Adjusted Mortality Rate (y). The predictors are Mean annual precipitation (X1), mean January temperature (X2), mean July temperature (X3), % population that is 65 years of age or over (X4), population per household (X5), median school years (X6), % of housing with facilities (X7), population per square mile (X8), % of population that is nonwhite (X9), % employment in white-collar occupations (X10), % of families with income under 3; 000 (X11), relative population potential (RPP) of hydrocarbons (X12), RPP of oxides of nitrogen (X13), RPP of sulfur dioxide (X14), and % relative humidity (X15).

From Tables 6 and 7, we have the following findings in terms of estimation accuracy and the effective model size:

- (1) In case of no contamination, it can be observed that RPACS methods give comparable results as PACS. In addition, it can be seen that RPACS.RMVN and RPACS.FCH achieve better performance than RPACS.KN and RPACS.SP.
- (2) In case of contamination, the performance of PACS is dramatically affected. Also, it is obvious that RPACS.RMVN and RPACS.FCH methods give very consistent results, even with the high contamination percentages. The performance of RPACS.KN and RPACS.SP is less efficient than RPACS.RMVN and RPACS.FCH especially for all the contamination percentages.

TABLE 6: The $\text{Corr}(\hat{\beta}, \hat{\beta}_{\text{PACS},0})$ and the effective model size values for the methods under consideration based on the NCAA sport data.

Methods		Outliers%				
		0	5	10	15	20
$\text{Corr}(\hat{\beta}, \hat{\beta}_{\text{PACS},0})$	PACS	1	0.9033	0.8069	0.4112	0.1345
	RPACS.KN	0.9843	0.9839	0.9530	0.9019	0.8499
	RPACS.SP	0.9840	0.9837	0.9526	0.9006	0.8490
	RPACS.FCH	0.9850	0.9846	0.9843	0.9841	0.9839
	RPACS.RMVN	0.9856	0.9852	0.9850	0.9847	0.9845
The effective model size	PACS	5	6	7	9	10
	RPACS.KN	5	5	6	6	7
	RPACS.SP	5	5	6	6	7
	RPACS.FCH	5	5	5	5	5
	RPACS.RMVN	5	5	5	5	5

TABLE 7: The $\text{Corr}(\hat{\beta}, \hat{\beta}_{\text{PACS},0})$ and the effective model size values for the methods under consideration based on the pollution data.

Methods		Outliers%				
		0	5	10	15	20
$\text{Corr}(\hat{\beta}, \hat{\beta}_{\text{PACS},0})$	PACS	1	0.9247	0.8259	0.7001	0.5925
	RPACS.KN	0.9882	0.9866	0.9552	0.9044	0.8518
	RPACS.SP	0.9877	0.9862	0.9545	0.9038	0.8511
	RPACS.FCH	0.9890	0.9887	0.9884	0.9882	0.9879
	RPACS.RMVN	0.9897	0.9895	0.9893	0.9890	0.9888
The effective model size	PACS	5	6	6	8	9
	RPACS.KN	5	5	6	7	7
	RPACS.SP	5	5	6	7	7
	RPACS.FCH	5	5	5	5	5
	RPACS.RMVN	5	5	5	5	5

6. Conclusions

In this paper, robust consistent group identification and VS procedures have been proposed (RPACS) which combine the strength of both robust and identifying relevant groups and VS procedure. The simulation studies and analysis of real data demonstrate that RPACS methods have better predictive accuracy and identifying relevant groups than PACS when outliers exist in the response variable and the predictors. In general, the preference is for RPACS.RMVN and RPACS.RFCH, respectively, for all the samples sizes.

Abbreviations

LASSO:	Least absolute shrinkage and selection operator
PACS:	Pairwise Absolute Clustering and Sparsity
RPACS:	Robust Pairwise Absolute Clustering and Sparsity
VS:	Variable selection
SCAD:	Smoothly clipped absolute deviation
Fused LASSO:	Fused least absolute shrinkage and selection operator
Adaptive LASSO:	Adaptive least absolute shrinkage and selection operator

Group LASSO:	Group least absolute shrinkage and selection operator
OSCAR:	Octagonal shrinkage and clustering algorithm for regression
MCP:	Minimax concave penalty
IC:	Indistinguishable coefficients
FCH:	Fast consistent high breakdown
RMVN:	Reweighted multivariate normal
SP:	Spearman's correlation
KN:	Kendall's correlation
MCD:	Minimum covariance determinant
CCA:	Canonical correlation analysis
NCAA:	National Collegiate Athletic Association
PD:	Pollution data.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] L. Breiman, "Heuristics of instability and stabilization in model selection," *The Annals of Statistics*, vol. 24, no. 6, pp. 2350–2383, 1996.

- [2] R. Tibshirani, "Regression shrinkage and selection via the lasso: A retrospective," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 73, no. 3, pp. 273–282, 1996.
- [3] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.
- [4] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005.
- [5] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society B: Statistical Methodology*, vol. 67, no. 1, pp. 91–108, 2005.
- [6] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.
- [7] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.
- [8] H. D. Bondell and B. J. Reich, "Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR," *Biometrics*, vol. 64, no. 1, pp. 115–123, 2008.
- [9] H. Zou and H. H. Zhang, "On the adaptive elastic-net with a diverging number of parameters," *The Annals of Statistics*, vol. 37, no. 4, pp. 1733–1751, 2009.
- [10] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, vol. 38, no. 2, pp. 894–942, 2010.
- [11] D. B. Sharma, H. D. Bondell, and H. H. Zhang, "Consistent group identification and variable selection in regression with correlated predictors," *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 319–340, 2013.
- [12] V. c. Yohai, "High breakdown-point and high efficiency robust estimates for regression," *The Annals of Statistics*, vol. 15, no. 2, pp. 642–656, 1987.
- [13] D. J. Olive and D. M. Hawkins, "Robust multivariate location and dispersion," <http://lagrange.math.siu.edu/Olive/pphbml.pdf>, 2010.
- [14] R. Wilcox, *Introduction to robust estimation and hypothesis testing*, Statistical Modeling and Decision Science, Academic press, 2005.
- [15] A. Alkenani and K. Yu, "A comparative study for robust canonical correlation methods," *Journal of Statistical Computation and Simulation*, vol. 83, no. 4, pp. 690–720, 2013.
- [16] W. D. Mangold, L. Bean, and D. Adams, "The Impact of Inter-collegiate Athletics on Graduation Rates among Major NCAA Division I Universities: Implications for College Persistence Theory and Practice," *Journal of Higher Education*, vol. 74, no. 5, pp. 540–563, 2003.
- [17] G. C. McDonald and R. C. Schwing, "Instabilities of regression estimates relating air pollution to mortality," *Technometrics*, vol. 15, no. 3, pp. 463–481, 1973.

Hindawi

Submit your manuscripts at
<https://www.hindawi.com>