Kuwait J.Sci., Vol.50, No.(2A),April.2023,pp(1-17)

Comparison of fast regression algorithms in large datasets

Sengul Cangur[1]*, Handan Ankarali[2]
[1] *Dept. of Biostatistics and Medical Informatics, Duzce University, Turkey*
[2] *Dept. of Biostatistics and Medical Informatics, Istanbul Medeniyet University, Turkey*
**Corresponding author: sengulcangur@duzce.edu.tr*

Abstract

The aim is to compare the performances of fast regression methods, namely dimensional reduction of correlation matrix (DRCM), nonparametric dimensional reduction of correlation matrix (N-DRCM), variance inflation factor (VIF) regression, and robust VIF (R-VIF) regression in the presence of multicollinearity and outliers problems. In all simulation-scenarios, all the target variables were chosen for final models using four methods. The DRCM and N-DRCM are the methods that reach the final model in the shortest time, respectively. The time to reach the final model using R-VIF regression was approximately twice shorter than that of VIF regression. In each method, as the number of variables and the level of outliers increased, the time taken to reach the final model increased. When the level of multicollinearity and the number of variables ($p > 500$) increased, the times to reach the final models using DRCM in datasets with outliers were slightly shorter than the those of N-DRCM. The largest numbers of noise variables were selected to the model using DRCM and N-DRCM, but the least number of them were selected to the model using the R-VIF regression. The RMSE values obtained using DRCM, N-DRCM and VIF regression were similar in each scenario. As a result of the real dataset, the final model selected using R-VIF regression had the highest R^2. It also had the lowest RMSE value among those obtained with other approaches excluding VIF regression. As such, the R-VIF regression method demonstrated a better performance than the others in all datasets.

Keywords: Dimensional reduction; large data; robust; variance inflation factor

1. Introduction

In many fields, large data are studied, where the number of variables and observations is quite high. Through the development of modern technology, recording and storing information has become significantly easier. However, many researchers still experience issues in relation to accessing suitable information using datasets. Common issues include associated time-limit, theoretical, and costs among others. Researchers currently seek new approaches or algorithms that will allow them to access information quickly with minimal errors and few features. As such, algorithms that are easy to implement, can select the most suitable features for predictive statistical complex models, find solutions to frequently run into problems in modeling researches and application, and reach the final model quickly are being investigated. The most efficient approaches are becoming increasingly popular.

A review of current literature suggests the following algorithms are the ones most frequently used in relation to huge datasets especially high-dimensional datasets: least absolute shrinkage selection operator (LASSO) (Tibshirani, 1996), adaptive LASSO (Zou, 2006), elastic net (Zou & Hastie, 2005), least angle regression (LARS) (Efron *et al.*, 2004), robust LARS (Khan *et al.*, 2007), Dantzig (Candes & Tao, 2007), iterative sure independent screening (ISIS) (Fan & Lv, 2008), generalized path-seeking algorithm (GPS) (Friedman , 2008), forward-backward greedy algorithm (FoBa) (Zhang, 2009), variance inflation factor (VIF) regression (Lin *et al.*, 2011), robust variance inflation factor (R-VIF) regression (Dupuis

& Victoria-Feser, 2013), dimensional reduction of correlation matrix (DRCM) (Midi & Uraibi, 2014), jack-knife robust LARS (JKR-LARS) (Shahriari *et al.*, 2014), and VIF regression screening algorithm (VIFRegS) (Uraibi, 2020). Fast algorithms that meet the needs of researchers working with large datasets are currently being developed. Researches include the recently developed VIF regression method that has been used in health research (Liu *et al.*, 2017; Cai *et al.*, 2018), the DRCM method that claims to be faster and simpler than the VIF regression estimator, and the R-VIF regression method that can overcome issues including multicollinearity, overfitting, and outliers.

As previously noted, the assumptions of fast regression algorithms are not always met in dataset, or although it is claimed that some algorithms can overcome prominent issues in the cases that the severities of the problems and the number of variables increase, few studies have investigated how fast algorithms perform. As such, this simulation examines whether fast regression algorithms such as VIF regression, DRCM, R-VIF regression and nonparametric DRCM (N-DRCM) perform as well as current research suggests, especially in relation to the dataset containing multicollinearity and outliers. In addition, N-DRCM, which is the nonparametric version of DRCM, is discussed in this study. Whether this method can compete with others as a fast estimator is examined through implementing a multiple scenario simulation.

2. Methods

2.1 Variance inflation factor (VIF) regression method

The VIF regression is an approach developed from the streamwise variable selection algorithm with the α- investing rule. The streamwise algorithm ensures that the method implemented is fast, while the α-investment control is to prevent model overfitting. This method was improved using the sparsity assumption ($k \ll p$) when k is the subset of p predictors, and can control marginal false discovery rate-mFDR (Zhou *et al.*, 2006; Foster & Stine, 2008). Lin *et al.* (2011) improved this method as stepwise regression remained unresolved in relation to the multicollinearity problem. The regression model $\mathbf{y} = \beta_0 + \beta_1\mathbf{x}_1 + \cdots + \beta_k\mathbf{x}_k + \beta_{new}\mathbf{x}_{new} + \boldsymbol{\varepsilon}$ ($\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$ was tested to obtain the predictive regression model through forward selection. In this model, $\mathbf{y}$ is the dependent variable, $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are independent variables, $\beta_0, \ldots, \beta_k$ are regression coefficients, and $\boldsymbol{\varepsilon}$ is error. Here, $\mathbf{X} = [\mathbf{1}_n\ \mathbf{x}_1\ \ldots \mathbf{x}_k]$, $\widetilde{\mathbf{X}} = [\mathbf{X}\ \mathbf{x}_{new}]$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)^T$, and $\widetilde{\boldsymbol{\beta}} = (\beta_0, \ldots, \beta_k, \beta_{new})^T$. The algorithm of this method is shown in Algorithm 1.

Algorithm 1.

Input: data $\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2, \ldots$ (centered);
Set: $\alpha_0 = 0.50$, and pay-out $\Delta\alpha = 0.05$, and subsample size m;
Initialize $S = \{0\}$; $\mathbf{r} = \mathbf{y} - \widehat{\mathbf{y}} = \mathbf{r} = \mathbf{y} - \mathbf{X}_S\left(\mathbf{X}_S^T\mathbf{X}_S\right)^{-1}\mathbf{X}_S^T\mathbf{y}$;
$\widehat{\sigma} = \mathrm{sd}(\mathbf{y}) = \|\mathbf{r}\| / \sqrt{(n - |S| - 1)}$; $j = 1$; $\alpha_1 = \alpha_0$; $f = 0$.
Sample $I = \{j_1, \ldots, j_m\} \in \{1, \ldots, n\}$. // the subsample ${}_I\mathbf{x}$ randomly selected from predictors $\mathbf{x}$
Compute $\widetilde{\gamma}_{new} = \langle \mathbf{r}, \mathbf{x}_{new} \rangle / \|\mathbf{x}_{new}\|$ and
${}_IR^2 = \mathbf{x}_{newI}^{\mathrm{T}}\mathbf{X}_S\left({}_I\mathbf{X}_S^{\mathrm{T}}{}_I\mathbf{X}_S\right)^{-1}{}_I\mathbf{X}_S^{\mathrm{T}}\mathbf{x}_{new}/\|\mathbf{x}_{new}\|^2$.
repeat
set threshold $\alpha_j = \alpha_j$ / (1 + j - f)
get $\hat{t}_j = \widetilde{\gamma}_{new}/\widehat{\sigma}\sqrt{(1 - {}_IR^2)}$ // compute corrected t-statistic
if $2\Phi\left(|t_j|\right) > 1 - \alpha_j$ // compare p-value to threshold **then**
$S = S \cup \{j\}$ // add feature to model
update $\mathbf{r} = \mathbf{y} - \widehat{\mathbf{y}}_S$, $\widehat{\sigma} = \mathrm{RMSE}_S$
$\alpha_{j+1} = \alpha_j + \Delta\alpha$
$f = j$
else

$\alpha_{j+1} = \alpha_j - \alpha_j / \left(1 - \alpha_j\right)$
end if
$j = j + 1$
until maximum CPU time or memory is reached.

α_0: the initial alpha-wealth according to α-investing rule, $\Delta\alpha$: if a hypothesis is rejected, the change of alpha-wealth value, **r**: residuals, S: the set of predictors, α_j: α value in the jth test, sd: standard deviation, f: the time at which the last hypothesis is rejected, I: subsample, Φ: the standard normal cumulative distribution, RMSE: root mean squared error, CPU: central processing unit

This method contains two components: evaluation and search. The evaluation step contains forward stagewise regression and evaluates variables using marginal correlations. The stagewise regression algorithm contains small step sizes and behaves similarly to l_1 algorithms such as Lasso and LARS. As such, it suffers from collinearities between the predictors. Lin *et al.* (2011) corrected this bias by selecting a small sample from the dataset to calculate the VIF of each variable. The resultant evaluation phase is fast and contains no significant loss of accuracy. In the search step, each variable is sequentially tested using the α-investing rule. This rule ensures that models do not overfit and can generate highly accurate results. VIF procedure can be combined with various algorithms such as stepwise regression, LARS, and FoBa. This algorithm is particularly useful when feature systems are created dynamically and the size of the candidate features collection is unknown or even infinite. It can also serve as an "online" algorithm for loading extremely large-scale data into RAM according to its properties (Lin *et al.*, 2011).

2.2 Robust VIF regression method

Robust VIF regression method is developed by Dupuis & Victoria-Feser (2013) as the classical VIF regression method can be adversely affected by outliers in the dataset. It contains all properties of the classic approach. Dupuis & Victoria-Feser (2013) used the robust weighted slope estimator and the fast robust t-statistic in this method. Therefore, this method is very robust against small model deviations. The R-VIF regression procedure, which is based on a streamwise variable selection algorithm and the α-investing rule, is shown in Algorithm 2.

Algorithm 2.

Input: data $\mathbf{y}$, $\mathbf{x}_1$, $\mathbf{x}_2$, . . . (standardized);
Set: initial wealth $\alpha_0 = 0.50$, and pay-out $\Delta\alpha$ =0.05, and subsample size m, and robustness constant c
Compute efficiency e_c^{-1} where e_c is as in

$$e_c = \left[\int_{-c}^{c}\left(5\left(\frac{r}{c}\right)^4 - 6\left(\frac{r}{c}\right)^2 + 1\right)d\Phi(r)\right]^2 / \int_{-c}^{c} r^2\left(\left(\frac{r}{c}\right)^2 - 1\right)^4 d\Phi(r)$$

Get all marginal weights w_{ij} by fitting p marginal models $y = \beta_{01} + \beta_{1x_1} + \varepsilon_1, \ldots, y = \beta_{0k} + \beta_k x_k + \varepsilon_k$ using $\sum_{i=1}^{n} w_i(r_i;c) r_i x_i = 0$ and $w_i(r_i;c) = min\left\{1;\ \frac{c}{|r_i|}\right\}$ (c=1.345)
Initialize $j = 1, S = \{0\}$, $\mathbf{X}_S = 1$, $\mathbf{X}_S^w = \text{diag}\left(\sqrt{w_{iS}^0}\right)\mathbf{X}_S$ and $\mathbf{y}^w = \text{diag}\left(\sqrt{w_{iS}^0}\right)\mathbf{y}$ where w_{iS}^0 is computed using

$$w_i(r_i;c) = \begin{cases} \left(\left(\frac{r_i}{c}\right)^2 - 1\right)^2 & if\ |r_i| \le c, \\ 0 & if\ |r_i| > c, \end{cases}$$

where $\mathbf{r}^0 = \left(\mathbf{y} - 1\widehat{\beta}^0\right)/\widehat{\sigma}^0$ using $\mathbf{X}_0^w = \mathbf{X}_0^{w2} = 1$, $\widehat{\beta}^0 = \left[\left(\mathbf{X}_0^w\right)^{\mathrm{T}}\mathbf{X}_0^w\right]^{-1}\left(\mathbf{X}_0^{w2}\right)^{\mathrm{T}}\mathbf{y}$,

$\widehat{\sigma}^0 = 1.483 med\left|\widetilde{\mathbf{r}}^0 - med(\widetilde{\mathbf{r}}^0)\right|$ and $\widetilde{\mathbf{r}}^0 = \mathbf{y} - 1\widehat{\beta}^0$.

repeat

set $\alpha_j = \alpha_j/(1+j-f)$

compute $\mathbf{r}_S^w = \mathbf{y}^w - \mathbf{X}_S^w\left(\mathbf{X}_S^{wT}\mathbf{X}_S^w\right)^{-1}\mathbf{X}_S^{wT}\mathbf{y}^w$ //start *Fast Robust Evaluation Procedure*

$\widehat{\gamma}_j^w = \left(\mathbf{z}_j^{wT}\mathbf{z}_j^w\right)^{-1}\mathbf{z}_j^{wT}\mathbf{r}_S^w$ and $\widehat{\sigma} = MAD\left(\mathbf{r}_S^w - \mathbf{z}_j^w\left(\mathbf{z}_j^{wT}\mathbf{z}_j^w\right)^{-1}\mathbf{z}_j^{wT}\mathbf{r}_S^w\right)$

where $\mathbf{z}_j^w = \mathrm{diag}\left(\sqrt{w_{ij}}\right)\mathbf{z}_j$

sample $I = \{i_1,\ldots,i_m\} \in \{1,\ldots,n\}$ // the subsample ${}_I\mathbf{x}$ randomly selected from predictors

get $R_{jS}^{w2} = {}_I\mathbf{z}_j^{wT}\,{}_I\mathbf{H}_S^w\,{}_I\mathbf{z}_j^w\left({}_I\mathbf{z}_j^{wT}\,{}_I\mathbf{z}_j^w\right)^{-1}$ // a robust R^2 coefficient

where ${}_I\mathbf{H}_S^w = {}_I\mathbf{X}_S^w\left({}_I\mathbf{X}_S^{wT}\,{}_I\mathbf{X}_S^w\right)^{-1}\,{}_I\mathbf{X}_S^{wT}$, and find $\rho^w = 1 - R_{jS}^{w2}$

get $T_w = (\rho^w)^{-1/2}\widehat{\gamma}_j^w / \sqrt{\widehat{\sigma}^2\left(\sum_i z_{ij}^{w2}\right)^{-1} e_c^{-1}}$ from *Fast Robust Evaluation Procedure*

//compute the approximate robust t-statistic

if $2(1-\Phi(T_w)) < \alpha_j$ **then**

$S = S \cup \{j\}$, $\mathbf{X}_S = \begin{bmatrix}\mathbf{1} & \mathbf{x}_j\end{bmatrix}$, $\mathbf{X}_S^w = diag\left(\sqrt{w_{iS}^0}\right)\mathbf{X}_S$, and $\mathbf{y}^w = diag\left(\sqrt{w_{iS}^0}\right)\mathbf{y}$,

where w_{iS}^0 is computed using $w_i(r_i;c) = \begin{cases} \left(\left(\frac{r_i}{c}\right)^2 - 1\right)^2 & if\ |r_i| \le c, \\ 0 & if\ |r_i| > c, \end{cases}$

where $\mathbf{r}^0 = \left(\mathbf{y} - \mathbf{X}_S\widehat{\beta}^0\right)/\widehat{\sigma}^0$ using $\mathbf{X}_0^w = \begin{bmatrix}1 & \sqrt{w_{ij}}\,x_{ij}\end{bmatrix}$, $\mathbf{X}_0^{w2} = \begin{bmatrix}1 & w_{ij}x_{ij}\end{bmatrix}$, i=1,…,n,

$\widehat{\beta}^0 = \left[\left(\mathbf{X}_0^w\right)^{\mathrm{T}}\mathbf{X}_0^w\right]^{-1}\left(\mathbf{X}_0^{w2}\right)^{\mathrm{T}}\mathbf{y}$,

where $\widehat{\sigma}^0 = 1.483 med\left|\widetilde{\mathbf{r}}^0 - med(\widetilde{\mathbf{r}}^0)\right|$ and $\widetilde{\mathbf{r}}^0 = \mathbf{y} - \mathbf{X}_S\widehat{\beta}^0$

$\alpha_{j+1} = \alpha_j + \Delta\alpha$

$f = j$

else $\alpha_{j+1} = \alpha_j - \alpha_j/\left(1-\alpha_j\right)$

end if

$j = j+1$

until all p covariates have been considered.

α_0: the initial alpha-wealth according to α-investing rule, $\Delta\alpha$: if a hypothesis is rejected, the change of alpha-wealth value, r and $\mathbf{r}$: residuals, S: the set of predictors, α_j: α value in the jth test, c = 4.685, w_i: Tukey's biweight weights, r_i: standardized residuals, Φ: the standard normal cumulative distribution, med: median, MAD: median absolute deviation, diag: diagonal, R_{jS}^{w2}: a robust R^2 coefficient proposed by Renaud and Victoria-Feser (2010)

2.3 Dimensional reduction of correlation matrix (DRCM) method

The DRCM method was suggested by Midi & Uraibi (2014). This method can reduce the time for selecting only the variables which provide important information to the response variable. The procedure consists of two steps: in the first step, DRCM tries to reduce the dimension of correlation matrix by including only those variables that have absolute correlations greater than a threshold value, in the potential model. In the second step, the p-values for the parameter estimates of potential model were computed using multiple linear regression method. The final regression model only includes those variables that are significant. The algorithm of this method, which is based on the regression method, is

shown in Algorithm 3.

Algorithm 3.

Input: data $\mathbf{y}$, $\mathbf{x}_1$, $\mathbf{x}_2$, . . . (standardized);
Initialize $S_1 = \{0\}$, $S = \{0\}$, $j = 1$,
$Cos(\theta_{\mathbf{x},\mathbf{y}}) = \frac{<\mathbf{x},\ \mathbf{y}>}{\|\mathbf{x}\|.\|\mathbf{y}\|} = \frac{Cov(\mathbf{x},\ \mathbf{y})}{\sqrt{Var(\mathbf{x})Var(\mathbf{y})}} = Corr(\mathbf{x}, \mathbf{y})$,
$\widehat{\beta}=\left[\mathbf{X}^{\mathrm{T}}\mathbf{X}\right]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$, $\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{y}=\frac{1}{n}\widehat{\beta}=R_{\mathbf{x}y}$, $\mathbf{r}=\mathbf{y}-\overline{\mathbf{y}}=\mathbf{r}=\mathbf{y}-\mathbf{X}\left[\mathbf{X}^{\mathrm{T}}\mathbf{X}\right]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$,
$Cov(\widehat{\beta}) = \sigma^2\left[\mathbf{X}^{\mathrm{T}}\mathbf{X}\right]^{-1}$, $\widehat{\sigma}^2 = \mathbf{y}^{\mathrm{T}}\mathbf{y}-\widehat{\beta}^{\mathbf{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y}$=MSE. // from the linear regression model $\mathbf{y} = \mathbf{x}\beta + \varepsilon$
Compute $Cos(\theta_{\mathbf{x},\mathbf{y}}) = \frac{<\mathbf{x},\ \mathbf{y}>}{\|\mathbf{x}\|.\|\mathbf{y}\|} = \frac{Cov(\mathbf{x},\ \mathbf{y})}{\sqrt{Var(\mathbf{x})Var(\mathbf{y})}} = Corr(\mathbf{x}, \mathbf{y})$ // First step
$\widehat{\beta}=\left[\mathbf{X}^{\mathrm{T}}\mathbf{X}\right]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$, $\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{y}=\frac{1}{n}\widehat{\beta}=R_{xy}$ // R_{xy} is the correlation between $\mathbf{x}$ and $\mathbf{y}$
// The value of $\left|R_{xy}\right|$ is between 0 and 1.
where $\mathbf{X}^{\mathrm{T}}\mathbf{X}=\mathbf{I}$
set threshold $M = \frac{\sum_{j=1}^{p}\left|R_{xy}\right|}{p}$ // Pearson correlation matrix $R_{\mathbf{x}y}$; the number of all candidate covariates p
if $\left|\widehat{\beta}\right| = \left|R_{\mathbf{x}y}\right| \geq M$
compare *Corr(***x**,**y***)* -values to threshold
// The dimension of the correlation matrix is reduced
then
$S_1 = S_1 \cup \{j\}$ // add candidate feature for model
end if
$j = j + 1$
until all p covariates have been considered.
// Second step
set $\alpha_j = \alpha_j$ //(1+*j*-*f*), $S_1 = \{0\}$, $f = j$
get $\hat{t} = \widehat{\beta}_j/(\widehat{\sigma}^2(\left[\mathbf{X}^{\mathrm{T}}\mathbf{X}\right]^{-1})$ // compute t-statistic
if $2\left(1-\Phi\left(\hat{t}\right)\right) < \alpha_j$ // compare p-value
then
$S = S \cup \{j\}$ // add feature from S_1 to model
else
$\alpha_{j+1} = \alpha_j - \alpha_j/\left(1-\alpha_j\right)$
end if
$j = j + 1$
until all covariates in S_1 have been considered.

S_1: the set of candidate predictors in first step, S: the set of predictors, $\left|R_{\mathbf{x}y}\right|$: the absolute values of correlation matrix, Φ: the standard normal cumulative distribution, f: the time at which the last hypothesis is rejected, α_j: α value in the jth test

2.4 Nonparametric DRCM (N-DRCM) method

The N-DRCM method is a nonparametric version of the DRCM method. The procedure consists of two steps. In the first step, Spearman correlation matrix is used to determine monotonic relationship between variables. These variables can be continuous or at least one of them can be ordinal. N-DRCM tries to reduce the dimension of correlation matrix by including only those variables that have absolute correlations greater than a threshold value, in the potential model. In the second step, the p-values for

the parameter estimates of potential model are computed by robust regression method using iteratively reweighted least squares (IRLS). The final regression model only includes those variables that are significant. This algorithm is shown in Algorithm 4.

Algorithm 4.

Input: data $\mathbf{y}$, $\mathbf{x}_1$, $\mathbf{x}_2$, . . .(standardized);
Initialize $S_1 = \{0\}$, $S = \{0\}$, $j = 1$,

$${}_{SP}Cos\left(\theta_{({}^{r}\mathbf{x}{}^{r}\mathbf{y})}\right) = \frac{Cov({}^{r}\mathbf{x},{}^{r}\mathbf{y})}{\sqrt{Var({}^{r}\mathbf{x})Var({}^{r}\mathbf{y})}} = {}_{SP}Corr({}^{r}\mathbf{x},{}^{r}\mathbf{y})$$

// Spearman correlation matrix for the ranked data ${}^{r}\mathbf{y}$, ${}^{r}\mathbf{x}_1$, ${}^{r}\mathbf{x}_2$,...

$${}^{r}\widehat{\beta} = \left[{}^{r}\mathbf{X}^{\mathrm{T}}{}^{r}\mathbf{X}\right]^{-1}{}^{r}\mathbf{X}^{\mathrm{T}}{}^{r}\mathbf{y}\,,\ \frac{1}{n}{}^{r}\mathbf{X}^{\mathrm{T}}{}^{r}\mathbf{y} = \frac{1}{n}{}^{r}\widehat{\beta}^{\mathrm{T}} = R_{({}^{r}\mathbf{x},{}^{r}\mathbf{y})},$$

$${}^{r}\mathbf{r} = {}^{r}\mathbf{y} - {}^{r}\overline{\mathbf{y}} = \mathbf{r} = {}^{r}\mathbf{y} - {}^{r}\mathbf{X}\left[{}^{r}\mathbf{X}^{\mathrm{T}}{}^{r}\mathbf{X}\right]^{-1}{}^{r}\mathbf{X}^{\mathrm{T}}{}^{r}\mathbf{y},$$

$$Cov({}^{r}\widehat{\beta}) = {}^{r}\sigma^{2}\left[{}^{r}\mathbf{X}^{\mathrm{T}}{}^{r}\mathbf{X}\right]^{-1},\ {}^{r}\widehat{\sigma}^{2} = {}^{r}\mathbf{y}^{\mathrm{T}}{}^{r}\mathbf{y} - {}^{r}\beta^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \text{MSE}.$$

// from the linear regression model $\mathbf{y} = \mathbf{x}\beta + \varepsilon$

Compute ${}_{SP}Corr({}^{r}\mathbf{x},{}^{r}\mathbf{y}) \frac{Cov({}^{r}\mathbf{x},{}^{r}\mathbf{y})}{\sqrt{Var({}^{r}\mathbf{x})Var({}^{r}\mathbf{y})}} = 1 - \frac{6\sum d_i^2}{\sqrt{Var({}^{r}\mathbf{x})Var({}^{r}\mathbf{y}))}}$ // First step

// Spearman correlation matrix

where $\sqrt{Var({}^{r}\mathbf{x})\,Var({}^{r}\mathbf{y})} = \begin{cases} n(n^2-1) & if \text{ all } n \text{ ranks are distinct integers,} \\ (n^2-1)/12 & if \text{ all ranks are distinct,} \end{cases}$

$${}^{r}\widehat{\beta} = \left[{}^{r}\mathbf{X}^{\mathrm{T}}{}^{r}\mathbf{X}\right]^{-1}{}^{r}\mathbf{X}^{\mathrm{T}}{}^{r}\mathbf{y}\,,\ \frac{1}{n}{}^{r}\mathbf{X}^{\mathrm{T}}{}^{r}\mathbf{y} = \frac{1}{n}{}^{r}\widehat{\beta}^{\mathrm{T}} = R_{({}^{r}\mathbf{x},{}^{r}\mathbf{y})}$$

${}_{SP}R_{xy}$ is the Spearman correlation between the ranked $\mathbf{x}$ and $\mathbf{y}$

// The value of $\left|{}_{SP}R_{({}^{r}\mathbf{x},{}^{r}\mathbf{y})}\right|$ is between 0 and 1.

where $\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{I}$

set threshold $M = \frac{\sum_{j=1}^{p}\left|{}_{SP}R_{({}^{r}\mathbf{x},{}^{r}\mathbf{y})}\right|}{p}$ // Spearman correlation matrix ${}_{SP}R_{({}^{r}\mathbf{x},{}^{r}\mathbf{y})}$;

// The number of all candidate covariates p

if $\left|{}_{SP}R_{({}^{r}\mathbf{x},{}^{r}\mathbf{y})}\right| \geq M$

compare ${}_{SP}Corr({}^{r}\mathbf{x},{}^{r}\mathbf{y})$ -values to threshold

// The dimension of the correlation matrix is reduced

then

$S_1 = S_1 \cup \{j\}$ // add candidate feature for model

end if

$j = j + 1$

until all p covariates have been considered.

// Second step

Compute $\underset{\beta}{min} \sum_{j=1}^{n} \rho\left(\frac{y_j - x_j^{\mathrm{T}}\beta}{\sigma}\right)$ // minimize β's using the standardized data from the linear model $\mathbf{y} = \mathbf{x}\beta + \varepsilon$

$\sum_{j=1}^{n} x_{ij}\psi\left(\frac{y_j - x_j^{\mathrm{T}}\beta}{\sigma}\right) = 0$ for all i=0,1,2,. . . ,p // solution using nonlinear optimization method – Iteratively reweighted least squares (IRLS)

where $\psi = \rho^{\mathrm{T}}$, $x_{i0} = 1$, $\sigma = \widehat{\sigma}^{0} = 1.483 med\left|\left(y_j - x_j\widehat{\beta}_0\right) - med(y_j - x_j\widehat{\beta}_0)\right|$,

$$\beta_{t+1} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{w}_t\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{w}_t$$

where $w_{jt} = \begin{cases} \frac{\psi\left[\left(y_j - x^{\mathrm{T}}\beta_{jt}\right)/\sigma_t\right]}{\left(y_j - x^{\mathrm{T}}\beta_{jt}\right)/\sigma_t} & if\ y_j \neq x^{\mathrm{T}}\beta_{jt} \\ 1 & if\ y_j = x^{\mathrm{T}}\beta_{jt} \end{cases}$

$w(u) = min\begin{cases} 1 & if\ |u| < 0 \\ \frac{c}{|u|} & if\ |u| \geq 0 \end{cases}$ // Huber' method (c=1.345)

$Cov\left(\widehat{\beta}\right) = \sigma^2 \frac{\sum_{i=1}^{n} \psi^2\left[\left(y_i - x_j^{\mathrm{T}}\beta\right)/\sigma\right]}{\left\{\sum_{i=1}^{n} \psi^{\mathrm{T}}\left[\left(y_i - x_j^{\mathrm{T}}\beta\right)/\sigma\right]\right\}^2}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}, Var\left(\widehat{\beta}\right) = \widehat{\sigma}^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{w}_t\mathbf{X}\right)^{-1}$
set $\alpha_j = \alpha_j$ /(1+*j*-*f*), $S_1 = \{0\}$, f $f = j$
get $\hat{t}_w = \left(\mathbf{X}^{\mathrm{T}}\mathbf{w}_t\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{w}_t\mathbf{y} \,/\, \sqrt{\widehat{\sigma}^2(\mathbf{X}^{\mathrm{T}}\mathbf{w}_t\mathbf{X})^{-1}}$ // compute the robust *t*-statistic
if $2\left(1 - \Phi\left(\hat{t}_w\right)\right) < \alpha_j$ // compare *p*-value
then
$S = S \cup \{j\}$ // add feature from S_1 to model
else
$\alpha_{j+1} = \alpha_j - \alpha_j/\left(1 - \alpha_j\right)$
end if
$j = j + 1$
until all covariates in S_1 have been considered.

S_1: the set of candidate predictors in first step, S: the set of predictors, d_i: difference in paired ranks, $\mathbf{w}_t$: diagonal matrix of weights, $\rho(.)$: likelihood function for a suitable choice of the distribution of the residuals, Φ: the standard normal cumulative distribution, f: the time at which the last hypothesis is rejected, α_j: α value in the *j*th test, ψ: influence function

2.5 Simulation study

This simulation study has been designed in a similar way to studies conducted by Rahman & Khan (2010) and Dupuis & Victoria-Feser (2013). A linear model was established as

$$y = x_1 + x_2 + \ldots + x_k + \sigma\varepsilon_j \tag{1}$$

where $x_1,\ x_2,\ldots,x_k$ are multivariate normal variables with $E\left(x_j\right) = 0$, $Var\left(x_j\right) = 1$, and $corr\left(x_j, x_i\right) = \theta$ $(i \neq j,\ \ i, j = 1,\ldots,k)$. θ is chosen to produce a range of theoretical $R^2 = \left(\mathrm{Var}\,(\mathrm{y}) - \sigma^2\right)/\mathrm{Var}\,(\mathrm{y})$ values for (1) and σ to give t values for target covariates of about 5-6 under normality. $x_1,\ x_2,\ldots,x_k$ represent k target covariates. ε is an independent standard normal variable. A set of p predictors was generated as follows:

$$x_{k+1} = x_1 + \delta e_{k+1}$$
$$x_{k+2} = x_1 + \delta e_{k+2}$$
$$\vdots$$
$$x_{3k} = x_k + \delta e_{3k} \tag{2}$$

Variables $x_{k+1},\ x_{k+2},\ \ldots, x_{3k}$ were noise covariates that correlated with target covariates. Variables $x_{3k+1},\ ,\ \ldots, x_p$ were the noise covariates that did not correlate with the target covariates ($x_j = e_j,\ j = 3k + 1,\ 3k + 2,\ldots,p$). $e_{k+1},\ldots,e_p$ were independent standard normal variables. In each scenario, the number of target covariates was set as five. The constant $\delta = 3.18$ was selected so that $corr\,(x_1,\ x_{k+1}) = corr\,(x_1, x_{k+2}) = \ldots = corr\,(x_k,\ x_{3k}) = 0.3$. The estimated final model was given in equation 3.

$$\mathbf{y} = \beta_0 + \beta_1\mathbf{x}_1 + \cdots + \beta_k\mathbf{x}_k + \beta_{new}\mathbf{x}_{new} + \varepsilon \tag{3}$$

The datasets consisted of "normal (no contamination)" and "outliers (with 5% and 10%)" to examine the effect outliers had on datasets. The datasets were generated using $\varepsilon \sim \mathrm{N}(0,\ 1)$ for normal data, $\varepsilon \sim 95\%\mathrm{N}(0,\ 1) + 5\%\mathrm{N}(30,\ 1)$ for the dataset with 5% outliers and $\varepsilon \sim 90\%\mathrm{N}(0,\ 1) + 10\%\mathrm{N}(30,\ 1)$ for the dataset with 10% outliers. To examine the effect of multicollinearity in datasets, correlations

among target regressors were specified as $\theta_1 = 0.1$ $\left(R^2 = 0.20\right)$ and $\theta_2 = 0.85$ $\left(R^2 = 0.80\right)$ so that $corr\left(x_j, x_i\right) = \theta, \quad (i \neq j \quad i,j = 1,\ldots,k)$. A total of 36 scenarios were created through combining different data types, including the uncontaminated dataset and the datasets with 5% and 10% outliers, with 50, 100, 250, 500, 750, and 1,000 independent variables. The sample size was 5,000 and the number of repetition was 100. A total of 14,400 models were examined. The initial-wealth and pay-out were respectively selected 0.5 and 0.05 for VIF and R-VIF regression methods. In each condition, the root mean square error (RMSE) values calculated through the four methods were recorded. This simulation was executed using the MATLAB/Simulink R2015a program (toolboxes: statistics and machine learning, curve fitting, optimization, and global optimization) by a computer with Intel(R) Core(TM) i7-6500U CPU @ 2.50 GHz, 2592 Mhz, two cores, and four logical processors.

2.6 Real data

Crime dataset taken from UCI Machine Learning Respiratory (Redmond, 2009) was used to compare the performances of DRCM, N-DRCM, VIF regression and R-VIF regression methods. This dataset consists socio-economic data from the 1990 US Census, law enforcement data from the 1990 US Law Enforcement Management and Administrative Statistics (LEMAS) survey, and crime data from the 1995 Federal Bureau of Investigation' Uniform Crime Reporting (FBI UCR). Crime dataset includes $n = 1994$ observations, the violent crime per capita variable ($\mathbf{y}$), and 122 predictors ($\mathbf{x}$) that have a possible relationship with crime in order to estimate ($\mathbf{y}$). The RMSE, R^2 and estimation values (beta, standard error, t-statistic, and p-value) of the final models selected using each method were calculated.

3. Results

3.1 Simulation

In case presences of multicollinearity and outliers, while the number of candidate covariates that can be included in the model increased, the values (average time, average numbers of covariates with different relationships) that show the performances of DRCM, N-DRCM, VIF regression and R-VIF regression methods are demonstrated in Table 1, Table 2, and Table 3, respectively.

In all scenarios, all the target independent variables were selected to the final models by four methods. Respectively, the DRCM and N-DRCM methods reached the final model in the shortest time.The plots of average times taken to reach the final models for fast regression methods in datasets with outliers for each theta value were given Fig. 1, respectively. When the number of variables was 250 or less in datasets with 5% and 10% outliers, the times taken to reach the final models for both DRCM and N-DRCM were similar. However, when theta value was 0.10 and the number of variables was 500 or more, the times taken to reach the final models in datasets with 10% outliers were significantly longer than the those of DRCM and N-DRCM in datasets with 5% outliers. Moreover, when theta value was 0.85 and the number of variables was 500 or more, the times to reach the final models using DRCM in datasets with outliers were slightly shorter than those of N-DRCM. When the number of variables was over 750 in both datasets with outliers, the times to reach the final models decreased in line with increasing theta values for both DCRM and N-DCRM. The decrement amount increased as the level of outliers increased. However, this was not observed in the R-VIF and VIF regression methods. The time to reach the final model using R-VIF regression was approximately two times shorter than that of VIF regression. The largest numbers of noise variables were selected to the final models using DRCM and N-DRCM methods.

The RMSE values obtained using DRCM, N-DRCM and VIF regression were similar in each scenario. The RMSE values calculated by each method were higher in the datasets with outliers compared to uncontaminated datasets. In addition, the RMSE values tended to decrease when the number of variables increased. This conclusion applies to the RMSE values obtained using R-VIF regression, except for when the number of variables in datasets with 10% outliers was 500 or above. When the number of variables in datasets with 10% outliers was 500 or above, the RMSE values obtained using R-VIF regression were lower than the values obtained in uncontaminated datasets.

When the number of variables was 500 or above, the approximate ratios of total noise variables chosen for the final models using DRCM and N-DRCM methods were found to be 2.1% in uncontaminated datasets, 2.3% in datasets with 5% outliers, and 4.1% in datasets with 10% outliers.In addition, when the number of variables was 500 or above, the approximate ratios of total noise variables chosen for the final model using R-VIF and VIF regression were 1.8‰in datasets with outliers and 1.7‰in uncontaminated datasets. In addition, in all datasets, when the number of variables was over 750, no noise variables were chosen for the final model by R-VIF regression. In all scenarios, the R-VIF regression method omitted noise covariates that did not correlate with the target variables in the final model. The time taken for each method to reach the final model was longer in datasets with outliers than in uncontaminated datasets. This became more evident as the number of variables increased. In addition, in dataset with 10% outliers, the time each method took to reach the final model was slightly higher than the time taken in dataset with 5% outliers. This became more evident when the number of variables was 500 or more.

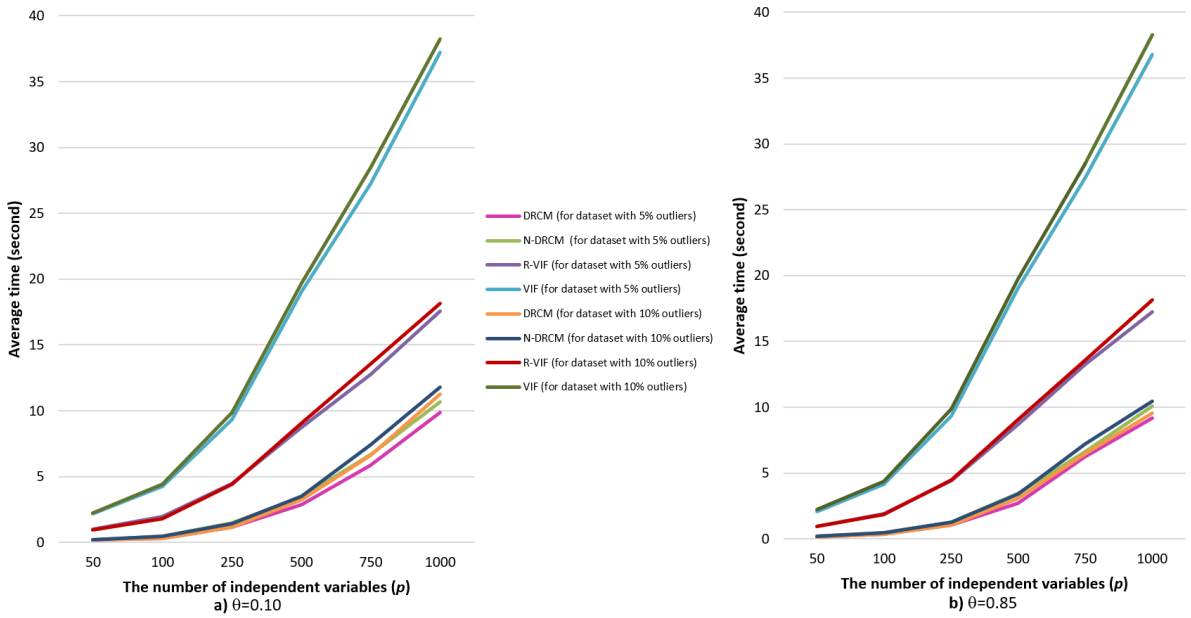

Fig. 1. The plots of average times taken to reach the final models for fast regression methods in the datasets with outliers for **a)** $\theta = 0.10$ and **b)** $\theta = 0.85$.

In all datasets, when the theta value was 0.85 and the number of variables was over 750, 19.8% of noise covariates that correlated with target variables was involved in the final models obtained by DRCM and N-DRCM. A further 9% were included in the final model when using VIF regression method. Also the numbers of total noise covariates selected to final models by both DRCM and N-DRCM methods increased slightly with increasing of multicollinearity level when the number of variables was over 100. It was determined that the numbers of total noise covariates selected to final models by the R-VIF and VIF regression methods decreased when the numbers of variables increased in both the datasets with outliers. The numbers of total noise covariates selected to finals model by both R-VIF and VIF regression methods had not changed considerably with increasing of multicollinearity level. Additionally, the numbers of total noise covariates selected to final models by the R-VIF and VIF regression methods were absent in uncontaminated dataset.

3.2 Real data

This large dataset with sample size (n=1994) and number of predictors ($p = 122$) was firstly examined in terms of multicollinearity and outliers. The VIF values of 88% of the variables were greater than 10, and their collinearity tolerance values were very close to zero. Condition index values of all dimensions except the twenty two dimensions were above 15. Moreover, the most of the variables were skew

Table 1. The performances of fast regression methods in uncontaminated dataset.

n=5000		**No Contamination**							
		R^2**=0.20,** (θ**=0.10**)				R^2**=0.80,** (θ**=0.85**)			
p	**Results**	***DRCM***	***N-DRCM***	***R-VIF***	***VIF***	***DRCM***	***N-DRCM***	***R-VIF***	***VIF***
50	Avg.Time	0.118	0.140	0.569	1.304	0.106	0.130	0.570	1.305
	A ($p_A = 5$) (%)	100	100	100	100	100	100	100	100
	B ($p_B = 10$) (%)	0	0	0	0	0	0	0	0
	C ($p_C = 35$) (%)	0.06	0.06	0	0	0.06	0.06	0	0
	D ($p_D = 45$) (%)	0.04	0.04	0	0	0.04	0.04	0	0
	RMSE	0.921	0.921	0.923	0.924	0.923	0.923	0.923	0.924
100	Avg.Time	0.182	0.250	1.092	2.480	0.159	0.224	1.119	2.487
	A ($p_A = 5$) (%)	100	100	100	100	100	100	100	100
	B ($p_B = 10$) (%)	0	0	0	0	0	0	0	0
	C ($p_C = 85$) (%)	0.86	0.85	0	0	1.29	1.16	0	0
	D ($p_D = 95$) (%)	0.77	0.76	0	0	1.15	1.04	0	0
	RMSE	0.921	0.921	0.919	0.920	0.920	0.920	0.910	0.920
250	Avg.Time	0.592	0.705	2.702	5.817	0.604	0.704	2.693	5.823
	A ($p_A = 5$) (%)	100	100	100	100	100	100	100	100
	B ($p_B = 10$) (%)	0	9.9	0	0	0	0	0	0
	C ($p_C = 235$) (%)	1.28	1.26	0	0	1.59	1.49	0	0
	D ($p_D = 245$) (%)	1.22	1.61	0	0	1.53	1.43	0	0
	RMSE	0.907	0.906	0.904	0.904	0.907	0.906	0.904	0.904
500	Avg.Time	1.903	2.145	5.335	11.416	1.888	1.999	5.425	11.460
	A ($p_A = 5$) (%)	100	100	100	100	100	100	100	100
	B ($p_B = 10$) (%)	0.1	0.1	0	0.1	0.1	10	0	0.1
	C ($p_C = 485$) (%)	1.86	1.65	0	0	2.47	2.27	0	0
	D ($p_D = 495$) (%)	1.82	1.62	0	0.002	2.42	2.43	0	0.002
	RMSE	0.908	0.908	0.907	0.908	0.905	0.905	0.904	0.904
750	Avg.Time	4.190	4.482	8.000	17.004	4.386	4.811	8.842	17.944
	A ($p_A = 5$) (%)	100	100	100	100	100	100	100	100
	B ($p_B = 10$) (%)	0	0	0	0	0	0	0	0
	C ($p_C = 735$) (%)	1.77	1.64	0	0	2.17	2.14	0	0
	D ($p_D = 745$) (%)	1.75	1.62	0	0	2.14	2.11	0	0
	RMSE	0.905	0.904	0.903	0.904	0.905	0.905	0.904	0.904
1000	Avg.Time	6.365	6.796	10.757	22.835	6.032	6.740	11.697	23.916
	A ($p_A = 5$) (%)	100	100	100	100	100	100	100	100
	B ($p_B = 10$) (%)	9.9	9.9	0	9.9	19.8	19.8	0	9.9
	C ($p_C = 985$) (%)	2.47	1.94	0	0	2.14	2.03	0	0
	D ($p_D = 995$) (%)	2.54	2.02	0	0.10	2.32	2.21	0	0.10
	RMSE	0.893	0.892	0.891	0.893	0.892	0.892	0.891	0.893

p: The number of predictors, Avg: Average, A: Average number of target covariates, B: Average number of noise covariates that correlated with target covariates, C: Average number of noise covariates that did not correlate with target covariates, D: Average number of total noise covariates, RMSEA: Root mean square error, VIF: Variance inflation factor, R-VIF: Robust VIF, DRCM: Dimensional reduction of correlation matrix, N-DRCM: Nonparametric DRCM

Table 2. The performances of fast regression methods in dataset with 5% outliers.

n=5000		**5% outliers**							
		R^2=0.20, (θ=0.10)				**R^2=0.80, (θ=0.85)**			
p	**Results**	***DRCM***	***N-DRCM***	***R-VIF***	***VIF***	***DRCM***	***N-DRCM***	***R-VIF***	***VIF***
50	Avg.Time	0.186	0.213	1.003	2.186	0.153	0.210	0.965	2.094
	A ($p_A = 5$) (%)	100	100	100	100	100	100	100	100
	B ($p_B = 10$) (%)	0.2	0.2	0.1	0.1	0	0	0	0
	C ($p_C = 35$) (%)	0	0	0	0	0	0	0	0
	D ($p_D = 45$) (%)	0.04	0.04	0.02	0.02	0	0	0	0
	RMSE	0.968	0.968	0.970	0.970	0.969	0.969	0.970	0.970
100	Avg.Time	0.315	0.486	1.983	4.258	0.330	0.462	1.882	4.149
	A ($p_A = 5$) (%)	100	100	100	100	100	100	100	100
	B ($p_B = 10$) (%)	0	0	0.1	0.1	0	0	0.1	0.1
	C ($p_C = 85$) (%)	1.41	1.43	0	0	1.62	1.55	0	0
	D ($p_D = 95$) (%)	1.26	1.28	0.01	0.01	1.45	1.39	0.01	0.01
	RMSE	0.967	0.967	0.966	0.966	0.967	0.967	0.966	0.966
250	Avg.Time	1.157	1.504	4.476	9.312	1.043	1.257	4.430	9.319
	A ($p_A = 5$) (%)	100	100	100	100	100	100	100	100
	B ($p_B = 10$) (%)	0	0	0.1	0.1	0	0	0.1	0.1
	C ($p_C = 235$) (%)	1.74	1.74	0	0	2.08	2.04	0	0
	D ($p_D = 245$) (%)	1.67	1.67	0.004	0.004	2.00	1.96	0.004	0.004
	RMSE	0.953	0.954	0.953	0.953	0.951	0.950	0.952	0.953
500	Avg.Time	2.871	3.457	8.734	19.079	2.708	3.442	8.713	19.029
	A ($p_A = 5$) (%)	100	100	100	100	100	100	100	100
	B ($p_B = 10$) (%)	0.1	0.2	0.1	0.1	10	10.2	0.1	0.1
	C ($p_C = 485$) (%)	2.08	2.06	0	0	2.47	2.27	0	0
	D ($p_D = 495$)(%)	2.04	2.02	0.002	0.002	2.62	2.43	0.002	0.002
	RMSE	0.952	0.951	0.949	0.950	0.952	0.951	0.949	0.949
750	Avg.Time	5.835	6.683	12.745	27.248	6.214	6.585	13.242	27.412
	A ($p_A = 5$) (%)	100	100	100	100	100	100	100	100
	B ($p_B = 10$)(%)	0.1	0	0.2	0.1	0	0.1	0.2	0.1
	C ($p_C = 735$) (%)	1.9	1.9	0	0	2.44	2.38	0	0
	D ($p_D = 745$) (%)	1.88	1.87	0.003	0.001	2.41	2.35	0.003	0.001
	RMSE	0.950	0.950	0.948	0.949	0.950	0.950	0.948	0.949
1000	Avg.Time	9.870	10.681	17.585	37.220	9.177	10.089	17.214	36.788
	A ($p_A = 5$) (%)	100	100	100	100	100	100	100	100
	B ($p_B = 10$) (%)	19.8	19.8	0	9.9	19.8	19.8	0	9.9
	C ($p_C = 985$) (%)	2.53	1.82	0	0	2.13	2.13	0	0
	D ($p_D = 995$) (%)	2.71	2.00	0	0.10	2.31	2.31	0	0.10
	RMSE	0.937	0.936	0.934	0.938	0.937	0.936	0.934	0.938

p: The number of predictors, Avg: Average, A: Average number of target covariates, B: Average number of noise covariates that correlated with target covariates, C: Average number of noise covariates that did not correlate with target covariates, D: Average number of total noise covariates, RMSEA: Root mean square error, VIF: Variance inflation factor, R-VIF: Robust VIF, DRCM: Dimensional reduction of correlation matrix, N-DRCM: Nonparametric DRCM

Table 3. performances of fast regression methods in dataset with 10% outliers.

n=5000		**10% outliers**							
		R^2=0.20, (θ=0.10)				**R^2=0.80, (θ=0.85)**			
p	**Results**	***DRCM***	***N-DRCM***	***R-VIF***	***VIF***	***DRCM***	***N-DRCM***	***R-VIF***	***VIF***
50	Avg.Time	0.202	0.219	0.945	2.232	0.156	0.200	0.939	2.213
	A ($p_A = 5$) (%)	100	100	100	100	100	100	100	100
	B ($p_B = 10$) (%)	0.2	0.2	0.1	0.1	0	0	0	0
	C ($p_C = 35$) (%)	0	0	0	0	0	0	0	0
	D ($p_D = 45$) (%)	0.04	0.04	0.02	0.02	0	0	0	0
	RMSE	1.019	1.020	1.019	1.020	1.019	1.019	1.019	1.020
100	Avg.Time	0.327	0.473	1.824	4.395	0.341	0.447	1.860	4.383
	A ($p_A = 5$) (%)	100	100	100	100	100	100	100	100
	B ($p_B = 10$) (%)	0	0	0.1	0.1	0	0	0.1	0.1
	C ($p_C = 85$) (%)	2.02	2.00	0	0	2.99	2.94	0	0
	D ($p_D = 95$) (%)	1.81	1.79	0.01	0.01	2.67	2.63	0.01	0.01
	RMSE	1.016	1.016	1.015	1.016	1.016	1.016	1.015	1.016
250	Avg.Time	1.152	1.438	4.408	9.881	1.030	1.251	4.455	9.853
	A ($p_A = 5$) (%)	100	100	100	100	100	100	100	100
	B ($p_B = 10$) (%)	0	0	0.1	0.1	0	0	0.1	0.1
	C ($p_C = 235$) (%)	3.10	3.07	0	0	3.25	3.25	0	0
	D ($p_D = 245$) (%)	2.98	2.95	0.004	0.004	3.12	3.11	0.004	0.004
	RMSE	1.003	1.005	1.001	1.002	1.002	1.002	1.001	1.001
500	Avg.Time	3.217	3.525	9.058	19.706	3.063	3.380	9.058	19.723
	A ($p_A = 5$) (%)	100	100	100	100	100	100	100	100
	B ($p_B = 10$) (%)	0.2	0.2	0.1	0.1	10	10	0.1	0.1
	C ($p_C = 485$) (%)	3.47	3.20	0	0	3.84	3.26	0	0
	D ($p_D = 495$)(%)	3.40	3.14	0.002	0.002	3.96	3.40	0.002	0.002
	RMSE	0.997	0.996	0.728	0.997	0.998	0.997	0.728	0.997
750	Avg.Time	6.625	7.392	13.568	28.448	6.463	7.189	13.571	28.481
	A ($p_A = 5$) (%)	100	100	100	100	100	100	100	100
	B ($p_B = 10$) (%)	0.2	0.1	0.2	0.2	10	9.9	0.2	0.2
	C ($p_C = 735$)(%)	3.91	3.88	0	0	4.17	4.16	0	0
	D ($p_D = 745$) (%)	3.86	3.83	0.003	0.003	4.25	4.03	0.003	0.003
	RMSE	0.996	0.996	0.725	0.998	0.996	0.996	0.725	0.998
1000	Avg.Time	11.236	11.799	18.142	38.231	9.528	10.459	18.138	38.265
	A ($p_A = 5$) (%)	100	100	100	100	100	100	100	100
	B ($p_B = 10$) (%)	9.9	19.8	0	9.9	19.8	19.8	0	9.9
	C ($p_C = 985$) (%)	4.54	4.42	0	0	5.20	5.05	0	0
	D ($p_D = 995$) (%)	4.59	4.57	0	0.1	5.35	5.20	0	0.1
	RMSE	0.984	0.986	0.718	0.986	0.986	0.986	0.718	0.985

p: The number of predictors, Avg: Average, A: Average number of target covariates, B: Average number of noise covariates that correlated with target covariates, C: Average number of noise covariates that did not correlate with target covariates, D: Average number of total noise covariates, RMSEA: Root mean square error, VIF: Variance inflation factor, R-VIF: Robust VIF, DRCM: Dimensional reduction of correlation matrix, N-DRCM: Nonparametric DRCM

distributed and contained outliers. The estimation values of the final models (without constant) selected using each method for "crime data" are shown Table 4.

The racepctblack (percentage of population that is African American), PctIlleg (percentage of kids born to never married), PctPersDenseHous (percent of persons in dense housing (more than 1 person per room)), NumStreet (number of homeless people counted in the street) variables were selected for the final models by all four methods. The number of predictors selected for the final models ranged from 14 to 16. Approximately the numbers of predictors selected by all methods to their final models were similar. In descending order, these methods were N-DRCM, R-VIF regression, VIF regression, and DRCM. The highest R^2 value was obtained by the R-VIF regression method, followed by the N-DRCM method. The R^2 values obtained by VIF regression and DRCM methods were similar and considerably lower than the values obtained by the other two methods. While the RMSE value obtained with the VIF regression method was the lowest, this method was followed by the R-VIF regression, N-DRCM, and DRCM methods, respectively. Overall, R-VIF regression performed better because its final model had the highest R^2 among those obtained with other methods, and the lowest RMSE value among those obtained with others excepting VIF regression.

4. Discussion and conclusion

When large datasets contain multicollinearity and outliers, the use of fast regression algorithms has become mandatory to address the lack of traditional methods and the loss of information that occurs when using traditional methods. In the literature review, it was noted that a limited number of researches about fast regression methods are being conducted. Lin *etal.* (2021) compared stepwise regression, LASSO, FoBa, GPS methods to test the performance of the VIF regression method they developed. They found the performance of VIF regression to be better than other algorithms in terms of computation speed, out-of-sample, out-of-sample error, mFDR control, etc. Dupuis & Victoria-Feser (2013) and Seo (2018) suggested using the R-VIF regression in place of classical VIF regression to obtain faster estimations when working with large datasets that contain outliers. In addition, Midi & Uraibi (2014) compared DRCM, VIF regression and Adaptive Lasso methods, and they obtained that the performance of DRCM method was more efficient than the others. Shahriari (2014) examined the performances of LARS, R-LARS, R-VIF and JKR-LARS methods in datasets with outliers and/or leverage points. Shahriari (2014) found that JKR-LARS performed similarly to R-LARS and R-VIF in datasets with outliers while outperforming R-LARS in datasets with high leverage points. However, according to her study, R-VIF failed to robustly sequence predictor variables in datasets with high leverage points. Uraibi (2020) investigated that VIFRegSd2, VIFRegSd3, and ISIS method in ultrahigh dimensional feature space when presence of collinearity structure. Uraibi (2020) found that VIFRegSd2 and VIFRegSd3 methods outperform ISIS, additionally VIFRegSd2 method can be used in practice for ultrahigh feature space and small sample size.

In this study, the performances of DRCM, N-DRCM, VIF regression, and R-VIF regression in relation to large datasets with varying levels of multicollinearity and outliers were examined in different scenarios. This study proposed that the N-DRCM method could be used as a fast regression estimator. As the number of variables and the level of outliers increased, the time taken to reach the final model by each method increased. When the number of variables was 500 or above and the level of outliers in the dataset increased, the times taken to reach the final models by DRCM and N-DRCM methods increased. When the level of multicollinearity and the number of variables ($p > 500$) increased, the times to reach the final models using DRCM in datasets with outliers were slightly shorter than the those of N-DRCM. However, in all scenarios, DRCM and N-DRCM were found to be the fastest methods to reach the final models. When the number of variables was over 750 in uncontaminated datasets, the times taken to reach the final models using DRCM and N-DRCM methods decreased with increasing of multicollinearity level. Moreover the numbers of total noise covariates selected to final models by both DRCM and N-DRCM methods increased slightly with increasing of multicollinearity level when the number of variables was over 100. It was observed that the numbers of total noise covariates and the numbers of total noise covariates that did not correlate with the target variables selected for the final models by the DRCM and N-DRCM methods were higher than those achieved via R-VIF and VIF

Table 4. The estimation values of final models selected using each methods (n=1994, p=122).

Methods	Variables	Beta	SE	t-statistic	p-value	R^2	RMSE
VIF R.	racepctblack	0.177	0.024	7.329	<0.001	0.640	0.140
	pctUrban	0.054	0.008	6.871	<0.001		
	pctWInvInc	-0.263	0.024	-10.763	<0.001		
	MalePctNevMarr	-0.104	0.023	-4.523	<0.001		
	PctWorkMom	-0.117	0.020	-5.962	<0.001		
	PctIlleg	0.345	0.034	10.289	<0.001		
	PersPerOccupHous	-0.356	0.036	-9.860	<0.001		
	PctPersDenseHous	0.281	0.036	7.756	<0.001		
	PctHousLess3BR	-0.139	0.034	-4.073	<0.001		
	MedNumBR	-0.053	0.018	-3.000	0.003		
	PctVacantBoarded	0.066	0.018	3.596	<0.001		
	MedOwnCostPctIncNoMtg	-0.061	0.018	-3.318	0.001		
	NumStreet	0.242	0.036	6.767	<0.001		
	LemasSwornFT	-0.275	0.074	-3.715	<0.001		
	PolicOperBudg	0.204	0.076	2.685	0.007		
R-VIF R.	racepctblack	0.220	0.021	10.681	<0.001	0.904	0.439
	agePct12t29	-0.185	0.044	-4.226	<0.001		
	agePct16t24	0.147	0.041	3.575	<0.001		
	numbUrban	-0.129	0.027	-4.859	<0.001		
	pctUrban	0.064	0.013	5.112	<0.001		
	pctWWage	-0.065	0.030	-2.175	0.030		
	pctWRetire	-0.033	0.015	-2.295	0.022		
	OtherPerCap	1.119	0.010	108.563	<0.001		
	PctEmploy	0.082	0.029	2.845	0.005		
	MalePctDivorce	0.070	0.020	3.485	0.001		
	PctKids2Par	-0.211	0.042	-4.979	<0.001		
	PctWorkMom	-0.053	0.013	-3.983	<0.001		
	PctIlleg	0.214	0.030	7.230	<0.001		
	PctPersDenseHous	0.220	0.015	14.621	<0.001		
	HousVacant	0.186	0.025	7.541	<0.001		
	NumStreet	0.111	0.014	8.090	<0.001		

R: Regression, SE: Standard error, RMSE: Residual mean square estimation, VIF: Variance inflation factor, R-VIF: Robust VIF, racepctblack: percentage of population that is African American, pctUrban: percentage of people living in areas classified as urban, pctWInvInc: percentage of households with investment, MalePctNevMarr: percentage of males who have never married, PctWorkMom: percentage of moms of kids under 18 in labor force, PctIlleg: percentage of kids born to never married, PersPerOccupHous: mean persons per household, PctPersDenseHous: percent of persons in dense housing (more than 1 person per room), PctHousLess3BR: percent of housing units with less than 3 bedrooms, MedNumBR: median number of bedrooms, PctVacantBoarded: percent of vacant housing that is boarded up, MedOwnCostPctIncNoMtg: median owners cost as a percentage of household income, NumStreet: number of homeless people counted in the Street, LemasSwornFT: number of sworn full time police officers, PolicOperBudg: police operating budget, agePct12t29: percentage of population that is 12-29 in age, agePct16t24: percentage of population that is 16-24 in age, numbUrban: number of people living in areas classified as urban, pctWWage: percentage of households with wage or salary income in 1989, pctWRetire: percentage of households with retirement income in 1989, OtherPerCap: per capita income for people with 'other' heritage, PctEmploy: percentage of people 16 and over who are employed, MalePctDivorce: percentage of males who are divorced, PctKids2Par: percentage of kids in family housing with two parents, HousVacant: number of vacant households

Table 4.(continue). The estimation values of final models selected using each methods (n=1994, p=122).

Methods	Variables	Beta	SE	t-statistic	p-value	R^2	RMSE
DRCM	racepctblack	0.873	0.103	8.490	<0.001	0.649	0.595
	agePct12t29	-1.768	0.115	-15.442	<0.001		
	pctUrban	0.141	0.036	3.904	<0.001		
	pctWPubAsst	0.324	0.113	2.875	0.004		
	PctLess9thGrade	-0.991	0.211	-4.689	<0.001		
	PctNotHSGrad	0.665	0.239	2.780	0.006		
	MalePctNevMarr	0.762	0.151	5.032	<0.001		
	PctIlleg	1.307	0.150	8.733	<0.001		
	PctPersDenseHous	0.956	0.095	10.063	<0.001		
	PctHousLess3BR	0.444	0.097	4.560	<0.001		
	HousVacant	0.681	0.119	5.711	<0.001		
	PctHousNoPhone	0.649	0.102	6.390	<0.001		
	MedOwnCostPctIncNoMtg	-0.533	0.074	-7.216	<0.001		
	NumStreet	0.559	0.170	3.298	0.001		
N-DRCM	racepctblack	0.232	0.022	10.574	<0.001	0.748	0.451
	agePct12t29	-0.135	0.025	-5.347	<0.001		
	Pct65up	-0.069	0.022	-3.081	0.002		
	pctWPubAsst	0.110	0.020	5.611	<0.001		
	PctLess9thGrade	-0.206	0.038	-5.441	<0.001		
	PctNotHSGrad	0.275	0.048	5.741	<0.001		
	PctOccupManu	-0.054	0.020	-2.707	0.007		
	MalePctNevMarr	0.052	0.024	2.179	0.030		
	PersPerFam	-0.136	0.020	-6.843	<0.001		
	PctIlleg	0.296	0.029	10.239	<0.001		
	PctNotSpeakEnglWell	-0.101	0.029	-3.445	0.001		
	PctPersDenseHous	0.376	0.030	12.487	<0.001		
	HousVacant	0.153	0.019	8.220	<0.001		
	NumStreet	0.076	0.014	5.535	<0.001		
	LemasSwornFT	-0.039	0.012	-3.268	0.001		
	LandArea	-0.034	0.016	-2.164	0.031		

R: Regression, SE: Standard error, RMSE: Residual mean square estimation, DRCM: Dimensional reduction of correlation matrix, N-DRCM: Nonparametric DRCM, racepctblack: percentage of population that is African American, agePct12t29: percentage of population that is 12-29 in age, pctUrban: percentage of people living in areas classified as urban, pctWPubAsst: percentage of households with public assistance income in 1989, PctLess9thGrade: percentage of people 25 and over with less than a 9th grade education, PctNotHSGrad: percentage of people 25 and over that are not high school graduates, PctHousNoPhone: percent of occupied housing units without phone, MedOwnCostPctIncNoMtg: median owners cost as a percentage of household income, MalePctNevMarr: percentage of males who have never married, PctIlleg: percentage of kids born to never married, PctPersDenseHous: percent of persons in dense housing, PctHousLess3BR: percent of housing units with less than 3 bedrooms, HousVacant: number of vacant households, PctHousNoPhone: percent of occupied housing units without phone, MedOwnCostPctIncNoMtg: median owners cost as a percentage of household income - for owners without a mortgage, NumStreet: number of homeless people counted in the street, agePct65up: percentage of population that is 65 and over in age, PctOccupManu: percentage of people 16 and over who are employed in manufacturing, PersPerFam: mean number of people per family, PctNotSpeakEnglWell: percent of people who do not speak English well, LandArea: land area in square miles, LemasSwornFT: number of sworn full time police officers

regression methods. As a result of the real dataset, the final model selected using R-VIF regression had the highest R^2. This model also had the lowest RMSE value among those obtained with other methods excluding VIF regression. Consequently, it was decided that the R-VIF regression method performed best in contaminated and uncontaminated datasets.

Due to recent technological advances, the authors of this study suggest to use fast regression methods instead of conventional methods. The R-VIF regression method is particularly recommended as a fast regression estimator in the datasets containing multicollinearity and outliers.

References

Cai, L., Huang, T., Su, J., Zhang, X., Chen, W., Zhang, F., He, L., & **Chou, K.C. (2018)**. Implications of newly identified brain eQTL genes and their interactors in schizophrenia. Molecular Therapy-Nucleic Acids, 12, 433-442.

Candes, E., & **Tao, T. (2007)**. The Dantzig selector: statistical estimation when p is much larger than n. The Annals of Statistics, 35(6), 2313-2351.

Dupuis, D.J., & **Victoria-Feser, M.P. (2013)**. Robust VIF regression with application to variable selection in large data sets. The Annals of Statistics, 7(1), 319-341.

Efron, B., Hastie, T., Johnstone, I., & **Tibshirani, R. (2004)**. Least angle regression. The Annals of Statistics, 32(2), 407-499.

Fan, J., & **Lv, J. (2008)**. Sure independence screening for ultra-high-dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(5), 849-911.

Friedman, J.H. (2008). Fast sparse regression and classification. Technical report. Stanford University, California. Available from: https://jerryfriedman.su.domains/ftp/GPSpaper.pdf.

Foster, D.P., & **Stine, R.A. (2008)**. α-investing: a procedure for sequential control of expected false discoveries. Journal of the Royal Statistical Society Series B, 70(2), 429-444.

Khan, J.A., Van Aelst, S., & **Zamar, R.H. (2007)**. Robust linear model selection based on least angle regression. Journal of the American Statistical Association, 102(480), 1289-1299.

Lin, D., Foster, D.P., & **Ungar, L.H. (2011)**. VIF regression: a fast regression algorithm for large data. Journal of the American Statistical Association, 106(493), 232-247.

Liu, C., Zhang, Y.-H., Deng, Q., Li, Y., Huang, T., Zhou, S., & **Cai, Y.-D. (2017)**. Cancer-related triplets of mRNA-lncRNA-miRNA revealed by integrative network in uterine corpus endometrial carcinoma. BioMed Research International, 2017, Article ID 3859582, 7 pages. http://dx.doi.org/10.1155/2017/3859582.

Midi, H., & **Uraibi, H.S. (2014)**. The dimensional reduction of correlation matrix for linear regression model selection. In *Mathematical and Computational Methods in Science and Engineering. Proceedings of the 16th International Conference on Mathematical and Computational Methods in Science and Engineering* (pp. 166-169). MACMESE '**14**. Kuala Lumpur, Malaysia: WSEAS Press.

Rahman, M.S., & **Khan, J.A. (2010)**. Robust stepwise algorithms for linear regression: a comparative study. Dhaka University Journal of Science, 58(2), 291-295.

Redmond, M. (2009). Communities and crime data set. UCI Machine Learning Repository. Available from: https://archive.ics.uci.edu/ml/datasets/communities+and+crime. Accessed: December, 2021.

Seo, H.S. (2018). Fast robust variable selection using VIF regression in large datasets. The Korean Journal of Applied Statistics, 31(4), 463-473.

Shahriari, S. (2014). Variable selection in linear regression models with large number of predictors. Ph.D. thesis, Universidade do Minho Escola de Ciências, Guimarães, Portugal.

Shahriari S., Faria S., Gonçalves A.M., & **Van Aelst S. (2014)**. Outlier detection and robust variable selection for least angle regression.In *Computational Science and Its Applications–ICCSA 2014. Lecture Notes in Computer Science*, vol 8581. Springer, **Cham**. pp.512-522.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B, 58(1), 267-288.

Uraibi, H.S. (2020). VIF-regression screening ultrahigh dimensional feature space. Journal of Modern Applied Statistical Methods, 19(1), eP2916.

Zhang, T. (2009). Adaptive forward–backward greedy algorithm for sparse learning with linear models. Advances in Neural Information Processing Systems, 21, 1921-1928.

Zhou, J., Foster, D.P., Stine, R.A., & **Ungar, L.H. (2006)**. Streamwise feature selection. Journal of Machine Learning Research, 7, 1861-1885.

Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101(476), 1418-1429.

Zou, H., & **Hastie, T. (2005)**. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B, 67(2), 301-320.

Submitted: 10/09/2021
Revised: 15/03/2022
Accepted: 22/03/2022
DOI: 10.48129/kjs.16159