https://doi.org/10.56093/ijas.v90i6.1047789

Indian Journal of Agricultural Sciences **90** (6): 1155–9, June 2020/Article

Effect of influential observation in genomic prediction using LASSO diagnostic

NEERAJ BUDHLAKOTI*, ANIL RAI and D C MISHRA

ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110 012, India

Received: 19 July 2019; Accepted: 23 September 2019

ABSTRACT

Detection of influential observation is one of the crucial steps of pre-processing to identify suspicious elements of data that may be due to error or some other unknown source. Several statistical measures are developed for detection of influential observation but still challenges are there to detect a true influential observation for high dimension data like gene expression, genotyping data. In this article we have demonstrated the effect of influential observation on genomic prediction accuracy by using recently proposed LASSO diagnostic, i.e. Df-Model, Df-Regpath, Df-Cvpath, Df-Lambda and Influence-LASSO. The effect of influential observation on genomic prediction accuracy was explored by observing the change in estimated and true accuracies for dataset with and without influential observation scenario. For this purpose we have used wheat and maize datasets which are available in public domain. It has been observed that influential observation had significant effects on the genomic prediction accuracy. In this study it has been shown that by implementing efficient diagnostic measure for influential observation detection, accuracy of genomic prediction can be improved.

Key words: GEBVs, Genomic prediction, Influential observation, LASSO, MSE, Prediction accuracy

Genomic Selection (GS) is one of the promising tools for improving genetic gain in animal and plants in today's scenario. GS is initially presented and applied by Meuwissen *et al.* (2001). Here effect of each marker is estimated and sum total of all marker effect is used for calculation of genotypic merit of individual, i.e. Genome Estimated Breeding Values (GEBVs). It is highly expected that performance of genomic prediction may be adversely affected by influential observation. Various diagnostic has been developed for detecting influential observation but they are generally restricted to usual regression. In genomic data where the number of markers (p) is much higher than the number of individuals or observations (n) leads to the problem of p>>n consequently in such cases usual regression technique cannot be applied. Although using a subset of markers can be alternative to this problem by compromising some level of accuracy. However in such situation penalized regression technique like, least absolute shrinkage and selection operator (LASSO) and ridge regression (RR) can be more effective. In such cases (p>>n), each observation may have huge and tremendous impact on model selection and further inference. In fact it is more desirable to address the issue of influential observation by using LASSO than the general linear regression set up, as it takes care of both parameters estimates and model selection. Detection of influential observation for linear regression set up, has been extensively explored by statistician in past (Cook 1977, 1979, Belsley *et al.* 1980, Geert 2000, Pena 2005, Loy *et al*. 2017). Zhao *et al.* (2013) has proposed for the solution to p>>n problem using influential observation by marginal correlation. Wang *et al.* (2016) has proposed a method for outlier detection based on distance correlation. Very recently, Rajaratnam *et al.* (2019) have been proposed the method for detection of influential observation in LASSO setup and proved that their diagnostic measures are performing better than the other existing methods especially in the case of high dimensional data. In this article, we have demonstrated the effect of influential observations on genomic prediction accuracy by using these proposed LASSO diagnostics.

*Corresponding author e-mail: neeraj35669@gmail.com

MATERIALS AND METHODS

LASSO was first time introduced by Tibshiranis (1996). LASSO minimizes the sum of squares of residuals subject to a constraint on sum of absolute values of regression coefficients. It can be written in the form of simple statistical model as:

$$\left(\hat{u},\hat{\beta}\right) = \arg\min(Y - X\beta)'(Y - X\beta), \text{ subject to } \sum\nolimits_j |\beta_j| \leq \lambda$$

where, $\beta = \Sigma_j |\beta_j|$ is $l1$ norm penalty on β which results in sparsity of solution and λ is a regularization parameter. Here Y is a vector of observation of phenotypic and X is a design matrix of marker genotype. As LASSO uses l_1 penalty

so as λ increases, the coefficient shrinks towards zero and some exactly zero. Due to this property LASSO is able to perform model selection and to generate interpretable model.

Here we have used a recently proposed LASSO diagnostic for measuring the influential observation (Rajaratnam *et al.* 2019). The measure used here are Df-Model- it measure the change in model selected; Df-Lambda: it measures the change in tuning parameter λ, Df-Regpath: it measures the changes observed in LASSO regularization path and Df-Cvpath which observe changes in LASSO cross-validation path. Details of these measures are given below;

Df-Model: By using this measure, it observes deviation in model selection using LASSO when an observation is dropped from model. This procedure is repeated every time until a single observation is dropped exactly once from the model. To declare any observation as influential, cut-off value ±2 is used for Df-Model.

Df-Lambda: This measure is obtained by observing the deviation in regularization parameter λ for full LASSO model vs. when i^{th} observation is discarded from model fitting. To measure this change is important as these very parameters tell at what extent selected LASSO model is shrinking the estimates. To declare any observation as influential cut-off value ±2 is justified for df-lambda.

Df-Regpath: It measures the deviation in the LASSO regularization path when an observation is discarded from LASSO path. If it shows a significant deviation from LASSO original path, this further suggests that discarded observation could have tremendous impact on LASSO estimates which further may affect the conclusion and interpretation for LASSO solution. For Df-Regpath, suitable cut-off for computed values can be justified at ±2.

Df-Cvpath: It observes the changes in predictive performance of LASSO when an observation is dropped from LASSO path. Quantifying this is crucial as if there is significant change in predictive performance of LASSO, suggests that it has infrequent response hence observation has huge impact on LASSO solution. It generates a cross-validation error curve $\gamma(s)$ which gives estimate of prediction error on test data after LASSO is trained on data for range of values for regularization parameter λ.Cut-off ±2 can be used for Df-Cvpath, which further identify the observation affecting LASSO cross-validation path. For more detailed discussion for described LASSO diagnostic above, please refer to Rajaratnam *et al.* (2019). In order to produce more consistent and robust method ofinfluential observation detection, i.e. that targets both model selection and model inference, Rajaratnam *et al.* (2019) proposed a two stage approach for diagnosis of influential observation named as Influence-LASSO. Procedure of same is given below;

Influence-LASSO

Step 1: use LASSO influence measure, i.e. Df-Model, Df-Regpath, Df-Cvpath, Df-Lambda on a complete dataset having dimension (n, p) to detect possible influential observation. Let say I_1 are no. of observation to be detected as influential.

Step 2: Now fit the LASSO to dataset having dimension of (n-I_1, p), let's assume there are q no. of variables to be included in the model selected by LASSO.

Step 3: Apply OLS Diagnostic on dataset of reduced dimension (n-I_1, q). Here assume that I_2 are no. of influential observation detected by OLS diagnostic.

Step 4: Fit OLS regression on dataset of dimension (n-I_1-I_2, q) where influential point detected in step 1 and 3 are discarded.

Experimental dataset

In order to check the performance of above mentioned methods, two datasets has been considered, i.e. wheat and maize. Former one comprises of CIMMYT 599 wheat lines and later one comprises 264 CIMMYT maize lines. Both the dataset are freely available in public domain. Brief details of both the dataset is given below;

Dataset 1: *Wheat*: This wheat data is generated through CIMMYT global wheat programme. Wheat lines were genotyped using 1447 Diversity Array Technology (DArT) markers generated by Triticarte Pty. Ltd. (Canberra, Australia; http://www.triticarte.com.au). Markers have two alternative values, i.e.1 (for their presence) or 0 (for absence). This data set includes 599 lines observed for trait grain yield (GY) for four mega environments. However for our convenience we have just considered GY for first mega environment. The final figure 1279 DArT markers were available after some editing hence same has been used in this study. This dataset has also used in earlier genomic prediction study (Crossa *et al.* 2010 and Cuevas *et al.* 2016).

Dataset 2: Maize: Maize dataset used in this study generated by CIMMYT's Global Maize Program (Crossa *et al.* 2010). It originally has 1148 SNP markers available for 300 maize line. Marker with maximum occurrence is coded as 0 and for minimum occurrence as 1. This study has trait Grain Yield (GY), evaluated under draught and watered conditions hence same has been used here. After some editing final figure 264 maize lines with 1135 SNPs markers were available for final study (Crossa *et al.* 2010).

For application to genomic prediction data Accuracy and Mean Squared Error (MSE) were used an evaluation measure. Accuracy can be defined as Pearson correlation coefficient between actual phenotypic value and predicted phenotypic value. If we assume $\hat{Y} = X\hat{\beta}$, where $\hat{Y}$ is estimated response and $\hat{}$ is estimated value of β, then correlation coefficient (r) can be expressed as:

$$r = \frac{S_{Y,\hat{Y}}}{S_Y S_{\hat{Y}}}$$

where $S_{Y,\hat{Y}}$ denotes the covariance between observed and predicted phenotypic value, S_Y is standard deviation of observed phenotype and $S_{\hat{Y}}$ denotes standard deviation of predicted phenotype. MSE can be calculated as average of squared difference between actual phenotypic value and predicted phenotypic value. MSE can be estimated as:

$$PE / MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

where, Y_i is observed response; $\hat{Y}_i$ is predicted phenotype value. Methodology described implemented using R (R Development Core Team 2019), package inflasso: Influence Diagnostics for Lasso Regression. R can be downloaded from http://www.r-project.org, the inflasso package can be accessed by typing library (inflasso) in R console.

RESULTS AND DISCUSSION

For understanding the effects of influential observation on the accuracy of genomic prediction, wheat and maize dataset has been used in the study. In order to see the effects, first we have fitted LASSO regression on original data under study say it as LASSO* for genomic prediction. Subsequently we have calculated Df-Model, Df-Regpath, Df-Cvpath, Df-Model and Influence-LASSO to identify possible influential observation in the data based on cutoff ±2 Identified influential observation and their corresponding genotype is dropped from the model and LASSO is refitted using the revised data. In order to see the performance of above methods, cross validation techniques is used. Here dataset is divided into two parts, i.e. training and testing sets such that training set contains of 70% data and remaining 30% pertaining to testing set. Former one is used for model construction and later one for checking the performance of developed model. Performance of methods under consideration was evaluated by estimating prediction accuracy and mean squared error (MSE). This process is repeated 100 times and prediction accuracy and prediction error is averaged and corresponding standard error (SE) calculated.

Table 1 gives the average prediction accuracies and MSE with their corresponding SE using different methods for dataset 1, whereas Table 2 reports the average prediction accuracies and MSE with their SE for dataset 2.

It can be observed (Table 1) (for wheat data) that among all the methods Influence-LASSO outperformed among other influential observation dignostic. However it can be noted that performance of Df-regpath, Df-Cvpath, Df-Lambda is almost similar to one other. Table 2 (for maize data) also conclude the similar results. Here Infuence-LASSO again performed in the same way. It is also observed that performance of Df-Model and Df-Lambda is found almost similar. It is observed that for both the dataset prediction accuracy is significantly increased (As best case in dataset 1, up to 34% and maize dataset 2, 54%) and MSE is significantly lowered (As best case in dataset 1 up to 34% and dataset 2, 28%). Same can be demonstrated using the graphical representation (Fig1, 2).

Fig 1 contain 12 box plot of prediction accuracy (six for each dataset highlighted using different color) for datasets 1, 2 respectively. These box plots show the distribution of prediction accuracy, estimated over 100 replications. Whereas, Fig 2 represents 12 box plot of prediction error (MSE) for datasets 1, 2 on the same pattern to boxplot of Fig 1. These box plots show the distribution of the MSE values over 100 runs.

It can be easily concluded from both Figures (Fig 1, 2) and Tables (Table 1, 2) that influential observation significantly affects the genomic prediction accuracy and mean squared error. These results demonstrate a clear cut advantage of identifying a possible influential observation and their treatment to further improve the genomic prediction accuracy. For both the datasets, i.e. Wheat and Maize dataset, prediction accuracy has been significantly improved and MSE get decreased. Above results shows the importance of detection of influential observation for estimating more accurate GEBVs. Accurate GEBVs will further lead to appropriate selection of individuals for breeding purpose. This article demonstrates the effect of influential observation on the accuracy of genomic prediction using real datasets.

Table 1 Mean and standard error of prediction accuracy and prediction error for different methods using dataset 1

Method	Accuracy	MSE	Accuracy (SE)	MSE (SE)	Percentage (%) gain in Accuracy	Percentage (%) reduction in MSE
LASSO*	0.44	0.82	0.06	0.08	NA	NA
Df-Model	0.47	0.83	0.06	0.09	6.8	0
Df-Regpath	0.55	0.60	0.05	0.07	25	27
Df-Cvpath	0.57	0.58	0.06	0.07	29.5	29.3
Df-Lambda	0.56	0.66	0.06	0.09	27.3	19.5
Influence LASSO	0.59	0.54	0.05	0.05	34	34.2

LASSO* represent LASSO regression fitted in original data (i.e. without any treatment to possible influential observation), next five methods in the table represent performance of LASSO in the absence of influential observation (i.e. possible influential observation and corresponding genotype marker genotype dropped from the model detected by above LASSO diagnostic).

Table 2 Mean and standard error of prediction accuracy and prediction error for different methods using dataset 2

Method	Accuracy	MSE	Accuracy (SE)	MSE (SE)	Percentage (%) gain in Accuracy	Percentage (%) reduction in MSE
LASSO*	0.26	0.96	0.09	0.14	NA	NA
Df-Model	0.36	0.96	0.11	0.16	38.5	0
Df-Regpath	0.28	1.01	0.10	0.14	7.7	0
Df-Cvpath	0.30	0.99	0.09	0.12	15.4	0
Df-Lambda	0.38	0.96	0.11	0.16	46.2	0
Influence LASSO	0.40	0.70	0.10	0.11	53.8	28.2

LASSO* represent LASSO regression fitted in original data (i.e. without any treatment to possible influential observation), next five methods in the table represent performance of LASSO in the absence of influential observation (i.e. possible influential observation and corresponding marker genotype dropped from the model detected above LASSO diagnostic).

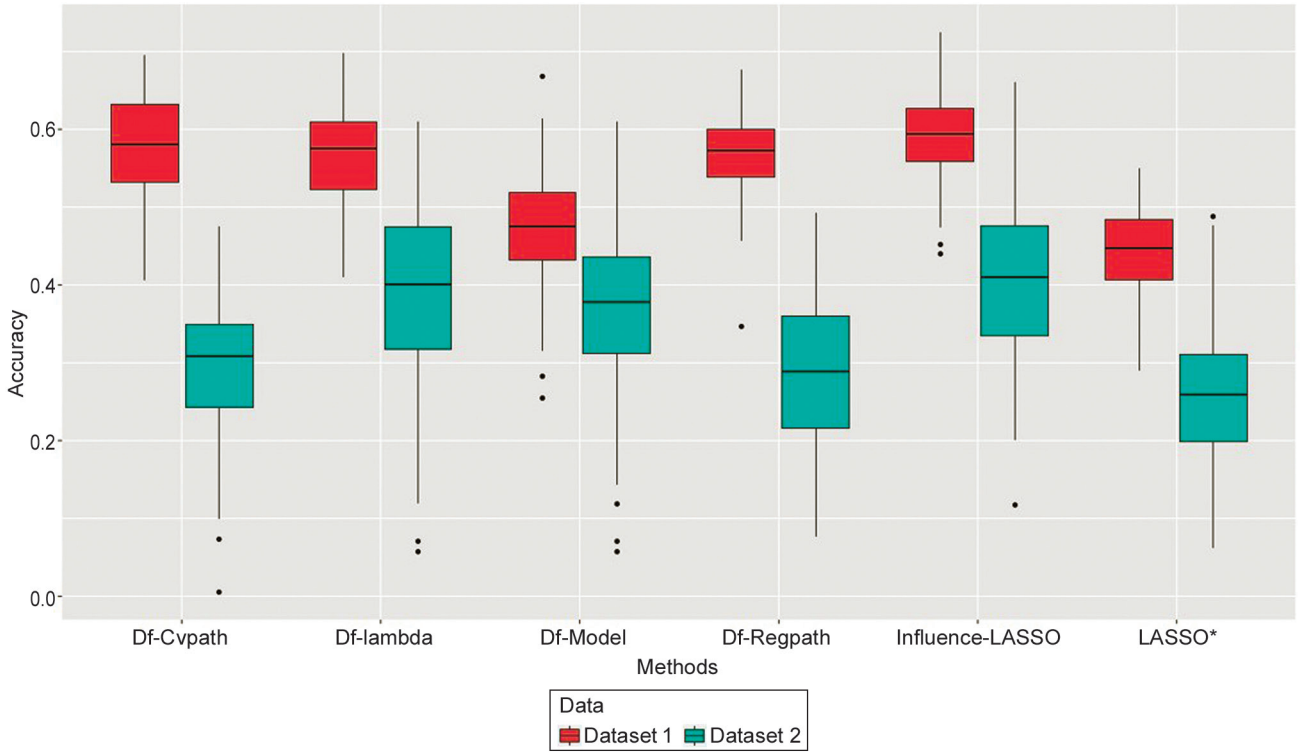

Fig 1 Box plot of genomic prediction accuracy for various methods under study using datasets 1, 2.

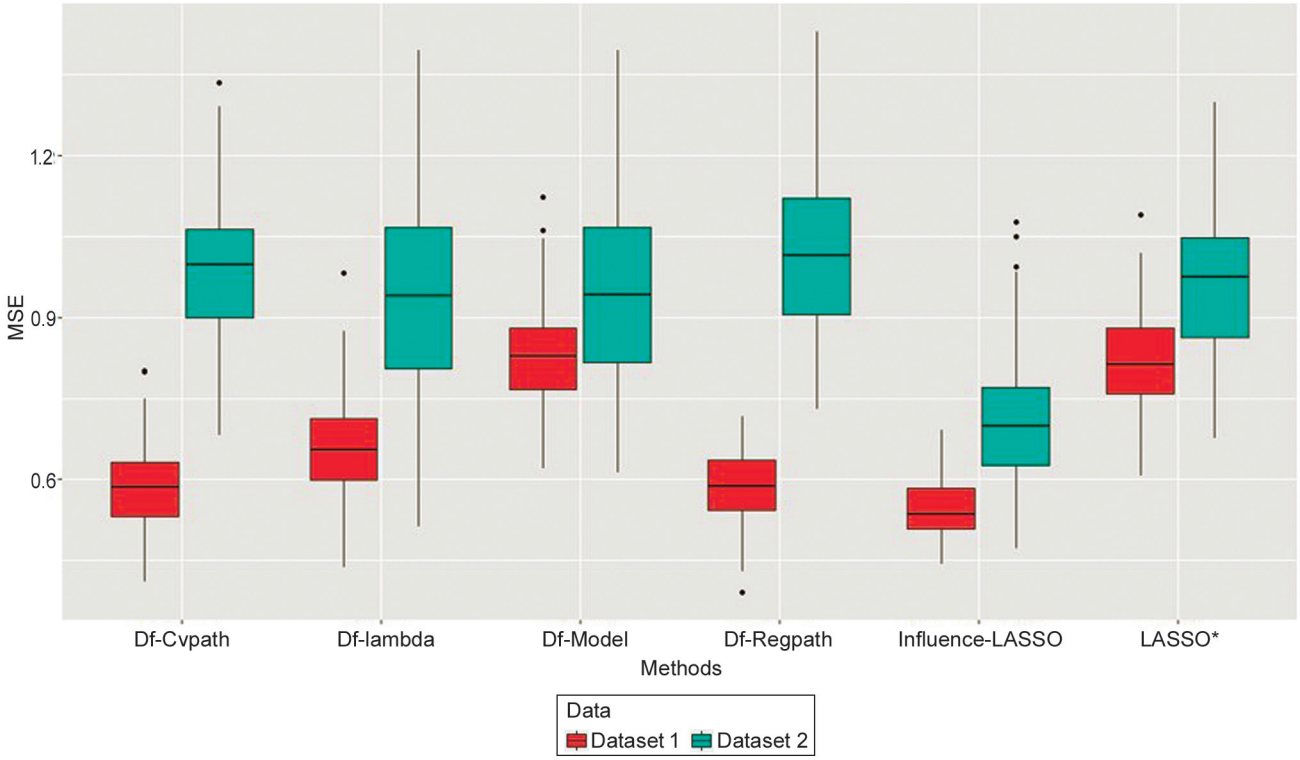

Fig 2 Box plot of MSE for various methods under study using datasets 1, 2.

Here we have discussed the results of influential observation detection methods available for high dimensional data. We have done the comparative analysis of existing methods and presented the results. It can be concluded that influential observations have significant effects on GEBVs.

REFERENCES

Belsley D A, Kuh E and Welsch R E. 1980. Regression Diagnostics:

Identifying Influential Data and Sources of Collinearity. New York: Wiley.

Cook R D. 1977. Detection of influential observation in linear regression. *Technometrics* **19**: 15–18.

Cook R D. 1979. Influential observations in linear regression. *Journal of the American Statistical Association* **74**: 169–74.

Crossa J, De Los Campos G, Pérez P, Gianola D and Burgueno J. 2010. Prediction of genetic values of quantitative traitsin plant breeding using pedigree and molecular markers. *Genetics* **186**: 713–24.

Cuevas J, Crossa J, Soberanis V, Perez-Elizalde S and Perez-Rodríguez P. 2016. Genomic prediction of genotype × environment interaction kernel regression models. *Plant Genome* **9**: 1–12.

Geert V and Geert M. 2000. Linear Mixed Models for Longitudinal Data. Springer Series in Statistics. DOI:10.1007/978-1-4419-0300-6.

Loy A, Hofmann H and Cook D. 2017. Model choice and diagnostics for linear mixed-effects models using statistics on street corners. *Journal of Computational and Graphical Statistics* **26**: 478–92.

Meuwissen T H E, Hayes B J and Goddard M E. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819–29.

Pena D. 2005. A new statistic for influence in linear regression. *Technometrics* **47**: 1–12.

R Development Core Team. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Rajaratnam B, Roberts S, Sparks D and Yu H. 2019. Influence diagnostics for high-dimensional LASSO regression. *Journal of Computational and Graphical Statistics* **28**(4): 877-90.DOI: 10.1080/10618600.2019.1598869.

Tibshirani R. 1996. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society* **58**: 267–88.

Wang T and Li Z. 2016. Outlier detection in high-dimensional regression model.*Communications in Statistics—Theory and Methods* **46:** 6947–58.

Zhao J, Leng C, Li L and Wang H. 2013. High-dimensional influence measure. *Annals of Statistics* **41**: 2639–67.