



OPEN

Statistical Approach for Improving Genomic Prediction Accuracy through Efficient Diagnostic Measure of Influential Observation

Neeraj Budhlakoti, Anil Rai & D. C. Mishra✉

It is expected the predictive performance of genomic prediction methods may be adversely affected in the presence of outliers. In agriculture science an outlier may arise due to wrong data imputation, outlying response, and in a series of trials over the time or location. Although several statistical procedures are already there in literature for identification of outlier but identification of true outlier is still a challenge especially in case of high dimensional genomic data. Here we have proposed an efficient approach for detecting outlier in high dimensional genomic data, our approach is p-value based combination methods to produce single p-value for detecting the outliers. Robustness of our approach has been tested using simulated data through the evaluation measures like precision, recall etc. It has been observed that significant improvement in the performance of genomic prediction has been obtained by detecting the outliers and handling them accordingly through our proposed approach using real data.

Genomic selection (GS) has been a popular choice for selection of appropriate candidates for breeding in the current research arena of plant and animal science. Various studies has been carried out in recent past. GS is an advance method of breeding where genome-wide dense markers information is used to predict genetic merit of an individuals in a breeding programme. In today's scenario GS is a promising tool for improving genetic gain of individuals under study. Genomic selection is firstly introduced by Meuwissen *et al.*¹. In this approach individual effect of each marker is estimated and sum of all markers effect is used for calculation of genotypic value i.e. Genome Estimated Breeding Value (GEBVs) of each individual.

GS process starts with building a statistical model from individuals having both genotypic and phenotypic data (i.e. training set), this model is further used for estimation of GEBVs for individuals having only genotypic information. Individuals are then ranked on the basis of GEBVs and subsequently superior individuals are selected. Genomic selection methods have been successfully applied for plants^{2,3} and animals⁴⁻⁷. However, success of genomic selection depends on the quality of the data suitable for implementing the various statistical models. But in practical situation genomic data quality seldom fulfill the ideal condition and often having many constraints such as presence of influential observations, missing points, noise etc.

Influential observations can potentially have devastating effects on genome estimated breeding values⁸. These influential observations can be the results of wrong data imputation, outlying response, and in a series of trials over the time or location. Detection of influential observation has been an extensive research area based on linear regression approach⁹⁻¹². Some of most widely used measures for this are Cook's D, DFBETA, DFFITS, Atkinson's Ci, COVRATIOi. Among them Cook's D is one of the most commonly used measure for outlier detection through linear regression technique¹⁰. Various statistical model with t-distributed error has been proposed (Bayesian with t-linear model¹³, Gaussian process with t-likelihood¹⁴, Regression with t-error¹⁵) as robust method against treating the outlier. Lange *et al.*¹⁵ have applied this model (Regression with t-error) to various datasets and concluded that it can handle outliers and address robustness concerns practically and routinely in a wide range of settings. However, discriminating true outlier from non-outlier is still a challenge especially in case of high dimensional genomic data. The key complication in handling the problem of outlier is that distinguishing mild outlier from regular observations and masking of true outlier¹⁶. In high dimensional genomic data, where no of markers (p)

Centre for Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute, 110012, New Delhi, India. ✉e-mail: dwij.mishra@gmail.com

are greater than no of individuals (n) creates a problem termed as large p small n problem ($p > n$). This is a very common phenomena in genomics and molecular biology research now a days. In such cases, penalized regression based approach such as Least Absolute Shrinkage and Selection Operator (LASSO) could be a preferable choice as it takes care of $n \ll p$ problem by shrinking the estimates of some less significant markers and dropping others from the model. Increased use of LASSO has been motivated by plenty of high dimensional biological data. But it becomes very crucial when some influential observations are present in high dimensional genomic data as each observation has tremendous effect on model selection and interpretation. So it is quite imperative to examine effect of influential observation before implementing the LASSO regression. Hence new measure for detection of influential observation in high dimensional genomic data is a need of hour for improving GEBVs.

Rajaratnam *et al.*¹⁷ recently developed approach for outlier detection for high dimensional data by considering the LASSO regression technique. In their approach they have proposed four measures i.e. df-model, df-lambda, df-regpath and df-cvpath for detection of influential observations influenced by different aspect of LASSO regression directly or indirectly. However, the results coming from these measures are not consistent i.e. different influential points are detected from these measures. In order to produce more concrete and consistent results, a meta-analysis based approach can be applied where an improved measure of outlier detection can be developed based on integration of these measures using p-values¹⁸⁻²⁰.

In this study, an improved measure for detection of influential observation has been developed using above mentioned approach. Performance of the developed measure has been empirically evaluated and it was observed that the outliers detected from this measure are more accurate. This developed method has been implemented in the case of genomic selection data (real and simulated) and results shows that there is remarkable improvement in the prediction accuracy of GEBVs.

Material and Methods

LASSO was first time introduced by Tibshirani²¹. LASSO minimizes the sum of squares of residuals subject to a constraint on sum of absolute values of regression coefficients. It is different from usual regression as it adds some additional penalty to usual regression estimator. So it diminishes the effect of less important β s (i.e. marker effect) and reduces least important β s as zero.

Statistical formulation of LASSO estimates can be defined as:

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p x_{ij} \beta_j \right)^2 \quad (1)$$

subject to $\sum_{j=1}^p |\beta_j| \leq t, i = 1, \dots, n$ (individuals), $j = 1, \dots, p$ (markers), Y_i is the phenotypic value for individual i , x_{ij} is an element of the incidence matrix corresponding to individual i and marker j , β_j is the marker effect for marker j . It has been assumed that response variable has zero mean. The constraints $\sum_{j=1}^p |\beta_j| \leq t$ shrinks effects of variables and sets some of them to zero.

We can also write the LASSO problem in the equivalent Lagrangian form:

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\} \quad (2)$$

Here $\sum_{j=1}^p |\beta_j|$ is l_1 norm penalty on β which results in sparsity of solution and λ is a regularization parameter. Computing the LASSO solution is quadratic programming problem which can be obtained through efficient algorithm like Least Angle Regression (LARS)²². Other important question to be addressed in this case is calculation of upper limit of sum of absolute value of predictor variable, for this cross validation approach can be used²³.

Here we have used a recently proposed approach for detection of influential observation based on LASSO technique¹⁷. They proposed four different measure i.e. df-model- it measure the change in model selected; df-lambda: it measure the change in λ , where λ is a regularization parameter in LASSO regression path, df-regpath: it measure the changes observed in LASSO regularization path and df-cvpath which observe changes in LASSO cross-validation path. These measures detects outlier from high dimensional genomic data based on LASSO regression. It can be observed that all these measures i.e. df-model, df-lambda, df-regpath and df-cvpath detects influential observations which affects model directly or indirectly, has difference in their results regarding detection of influential observation, it means that there is lack of concordance among them. In order to overcome this limitation, we have proposed a more robust measure for detection of influential observation by integrating above discussed measure using p-values based meta-analysis approach.

Approach of proposed measure. In order to develop a robust statistics for detection of influential measure, we have used p-value based meta-analysis approach. In this approach, we have combined the above mentioned four measures on the basis of their p-values. We used various methods for combining these p-values and explored the performance of each method. The brief description of this approach has been as follows. Let's say, there are K independent test and their corresponding p-values are p_1, p_2, \dots, p_K . Under H_0 , it is assumed that p-values from different methods (for individual observations) are uniformly distributed between 0 and 1 (i.e. $p_k \sim U [0, 1]$). To get overall statistical significance for the hypothesis under test (H_0 i.e. null hypothesis vs. H_1 alternative hypothesis), individual p-values for each observation/genotype from different methods (i.e. df-model, df-lambda, df-regpath and df-cvpath) can be combined. Methods used for this purpose has been summarized in Table 1.

Methods	Test Statistic	Transformed Variable	Dist. under H_0	Reference
Inverse Chi-Square/Fisher	$L = \sum_{k=1}^K Z_k$	$Z_k = -2\log p_k$	χ_{2K}^2	20,39
Logit/Pearson	$S = \sum_{k=1}^K S_k$	$S_k = \log\left[\frac{p_k}{1-p_k}\right]$	t_{5K+4}	40
Meanp	$W = (0.5 - \bar{p})\sqrt{12k}$	$\bar{p} = \sum_{k=1}^K p_k / K$	$N(0, 1)$	19
Sumz/Stouffer's method	$Z = \frac{\sum_{k=1}^K w_k z(p_k)}{\sqrt{\sum_{k=1}^K w_k^2}}$	NA	$N(0, 1)$	41

Table 1. List of methods used in study for combining p-value to calculate overall significance. where p_k : Statistical significance value from k^{th} methods for an individual or genotype; K: Different methods for which p-values to be combined; df: degrees of freedom; N(): Normal distribution; t: Central t-distribution; χ^2 : Central Chi-square-distribution.

Using this approach (Table 1), the final statistical significance value i.e. combined p-values for selected observation/genotype has been calculated and influential observation is identified based on suitable p-value cut-off. Source code for our proposed approach can also be accessed from github repository at <https://github.com/BudhlakotiN/OGS>.

Experimental dataset. In order to check the robustness of our approach the same has been validated using real data. We have used total six datasets in the current study. Detailed discussion regarding each of dataset is given below.

Dataset 1: Wheat. Wheat lines were genotyped using 1447 Diversity Array Technology markers generated by Triticaret Pty. Ltd. (Canberra, Australia; <http://www.triticarte.com.au>). These markers may take two different values i.e. their presence (1) or absence (0). 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 number of DArT markers after edition was 1279 hence same has been used in this study. Same has been also used in genomic prediction study^{24,25}.

Dataset 2: Maize. Maize dataset is generated by CIMMYT's Global Maize Program²⁴. It originally include 300 maize line with 1148 SNP markers. For marker with highest frequency is coded as 0 and lowest frequency as 1. Here trait under study is also GY, evaluated under draught and watered conditions. The average minor allele frequency in these data sets was 0.20. After some editing 264 maize lines with 1135 SNPs markers were available for final study²⁴.

Dataset 3–6: Wheat. This wheat dataset is generated from CIMMYT semiarid wheat breeding program which is comprised of 254 advanced wheat breeding lines genotyped for 1726 DArT markers²⁶. This dataset is recorded for four phenotypic traits i.e. Days to heading (DTH), Thousand Kernel Weight (TKW), Yield (under irrigated condition hence denoted as Y_I), Yield (under draught condition i.e. Y_D). For convenience, here trait DTH is considered as Dataset-3, trait TKW as Dataset-4, trait Y_I as Dataset-5 and trait Y_D as Dataset-6.

Simulation. For illustration simulated data were generated using QTL Bayesian interval mapping (“qtlbim”)²⁷, a R based (R Development Core Team 2019) package. R is available at <http://www.r-project.org> and qtlbim package can be loaded from R library. This package has been used in various studies for simulation of data related to genomic selection^{28–30}. The qtlbim package uses Cockerham's model as the underlying genetic model. We have simulated a total of three data sets for genotypic and phenotype information. Here we have created range of diversified genetic architecture i.e. with very low heritability 0.10 to medium 0.5 and high heritability 0.7. Accordingly, we have simulated data at these particular heritability levels. For each stage we have simulated data for 1000 SNPs for 200 individuals. Simulated data have 10 chromosomes with 100 SNPs in each with specified length. Total 1000 markers are distributed over the all 10 chromosomes in such a way that each marker is equi-spaced over the chromosome. We have simulated normally distributed phenotype, with further no genotype or phenotype information missing. In order to check the sensitivity of all methods to detect true outlier, we have replaced 5% of observation and made them outlier (i.e. beyond mean $\pm 3*SD$). Overview of whole workflow of the current study presented in Fig. 1.

Evaluation measure. As an evaluation measure, prediction accuracy and prediction error were used. Prediction accuracy can be defined as Pearson correlation coefficient (r) between observed phenotypic value and predicted phenotypic value.

If $\hat{Y} = X\hat{\beta}$, where \hat{Y} is estimated response and $\hat{\beta}$ is estimated value of β , then correlation coefficient (r) can be expressed in following form:

$$r = \frac{S_Y \hat{Y}}{S_Y S_{\hat{Y}}} \quad (3)$$

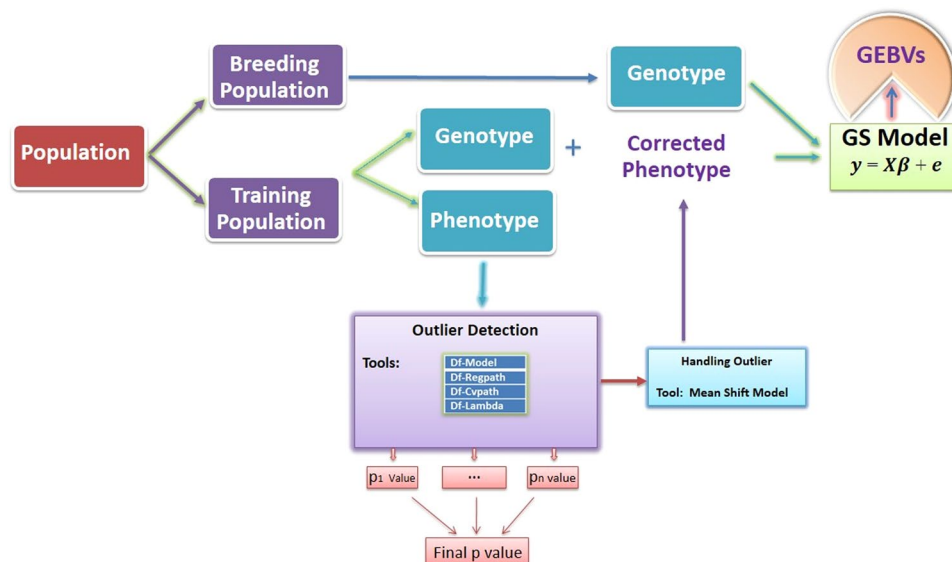


Figure 1. Operational workflow of the whole procedure used in the study.

Heritability	Methods	Precision	Recall	F1
0.1	Df-Model	0.55	0.6	0.57
	Df-Regpath	0.875	0.7	0.78
	Df-Cvpath	0.85	0.6	0.71
	Df-lambda	0.46	0.6	0.52
	sumz	0.8	0.8	0.8
	Inverse Chi	0.89	0.8	0.84
0.5	Df-Model	0.875	0.7	0.78
	Df-Regpath	0.8	0.8	0.8
	Df-Cvpath	1	0.7	0.83
	Df-lambda	0.4	0.4	0.4
	sumz	0.7	0.9	0.8
	Inverse Chi	0.8	1	0.9
0.7	Df-Model	0.73	0.8	0.76
	Df-Regpath	0.89	0.8	0.84
	Df-Cvpath	1	0.6	0.77
	Df-lambda	0.66	0.6	0.63
	sumz	0.8	0.8	0.8
	Inverse Chi	0.9	0.9	0.95

Table 2. Performance of different methods (in terms of Precision, Recall and F1 score) for different simulated datasets.

where $S_{\hat{y}}$ 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. Prediction Error (PE) can be simply defined as mean sum of square error (MSE) between observed phenotypic value and predicted phenotypic value. Same can be expressed using following formula (Eq. 3).

$$PE/MSE = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (Y_i - \hat{Y}_i)^2 \tag{4}$$

Where Y_i is observed response; \hat{Y}_i is predicted phenotypic value. It can be understood that n is the total no. of individual's i.e. $n = n_{train} + n_{test}$, here n_{train} denotes no of individuals in the training set and n_{test} is no. of individuals in test set.

In order to assess performance of methods to identify true outlier (observation with added noise) and non-outlier (observation without any noise), we have used precision (i.e. proportion of True Positive (TP) to total positives (i.e. sum total of true positive (TP) and False Negative (FN), Eq. 5), recall (i.e. proportion of TP to TP

No. of individuals (n)	No. of markers (p)	Time (Minutes)
500	2000	305
	5000	560
	10000	965
750	2000	638
	5000	1367
	10000	2191
1000	2000	1182
	5000	2292
	10000	2570

Table 3. Time required for running the datasets of varying combination of dimension using our proposed approach.

and False Negative (FN), Eq. 6) and F1 score (i.e. harmonic mean of precision and recall, Eq. 7). All these can be computed using the following expressions:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (5)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (6)$$

$$\text{F1} = \sqrt{\text{Precision} \times \text{Recall}} \quad (7)$$

To calculate the overall performance of different methods, Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) has been used. It is a multi-criteria based decision making method given by Hwang and Yoon³¹. It is based on the impression that the selected alternative should have the shortest geometric distance from the positive ideal solution (PIS) and the longest geometric distance from the negative ideal solution (NIS)³². TOPSIS compares set of alternatives by giving some weights to each criteria followed by normalization of each single criteria and calculates the geometric distance between each alternative and ideal alternative. TOPSIS is based on the assumption of that criteria are monotonically increasing or decreasing. Here final rank has been calculated using R package ‘topsis’ motivated from TOPSIS method.

Results and Discussion

Performance of our proposed method that how well it distinguishes true outlier from non-outlier, various measures like precision, recall and F1 score has been calculated for different datasets (generated at different heritability $h^2 = 0.1, 0.5$ and 0.7) and presented in Table 2. It suggest that our proposed approach i.e. based on combining p-value outperformed in almost every scenario.

Computational efficiency. The time required to compute our proposed measures is calculated using an Intel(R) Core(TM) i7-5500U CPU@2.40 GHz processor on a dataset with varying dimension (i.e. no. of individuals (500, 750 & 1000) and markers (2000, 5000 & 10000) with all possible combination). Results of same is presented below in Table 3.

In order to understand the effect of outlier on the genomic prediction accuracy, we have studied their effects on real dataset. First of all we have fitted LASSO regression with original experimental data say it as LASSO*. Then using the approach given by Rajaratnam *et al.*¹⁷, we have calculated p-values for all the four measures i.e. df-model, df-lambda, df-regpath, df-cvpath followed by combining these p-values into single value for each observation/genotype. Using the same we have identified the outlier in the response. The outlier and their corresponding marker genotype were dropped from the model and again LASSO is refitted using the modified data. In order to check robustness of our proposed approach, we have also fitted some of most commonly used methods for genomic selection i.e. Ridge Regression, Best Linear Unbiased Prediction (BLUP), Genomic-BLUP (GBLUP) and Bayesian methods. BLUP i.e. Best Linear Unbiased Prediction introduced by Henderson³³ is used in a linear mixed model for prediction of random effects. GBLUP is an improved version of BLUP where additive genomic relationship matrix (G) is used as a variance-covariance matrix of random effect in the model³⁴. For performance evaluations of methods under study, cross validation techniques is used. Data is divided into two parts i.e. training and testing sets such that training set comprises of 70% data and testing set of 30%. Former one is used for model building and later one for model evaluation. The performance of methods was evaluated by calculating prediction accuracy and prediction error. Whole procedures is repeated 100 times and prediction accuracy and prediction error is averaged and their respective standard error is calculated. Results of the same has been discussed below. Here Tables 4–9 reports the average prediction accuracy and prediction error (i.e. MSE) with their sampling variability (SE i.e. standard error) of the methods under study for dataset 1–6. In order to calculate gain in prediction accuracy all the fitted model were compared to baseline model i.e. LASSO and percentage change in prediction accuracy is calculated. In same way percentage reduction in MSE is also calculated.

Methods	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
Inverse Chi	0.62	0.52	0.05	0.06	41	36.6
Logit	0.60	0.53	0.04	0.05	36.3	35.4
Meanp	0.59	0.56	0.06	0.05	34	31.7
Sumz	0.59	0.54	0.05	0.05	34	34.2
Regression with t-error	0.47	1.21	0.06	0.06	6.8	0
RR	0.56	0.60	0.05	0.06	27.3	27
GBLUP	0.60	0.81	0.05	0.06	36.3	0
Bayesian LASSO	0.57	0.61	0.06	0.07	29.5	25.6

Table 4. Mean and standard error of prediction accuracy and prediction error for various methods using dataset 1.

Methods	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
Inverse Chi	0.43	0.69	0.08	0.10	66	28.1
Logit	0.40	0.70	0.10	0.11	53.8	28.2
Meanp	0.34	0.83	0.08	0.14	30.8	13.5
Sumz	0.44	0.70	0.09	0.13	69	28
Regression with t-error	0.36	7.2	0.09	0.10	38.5	0
RR	0.46	0.72	0.08	0.11	77	26
GBLUP	0.48	0.71	0.07	0.11	84.6	26
Bayesian LASSO	0.45	0.68	0.09	0.10	73.1	29.2

Table 5. Mean and standard error of prediction accuracy and prediction error for various methods using dataset 2.

Here all the analysis has been carried out using R (R Development Core Team 2019). LASSO model is fitted using R package `glmnet`³⁵, other methods like BLUP, GBLUP are fitted using `rrBLUP` package³⁶ with `mixed.solve` and `kin.blup` function respectively. Ridge regression is fitted using Gustavo de los Campos R code, fitting this require heritability of underlying trait. For better description, heritability for each traits under study is provided in the supplementary material (Table S1). Regression with t-error fitted using R package “hett” (using `tlm` function)³⁷. Degree of freedom is estimated for different dataset used in study by using the `tlm` function with option (`estDof = TRUE`), available in R package “hett” and then t-regression is fitted.

In this Table 4 and others (Tables 5–9) LASSO* represents LASSO regression fitted in original data (i.e. without any treatment to possible outlier), next four methods in the table represent performance of LASSO in the absence of outlier (i.e. possible outlier and corresponding genotype marker genotype dropped from the model detected using LASSO diagnostic) whereas next four methods in the table represent performance of LASSO in the absence of outlier (i.e. possible outlier and their corresponding marker genotype to be dropped from original data detected by our various p-value based meta-analysis approach). Last four methods shows the performance of other methods on our proposed approach.

In order to assess gain in the prediction accuracy for different datasets under study, It could be observed that there is significant amount of gain in prediction accuracy (Tables 4–9) as compare to their counterparts (41% increase in case of dataset 1, 69% for dataset2, 31% for dataset 3, 57% for dataset 4, 36% for dataset 5 and 27% for dataset 6). In case of Prediction error it can be observed from results (Tables 4–9) that MSE for our proposed approach has been significantly reduced (i.e. 37% for dataset 1, 28% for dataset 2, 46% for dataset 3, 57% for dataset 4, 40% for dataset 5 and 43% for dataset 6). It shows clear advantage of our integrated approach (i.e. p-value based meta-analysis method) over the existing approach. In order to see that gain in terms of predictions performance is not only restricted to LASSO, we have also investigated the performance of integrated approach by

Methods	Accuracy	MSE	Accuracy SE	MSE SE	Percentage(%) gain in Accuracy	Percentage(%) reduction in MSE
LASSO*	0.52	13.3	0.1	2.7	NA	NA
Df-Model	0.60	9.5	0.09	2.2	15.4	28.6
Df-Regpath	0.57	10.8	0.08	1.9	9.6	18.8
Df-Cvpath	0.60	9.4	0.08	2.1	15.4	29.3
Df-Lambda	0.58	10.7	0.08	2.2	11.5	19.5
Inverse Chi	0.68	7.2	0.08	1.4	30.8	45.8
Logit	0.67	7.5	0.08	1.5	28.8	43.6
Meanp	0.66	7.9	0.08	1.5	26.9	40.6
Sumz	0.68	7.3	0.07	1.6	30.8	45.1
Regression with t-error	0.61	7.4	0.08	1.8	17.3	44.4
RR	0.68	7.5	0.07	1.3	30.8	43.6
GBLUP	0.65	7.5	0.07	1.5	25	43.6
Bayesian LASSO	0.62	8.9	0.8	1.6	19.2	33.1

Table 6. Mean and standard error of prediction accuracy and prediction error for various methods using dataset 3.

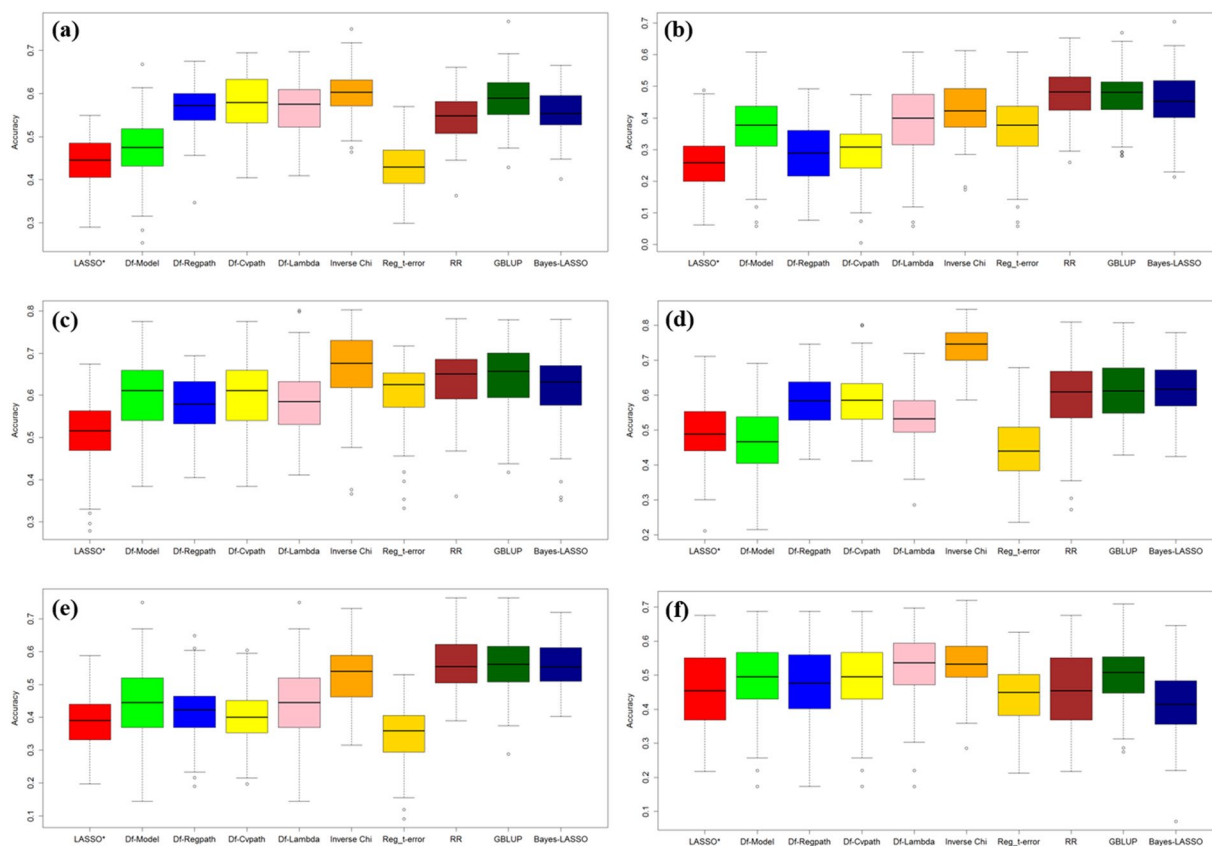


Figure 2. Box plot of prediction accuracy for different methods under study using various datasets (a) dataset 1 (b) dataset 2 (c) dataset 3 (d) dataset 4 (e) dataset 5 (f) dataset 6.

using most commonly used GS models (RR, GBLUP etc.). It can be marked with confidence that gain in terms of prediction performance has been maintained to other methods also (Tables 4–9).

In Fig. 2, each graph (Fig. 2a–f) contains the ten box plot for prediction accuracy for dataset 1–6 respectively. In each figure first box plot shows the prediction accuracy by fitting simple LASSO regression, next four box plot shows the prediction accuracy calculated following the approach of Rajaratnam *et al.*¹⁷ and next method (Inverse Chi) represent performance of LASSO in the absence of outlier (i.e. possible outlier and their corresponding marker genotype to be dropped from original data detected by p-value based meta-analysis approach i.e. Inverse Chi). Last four methods shows the performance of other GS methods on our proposed approach. These Box plots shows the distribution of prediction accuracy with their SE, estimated over 100 replications.

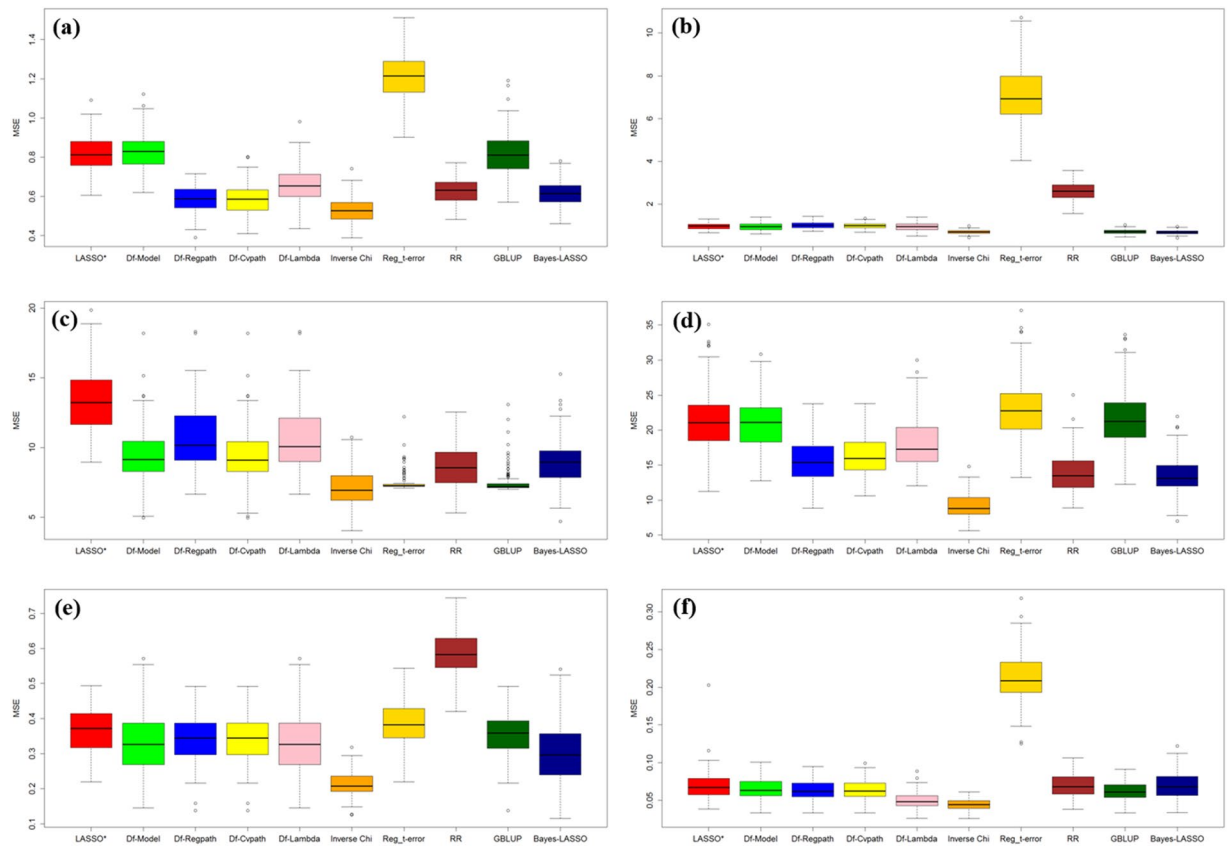


Figure 3. Box plot of prediction error (MSE) for different methods under study using various datasets (a) dataset 1 (b) dataset 2 (c) dataset 3 (d) dataset 4 (e) dataset 5 (f) dataset 6.

Methods	Accuracy	MSE	Accuracy SE	MSE SE	Percentage(%) gain in Accuracy	Percentage(%) reduction in MSE
LASSO*	0.47	21.6	0.09	4.5	NA	NA
Df-Model	0.47	20.8	0.10	3.9	0	6.3
Df-Regpath	0.59	15.8	0.07	3.1	25.5	28.8
Df-Cvpath	0.58	16.4	0.07	3.4	23.4	26.1
Df-Lambda	0.55	18.2	0.07	3.9	17	18
Inverse Chi	0.74	9.2	0.06	1.8	57.5	57.2
Logit	0.73	9.9	0.07	2.2	55.3	55.4
Meanp	0.70	11.2	0.07	2.5	48.9	49.5
Sumz	0.72	10.6	0.06	2.3	53.2	52.3
Regression with t-error	0.45	23.2	0.10	4.5	0	0
RR	0.70	10.9	0.06	2.5	48.9	49.5
GBLUP	0.64	21.7	0.08	5.3	36.2	2.3
Bayesian LASSO	0.62	13.5	0.08	2.7	31.9	39.2

Table 7. Mean and standard error of prediction accuracy and prediction error for various methods using dataset 4.

In Fig. 3, each graph (Fig. 3a–f) represents ten box plot for prediction error for dataset 1–6 on the same pattern of boxplot to Fig. 2. These boxplots represents the distribution of the MSE values over 100 runs. These plots (Figs. 2 and 3) show a clear cut advantage of our proposed approach over the LASSO diagnostic given by Rajaratnam *et al.*¹⁷, in improving genomic prediction accuracy and other existing approach. In almost every scenario i.e. wheat and maize dataset (dataset 1–6), prediction accuracy has been improved and prediction error get minimized. Clear distinctions of estimated accuracy and prediction error shows the importance of outlier detection for estimating more accurate GEBVs leads to enhanced prediction accuracy. It can be summarized from the Tables 4–9 that among p-value combination methods Inverse Chi, logit and sumz performed equally although advantage goes to Inverse Chi and sumz over logit and meanp method.

Methods	Accuracy	MSE	Accuracy SE	MSE SE	Percentage(%) gain in Accuracy	Percentage(%) reduction in MSE
LASSO*	0.38	0.37	0.10	0.06	NA	NA
Df-Model	0.44	0.33	0.09	0.06	13.6	10.8
Df-Regpath	0.42	0.35	0.09	0.08	9.1	5.4
Df-Cvpath	0.40	0.35	0.09	0.07	4.5	5.4
Df-Lambda	0.44	0.33	0.09	0.08	13.6	10.8
Inverse Chi	0.54	0.22	0.08	0.03	36.4	40.5
Logit	0.54	0.22	0.08	0.03	36.4	40.5
Meanp	0.52	0.25	0.08	0.03	31.8	32.4
Sumz	0.54	0.22	0.08	0.03	36.4	40.5
Regression with t-error	0.34	0.38	0.10	0.06	0	0
RR	0.53	0.22	0.08	0.10	39.5	40.5
GBLUP	0.57	0.35	0.09	0.10	50	5.4
Bayesian LASSO	0.55	0.30	0.10	0.10	44.73	18.9

Table 8. Mean and standard error of prediction accuracy and prediction error for various methods using dataset 5.

Methods	Accuracy	MSE	Accuracy SE	MSE SE	Percentage(%) gain in Accuracy	Percentage(%) reduction in MSE
LASSO*	0.45	0.07	0.10	0.02	NA	NA
Df-Model	0.49	0.06	0.10	0.02	8.9	14.3
Df-Regpath	0.47	0.06	0.10	0.02	4.5	14.3
Df-Cvpath	0.49	0.06	0.10	0.02	8.9	14.3
Df-Lambda	0.51	0.05	0.09	0.009	13.4	28.6
Inverse Chi	0.55	0.04	0.07	0.006	22.2	42.9
Logit	0.54	0.04	0.07	0.007	20	42.9
Meanp	0.57	0.04	0.08	0.007	26.67	42.9
Sumz	0.54	0.04	0.08	0.007	20	42.9
Regression with t-error	0.44	0.22	0.10	0.10	0	0
RR	0.52	0.04	0.08	0.01	15.5	42.9
GBLUP	0.50	0.06	0.09	0.02	11	14.3
Bayesian LASSO	0.42	0.07	0.10	0.01	0	0

Table 9. Mean and standard error of prediction accuracy and prediction error for various methods using dataset 6.

Accuracy/Data	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Dataset 6
Original Data fitted using BLUP	0.50	0.42	0.60	0.62	0.45	0.40
Mean Shift fitted using BLUP	0.97	0.68	0.93	0.81	0.85	0.82
Mean Shift fitted using GBLUP	0.98	0.77	0.95	0.88	0.94	0.86

Table 10. Effect of Mean Shift Model over baseline model on accuracy of the genomic prediction using BLUP.

Ranking of the various methods used for performance evaluation has been done using multi criteria based decision method called TOPSIS. Result of same has been given in Tables S2 and S3 (Supplementary Information). It can be concluded from Tables S2 and S3 that our integrated approach (based on p-value meta-analysis) using Inverse Chi method ranked first among other p-value based meta-analysis methods (i.e. logit, meanp and sumz) for both in case of dataset 1 and dataset 2 and same pattern has been observed for other datasets also using TOPSIS methods based on multi criteria.

Mean shift as substitute for deletion. Instead for deleting the observation flagged as outlier here we have substituted the outlier with the mean shift of data using mean shift outlier model (MSOM)³⁸. Here one or more observation is assumed to be introduced from a shifted location as compare to remaining observation. This method can be important for robust modelling where we identify the observation flagged as outlier with separate mean shift effect instead of dropping them from model. Earlier we have fitted the model to real and simulated data and for each observation outliers are identified (p-value < 0.05) based on p-value combination approach.

Here instead of deleting the observation flagged as outlier, we have replaced them with separate mean shift effect (using MSOM).

Best linear unbiased prediction i.e. BLUP³³ and GBLUP³⁴ model is fitted on original data and data where outliers are treated with MSOM. A Significant improvement in the accuracy over baseline model (using original data as such) has been observed. Details of same is presented in Table 10.

Conclusion

Impact of outlier on genomic prediction accuracy has been explored. In this study, a new efficient method using meta-analysis for outlier detection in genomic data has been proposed. It has been shown that by implementing efficient diagnostic measure for outlier detection, accuracy of GS model can be improved. Comparative study has been made among various existing methods of outlier detection in high dimensional genomic data for their impact on accuracy of genomic estimated breeding value. It has been observed that our proposed method outperformed among existing methods.

Data availability

All secondary datasets used in this study are publicly available.

Received: 16 October 2019; Accepted: 28 April 2020;

Published online: 21 May 2020

References

- Hayes, B. & Goddard, M. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829 (2001).
- Jannink, J.-L., Lorenz, A. J. & Iwata, H. Genomic selection in plant breeding: from theory to practice. *Briefings in functional genomics* **9**, 166–177 (2010).
- Zhao, Y., Mente, M. F. & Reif, J. C. Genomic selection in hybrid breeding. *Plant Breeding* **134**, 1–10 (2015).
- Hayes, B. J., Bowman, P. J., Chamberlain, A. & Goddard, M. Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of dairy science* **92**, 433–443 (2009).
- Daetwyler, H. D., Swan, A. A., van der Werf, J. H. & Hayes, B. J. Accuracy of pedigree and genomic predictions of carcass and novel meat quality traits in multi-breed sheep data assessed by cross-validation. *Genetics Selection Evolution* **44**, 33 (2012).
- Daetwyler, H., Kemper, K., Van der Werf, J. & Hayes, B. Components of the accuracy of genomic prediction in a multi-breed sheep population. *Journal of animal science* **90**, 3375–3384 (2012).
- Wang, C. *et al.* Accuracy of genomic prediction using an evenly spaced, low-density single nucleotide polymorphism panel in broiler chickens. *Poultry science* **92**, 1712–1723 (2013).
- Atkinson, A. & PLOTS, T. Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis. *Oxford Statistical Science Series, Oxford University Press: Oxford* (1985).
- Belsley, D. A., Kuh, E. & Welsch, R. Identifying influential data and sources of collinearity. *Regression Diagnostics* (1980).
- Cook, R. D. Detection of influential observation in linear regression. *Technometrics* **19**, 15–18 (1977).
- Cook, R. D. Influential observations in linear regression. *Journal of the American Statistical Association* **74**, 169–174 (1979).
- Peña, D. A new statistic for influence in linear regression. *Technometrics* **47**, 1–12 (2005).
- Geweke, J. Bayesian treatment of the independent Student-t linear model. *Journal of applied econometrics* **8**, S19–S40 (1993).
- Jylänki, P., Vanhatalo, J. & Vehtari, A. Robust Gaussian process regression with a Student-t likelihood. *Journal of Machine Learning Research* **12**, 3227–3257 (2011).
- Lange, K. L., Little, R. J. & Taylor, J. M. Robust statistical modeling using the t distribution. *Journal of the American Statistical Association* **84**, 881–896 (1989).
- Lourenço, V. M. & Pires, A. M. M-regression, false discovery rates and outlier detection with application to genetic association studies. *Computational Statistics & Data Analysis* **78**, 33–42 (2014).
- Rajaratnam, B., Roberts, S., Sparks, D. & Yu, H. Influence Diagnostics for High-Dimensional Lasso Regression. *Journal of Computational and Graphical Statistics*, 1–14 (2019).
- Edgington, E. S. An additive method for combining probability values from independent experiments. *The Journal of Psychology* **80**, 351–363 (1972).
- Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A. & Song, F. *Methods for meta-analysis in medical research*. Vol. 348 (Wiley Chichester, 2000).
- Won, S., Morris, N., Lu, Q. & Elston, R. C. Choosing an optimal method to combine P-values. *Statistics in medicine* **28**, 1537–1553 (2009).
- Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**, 267–288 (1996).
- Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. Least angle regression. *The Annals of statistics* **32**, 407–499 (2004).
- Usai, M. G., Goddard, M. E. & Hayes, B. J. LASSO with cross-validation for genomic selection. *Genetics research* **91**, 427–436 (2009).
- Crossa, J. *et al.* Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* **186**, 713–724 (2010).
- Cuevas, J. *et al.* Genomic prediction of genotype × environment interaction kernel regression models. *The Plant Genome* **9** (2016).
- Poland, J. *et al.* Genomic selection in wheat breeding using genotyping-by-sequencing. *The Plant Genome* **5**, 103–113 (2012).
- Yandell, B. S. *et al.* R/qtlbim: QTL with Bayesian interval mapping in experimental crosses. *Bioinformatics* **23**, 641–643 (2007).
- Yi, N. *et al.* An efficient Bayesian model selection approach for interacting quantitative trait loci models with many effects. *Genetics* **176**, 1865–1877 (2007).
- Yi, N. & Banerjee, S. Hierarchical generalized linear models for multiple quantitative trait locus mapping. *Genetics* **181**, 1101–1113 (2009).
- Piao, Z. *et al.* Bayesian dissection for genetic architecture of traits associated with nitrogen utilization efficiency in rice. *African Journal of Biotechnology* **8** (2009).
- Hwang, C.-L. & Yoon, K. In *Multiple attribute decision making* 58–191 (Springer, 1981).
- Assari, A. & Assari, E. Role of public participation in sustainability of historical city: usage of TOPSIS method. *Indian Journal of Science and Technology* **5**, 2289–2294 (2012).
- Henderson, C. R. Estimation of changes in herd environment. *Journal of Dairy Science* **32**, 706–715 (1949).
- Endelman, J. B. & Jannink, J.-L. Shrinkage estimation of the realized relationship matrix. *G3: Genes, Genomes, Genetics* **2**, 1405–1413 (2012).
- Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* **33**, 1 (2010).

36. Endelman, J. B. Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome* **4**, 250–255 (2011).
37. Taylor, J. & Taylor, M. J. hett: Heteroscedastic t-Regression. R package version 0.3-2. <https://CRAN.R-project.org/package=hett>. (2018).
38. Tanaka, E. Simple robust genomic prediction and outlier detection for a multi-environmental field trial. *arXiv preprint arXiv:1807.07268* (2018).
39. Fisher, R. (Edinburgh, 1932).
40. Mudholkar, G. & George, E. In *Symposium on optimizing methods in statistics*. 345–366 (Academic Press New York).
41. Stouffer, S., Suchman, E., Devinney, L., Star, S. & Williams, R. (Princeton: Princeton University Press).

Author contributions

Conceived the idea: A.R., N.B.; Designed the study: A.R., D.C.M., N.B.; Collected and analyzed the data: N.B., D.C.M.; Developed the methodology/approach: N.B., D.C.M.; Drafted the manuscript: N.B.; corrected the manuscript: N.B., D.C.M. and A.R. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41598-020-65323-3>.

Correspondence and requests for materials should be addressed to D.C.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020