*Indian Journal of Agricultural Sciences* **89** (10): 1708–13, October 2019/Article
https://doi.org/10.56093/ijas.v89i10.94633

# Identification of genetic markers for increasing agricultural productivity: An empirical study

SAYANTI GUHA MAJUMDAR[1], ANIL RAI[2] and D C MISHRA[3]

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

Received: 26 February 2019; Accepted: 10 April 2019

ABSTRACT

Genomic selection (GS) has been used globally for increasing agricultural production and productivity. It has been used for complex quantitative traits by selecting breeding material after predicting Genomic Estimated Breeding Values (GEBVs) of target species. The accuracy of GS for estimation of GEBVs depends on various factors including sampling population, genetic architecture of target species, statistical models, etc. The feature (marker) selection is one of the important steps in development of GS models. There are large numbers of models proposed in the literature for GS. However, applicability of these models is based on many factors including extent of additive and epistatic effects of breeding population. Therefore, there is strong need to evaluate the performance of these models and techniques of feature selection under different situations. In this study, performance of linear/additive effect models, viz. linear least squared regression, BLUP, LASSO, ridge regression, SpAM as well as non-linear/epistatic effect models, viz. mRMR, HSIC LASSO have been evaluated through a simulation study in R platform. In general, performance of SpAM was found to be superior for GS than all other models considered in this study in case of presence of additive effect and absence of epistatic effect. However, in case of low heritability and high epistatic effect the HSIC LASSO outperformed all models. This study will assist researcher in selection of appropriate feature selection technique for a given situation.

**Key words:** BLUP, Genomic Selection, LASSO, mRMR, QTL, Regression, SpAM

Genomic selection (GS) is an alternative form of marker-assisted selection which uses whole-genome molecular markers so that all Quantitative Trait Loci (QTL) remain in linkage disequilibrium with at least one marker. The use of high-density markers is one of the fundamental features of GS. The advancement of next generation sequencing technology made genomic selection method more popular for breeding, as it accelerates the genetic gain at reduced cost by shortening the breeding cycle. The GS has been used for complex quantitative traits by using whole-genome markers to predict Genomic Estimated Breeding Values (GEBVs) of target species. The accuracy of GS for estimation of GEBVs depends on various factors including—(i) selected model, (ii) training sample size, (iii) relatedness of training and breeding population, (iv) marker density, (v) gene effects, (vi) heritability and genetic architecture, and (vii) extent and distribution of LD between markers and QTL. Accuracy also varies among GS models according to their assumptions and treatments of marker effects.

The feature (marker) selection is one of the important steps in development of GS models. Some models perform well with markers having additive effect while other models are useful in presence of epistasis. Therefore, there is strong need to evaluate the performance of these models and techniques of feature selection under different situations. In this article, performance of linear/additive effect models and non-linear/epistatic effect models have been evaluated through a simulation study. Recently, numbers of linear statistical models were proposed for Genomics Selection (GS) which are useful for modeling additive effects and do not capture epistatic genetic effect. Some important and popular linear/additive models which we have considered in this study are (i) Linear least-squared regression (LSR) applied by Meuwissen *et al.* (2001), (ii) Least Absolute Shrinkage and Selection Operator (LASSO) of Tibshirani (1996), (iii) Ridge regression by Hoerl and Kennard (1970) and (iv) Best Linear Unbiased Prediction (BLUP) of Henderson (1949). Non-linear GS models usually suitable for modeling epistatic effects except Sparse Additive Model which is non-linear but it performs very well in case of additive data. Some important non-linear models which we have considered are: (i) Sparse Additive Models (SpAM) proposed by Ravikumar *et al.* (2009) and subsequently used by Liu *et al.* (2009), Raskutti *et al.* (2012),Suzuki and Sugiyama (2013); (ii) Minimum Redundancy Maximum Relevance (mRMR) of Ding *et al.*

Present address: [1]Ph D Scholar (sayanti23gm@gmail.com), [2]Head and Principal Scientist (anilrai64@gmail.com), [3]Scientist (dwij.mishra@gmail.com), Division of Bioinformatics, ICAR-IASRI.

(2005) and (iii) Hilbert-Schmidt Independence Criterion (HSIC) (Gretton *et al.* 2005) LASSO proposed by Yamada *et al.* (2013). The SpAM is used for high dimensional feature selection. A disadvantage of SpAM is that it can only deal with additive effects. In case of epistatic effects in the data, SpAM may fail to select significant markers. Ding and Peng employed an approach of mRMR to find the optimal subset of multiple genes. This method can rank features based on their relevance to the target and, at the same time, the redundancy of features are also penalized. This method only selects significant markers but does not provide any coefficient value to predict the marker effect. So, if we want to estimate genomic estimated breeding value (GEBV), we need to use another model after feature selection to estimate markers effects. Here, the above GS models have been considered for evaluation of their performances through a simulation study.

## MATERIALS AND METHODS

The data for this study was simulated using R package QTL Bayesian interval mapping ("qtlbim") (Yandell *et al.* 2012). R was downloaded from http://www.r-project.org, the *qtlbim* package was accessed through library (qtlbim) in R. The description of the package can be found at http://cran.rproject.org/web/packages/qtlbim/qtlbim.pdf. The package *qtlbim* follows Cockerham's model for simulating quantitative trait loci with epistasis (Kao *et al.* 2002).

In this study we have simulated 24 different datasets with genotypic and phenotypic information for $F_2$ population. Each dataset has different combination of genetic architecture, i.e. different heritability (narrow sense heritability) levels, viz. 0.1, 0.2, 0.3, 0.5, 0.7, 0.9 with various epistatic effects, viz. 0, 5, 10, 15. For each dataset, we have generated 50 replicates and for each replicate, we randomly selected 80% of data for training and 20% data was kept for testing. This randomly selected training-testing dataset was repeated 50 times for each replicate. Each dataset contains 200 individual and 1000 biallelic markers. Out of the 200 individuals, 160 were chosen randomly for the training set to fit the model and 40 individuals remain in the testing set to predict the phenotype. A genome with 10 chromosomes is simulated, each chromosome having a specified length. The 1000 markers were distributed throughout the genome so that each chromosome had 100 markers which were equally spaced over the chromosomes. Also no missing genotypic values and no missing phenotypic values are considered for simulation. The phenotypic values are normally distributed. For the additive model, there is one QTL in each chromosome with either positive or negative additive effect and no epistatic interaction. For epistatic model, we assumed two QTLs on each of the five chromosome and remaining five chromosomes have no QTL. Thus we got 5 two-way epistatic interactions.

We evaluated linear feature selection methods including least squares regression, ridge regression, BLUP, LASSO and non-linear feature selection methods such as sparse additive model (SpAM), minimum redundancy maximum relevance (mRMR) and HSIC LASSO. Then the performance of each method was evaluated by the accuracy of prediction, the mean square error (MSE), fraction of correctly selected features and the redundancy rate (RED). Accuracy of prediction is defined as the correlation between the actual phenotypic values and the predicted phenotypic values (Howard *et al.*, 2014).

$$\textit{Prediction accuracy (PA)} = \textit{correlation}\ (y_{actual}, y_{pred})$$

The RED score (Zhao *et al.* 2010) is determined by

$$RED = \frac{1}{m(m-1)} \sum_{u_k, u_j, k>l} \left|\rho_{k,l}\right|$$

where $\rho_{k,l}$ is the correlation coefficient between the $k$ -th and $l$ -th features. A large RED score signifies that selected features are more strongly correlated to each other which means many redundant features are selected. Thus, a small redundancy rate is preferable for feature selection.

Fraction of correctly selected features is measured by

$$F = \frac{N_c}{N_t}$$

where $F$ is fraction of correctly selected features, $N_c$ is the number of correctly selected features, $N_t$ is the number of total selected features.

Then the performance of the models will be evaluated by estimating mean squared error, given as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

where $\hat{y}$ is the predicted value of phenotype y, $y$ is the actual value of phenotype y, and $n$ is the number of observations.

In order to implement the linear and non-linear methods, the statistical software R was used. Specifications of the parameters and inputs for each method are described in Table 1.

## RESULTS AND DISCUSSION

In this study, we have compared the performance of four linear feature selection methods i.e. LSR, LASSO, RR and BLUP and three non-linear feature selection methods, i.e. SpAM, mRMR and HSIC LASSO on the basis of Prediction Accuracy (PA), Fraction of correctly selected features (F), Redundancy rate (RED) and Mean Squared Error (MSE). The standard error (SE) of mean is also calculated for PA, F, RED and MSE. Each feature selection method was applied to 50 sets of simulated progeny for each combination of genetic architecture. Each set contains 50 training-testing data sets which yield 2500 total replicates for each combination.

We evaluated the GS models for four different epistatic effects (E) for each level of heritability ($h^2$). The results are represented graphically.

*Prediction Accuracy:* In case of additive effect, i.e. when epistatic effect is 0, the Prediction Accuracy (PA) of Least Squared Regression (LSR) and Sparse Additive Model (SpAM) is highest and comparable. The performance of BLUP is also found to be reasonable. The PA increases

Table 1 Implementation details of different feature selection methods for Genomic Selection

| Method | Type | Implementation in R |
|---|---|---|
| Least squares regression (LSR) | Linear | It may be noted that, in datasets we have more number of markers than the number of individuals. The approach of Meuwissen *et al.* (2001) was followed and simple linear regression model was fitted for each of the markers and QTL i.e. we have fitted 1010 simple linear regression models. Then we chose 100 of the markers having most significant P-values. Further, these 100 markers were included into a final model to simultaneously fit a linear regression model. In order to fit this linear regression model, the lm function was used that can be found in the stats package (R Development Core Team 2017) in R. Out of these 100 markers we chose top 10 highly significant 10 markers as final selected feature. Finally, phenotypic values were predicted for testing data set by using available marker data and estimated regression coefficient of marker effects. |
| BLUP | Linear | The mixed.solve function in R has been used to implement BLUP. The function was available in the rrBLUP package (Endelman 2011). The marker data were used as the design matrix for the random effects, and there was no fixed effect in this model. The model was fitted using training data. Based on top ten highly significant coefficients of this model, markers were selected as features of this model. Then the prediction of phenotypic value was performed using the marker data for the testing data set and the predicted coefficients of marker effects from this model. |
| LASSO | Linear | In order to implement the LASSO method, we used the glmnet function of the glmnet package (Friedman *et al.* 2010) in R with default parameter values. Markers with non-zero coefficients are chosen as selected features. The prediction was performed with the help of tuning parameter $\lambda$ which minimized the average cross-validation error (cvm). |
| Ridge Regression (RR) | Linear | Ridge regression can also be implemented through glmnet function of the glmnet package (Friedman *et al.* 2010) in R with the value of alpha equal to zero. In this case also, top ten highly significant coefficients were used to select corresponding markers. Then the prediction was performed in testing set with the help of predict function using the previously fitted model in the training set. |
| SpAM | Additive non-linear | The sparse additive model was fitted using samQL function of the SAM package (Zhao *et al.* 2014) in R with default parameter values. In this case also, top 10 highly significant markers are chosen as selected features based on corresponding coefficients. The prediction of phenotypic values was performed with the help of predict function using the model fitted on the training dataset. |
| mRMR | Non-linear | Minimum redundancy maximum relevance method was implemented through the mRMRe package (Jay *et al.* 2017). In this case, first mRMR data function was used to create an mRMR data object and then with this data we performed feature selection with the help of mRMR.classic function. |
| HSIC LASSO | Non-linear | In order to perform HSIC LASSO, we have written a function in R to kernelize the input variable matrix and the output vector. Then penalized function of penalized package (Goeman *et al.* 2018) has been used to fit this Kernelized LASSO model. In order to maintain uniformity and compare the performances of different method, we have selected most significant 10 markers as selected feature among all non-zero coefficients. Then predict function is used to predict the phenotypic value of the testing dataset. |

with increase in heritability, when, epistatic effect is 0, for both models. We can observe the above mentioned findings in Fig 1(a). It has also been observed that standard error of PA is quite low for both models. In case of datasets with epistatic effects, highest prediction accuracy is recorded for HSIC LASSO (Fig 1(b-d)) with an exception for the datasets having heritability level 0.7 with epistatic effect 5 and heritability level 0.9 with epistatic effects 5 and 10 (Fig 1b, 1c). This may be due to the fact that in case of high heritability level there may be domination of additive effect over epistatic effect as in this case of LSR, SpAM and BLUP performs better than HSIC LASSO. Also, the PA increases with increase in epistatic effect at a fixed level of heritability for HSIC LASSO and PA of other models tend to decrease with increase in epistatic effect at a fixed level of heritability. From Fig 1 it can be observed that the PA increases for different models when heritability increases at a particular level of epistatic effect. It can also be observed from these results that HSIC LASSO performs much better than other models in case of low heritability and high epistatic effect (Fig 1c, 1d), whereas, Sparse Additive Models perform well for highly additive data in most of the cases (Fig 1a).

*Fraction of Correctly Selected Features:* In case of additive effect, i.e. 0 value of epistatic effect, SpAM is able to select highest number of features correctly followed by LSR (Fig 2(a)). The performance of BLUP and mRMR are closer to each other and both are able to select same

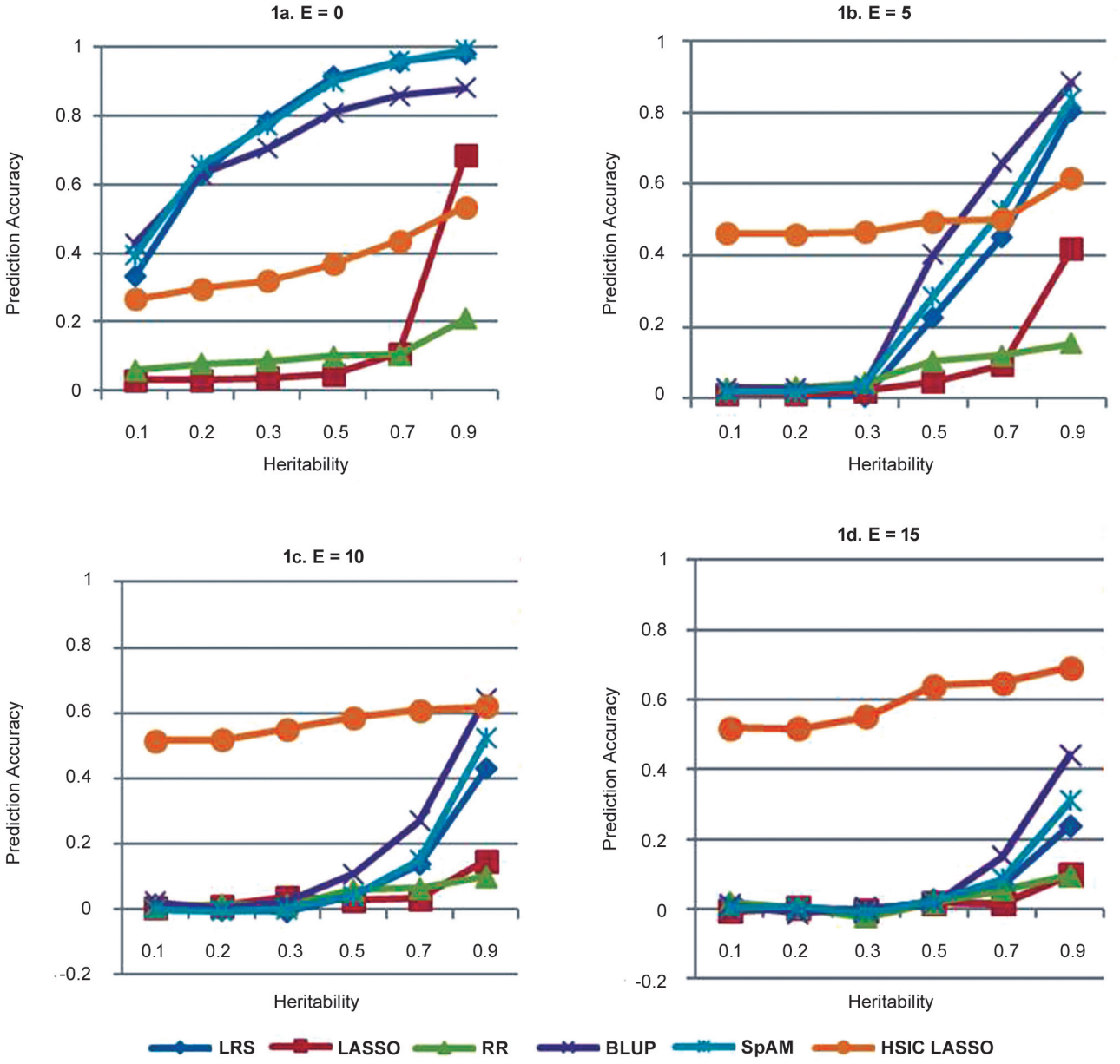


Fig 1 Prediction Accuracy at different level of epistatic effect (E)

number features. Also as the heritability increases, SpAM is able to select all features correctly above 0.3 heritability value. The fraction of correctly selected features in case of HSIC LASSO increases with increase in epistatic effect, however, in case of moderate and high heritability levels, the fraction of correctly selected features ($F$) for BLUP is highest irrespective of level of epistatic effect except E=0 (Fig 2b, 2c, 2d).

*RED and MSE:* It has been observed from the study that the RED, i.e. the redundancy of the selected features has not been affected by increasing heritability or epistatic effect and it is almost same across different methods. The MSE of HSIC LASSO was found to be least among all the methods considered in this study irrespective of level of heritability and epistatic effect. This may be due to the fact that, it is a non-linear method, where, MSE may not be highly desirable. However, among the additive effect models, performance of LSR with respect to MSE is highest, i.e. it has lowest MSE irrespective of level of heritability and epistatic effects.

In the above study, the performance of different feature selection methods was assessed for GS through a simulation study. First, the data was simulated with a carefully defined architecture and then seven different linear and non-linear methods of GS were applied to these datasets. The results were compared with the help of different relevant measurement and assessment techniques. From the results, it can be conclude that on the basis of the entire performance criterion, SpAM performs overall best in case of additive data, i.e. when epistatic effect is 0. In case of epistatic data, HSIC LASSO was found to perform superior on the basis all criterion considered for this study. Also, it is desirable

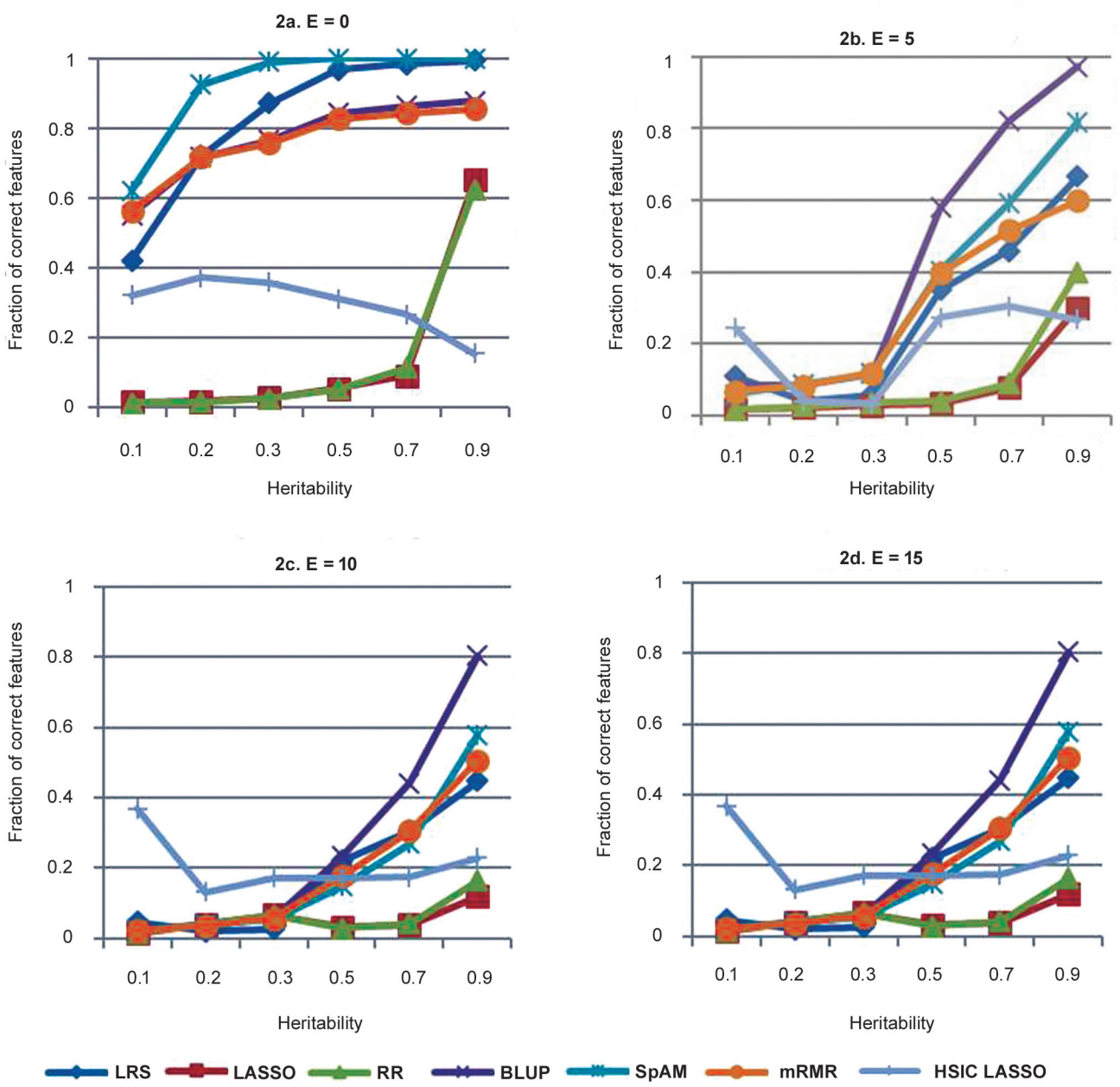


Fig 2 Fraction of correctly selected features at different level of epistatic effect (E)

to apply HSIC LASSO in case of low heritability and high epistatic effect.

## ACKNOWLEDGEMENTS

The first author is thankful for the fellowship received from ICAR-IASRI for the PhD programme. The facilities provided by ICAR-IARI and ICAR-IASRI are duly acknowledged.

## REFERENCES

Ding C and Peng H. 2005. Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology* **3**(2): 185–205.

Endelman J B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* **4**: 250–5.

Friedman J, Hastie T and Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**: 1–22.

Goeman J J. 2010. L1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal* **52**(1): 70–84.

Gretton A, Bousquet O, Smola A and Scholkopf B. 2005. Measuring statistical dependence with Hilbert-Schmidt norms, pp 63–77. *Algorithmic Learning Theory*. Springer.

Henderson C R. 1975. Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**(2): 423–47.

Hoerl A E and Kennard R W. 1970. Ridge regression: biased estimation for non-orthogonal problems. *Technometrics* **12**: 55–67.

Hoerl A E and Kennard R W. 1970. Ridge regression: applications to non-orthogonal problems. *Technometrics* **12**: 69–82.

Howard R, Carriquiry A L and Beavis W D. 2014. Parametric

and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. *G3 (Bethesda)* **4**(6): 1027–46.

Jay N D, Cavanagh S P, Olsen C, Hachem N E, Bontempi G and Haibe-Kains B. 2013. mRMRe: an R package for parallelized mRMR ensemble feature selection. *Bioinformatics* **29**(18): 2365–68.

Kao C H and Zeng Z B. 2002. Modeling epistasis of quantitative trait loci using Cockerham's model. *Genetics* **160**: 1243–61.

Liu H, Lafferty J and Wasserman L. 2009. Nonparametric regression and classification with joint sparsity constraints, pp 969–76. (In) Advances in Neural Information Processing Systems.

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.

Peng, H, Long, F and Ding, C. 2005. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**: 1226–37.

R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

Raskutti G, Wainwright M and Yu B. 2012. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research* **13**: 389–427.

Ravikumar P, Lafferty J, Liu H and Wasserman L. 2009. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**(5): 1009–30.

Suzuki T and Sugiyama M. 2013. Fast learning rate of multiple kernel learning: Trade-off between sparsity and smoothness. *The Annals of Statistics* **41**(3): 1381–405.

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

Yamada M, Jitkrittum W, Sigal L, Xing E P and Sugiyama M. 2014. High-dimensional feature selection by feature-wise kernelized Lasso. *Neural Computation* **26**: 185–207.

Yandell B S, Mehta T, Banerjee S, Shriner D, Venkataraman R *et al.* 2007. R/qtlbim: QTL with Bayesian Interval Mapping in experimental crosses. *Bioinformatics* **23**: 641–43.

Yandell B S, Nengjun Y, Mehta T, Banerjee S, Shriner D *et al.* 2012. qtlbim: QTL Bayesian Interval Mapping. R package version 2.0.5. URL: http://CRAN.R-project.org/package=qtlbim

Zhao Z, Wang L and Li H. 2010. Efficient spectral feature selection with minimum redundancy, pp 673–78. (In) AAAI Conference on Artificial Intelligence.

Zhao T, Li X, Liu H and Roeder K. 2014. SAM: Sparse Additive Modelling. R package version 1.0.5. URL: https://CRAN.R-project.org/package=SAM