Special Proceedings: ISBN #: 978-81-950383-0-5
23rd Annual Conference, 24-28 February 2021; pp 101-111

Innovations in Genomic Selection: Statistical Perspective

Dwijesh Chandra Mishra, Neeraj Budhlakoti, Sayanti Guha Majumdar and Anil Rai
ICAR-Indian Agricultural Statistics Research Institute, New Delhi

Received: 08 June 2021; Revised: 25 June 2021; Accepted: 29 June 2021

Abstract

Genomic selection is a modified form of Marker Assisted Selection in which the markers from the whole genome is used to estimate Genomic Estimated Breeding Value (GEBV). The population of individuals with both phenotypic and genotypic data is used to estimate model parameters that are subsequently be used to calculate GEBVs of selection candidates having only genotypic data. These GEBVs are then further be used to select the individuals for the purpose of advancement in the breeding cycle. Several estimators are available to estimate GEBV. However, various issues such as high dimensionality of the data, multicollinearity among the markers, a smaller number of individuals and a greater number of markers (large *p* and small *n* problem) are the major challenges in estimation of GEBVs. This paper discusses most commonly used methods for estimation of GEBVs, *viz*., Ridge Regression, Genomic Best Linear Unbiased Prediction (GBLUP), Bayesian Alphabets and Least Absolute Shrinkage and Selection Operator (LASSO) with the aim to meet the challenges associated with estimation of GEBVs. Apart from this, some semi and non-parametric methods of genomic selection have been discussed as well. Moreover, another problem like presence of outliers in the data of genomic selection has also been conversed. Furthermore, a case study deals with non-linearity of the data has also been presented and illustrated using multi traits data. At the end, some future directives of research in this area are highlighted.

Key words: Genomic estimated breeding values; Statistical models; Genomic best linear unbiased prediction (GBLUP); Bayesian methods; Least absolute shrinkage and selection operator (LASSO).

1. Introduction

Right from the beginning of agriculture, farmers used to select the best plants on the basis of their phenotypic characters such as higher yields, larger seeds, or sweeter fruits for the purpose of growing them for next season. In this way, they tried to alter the genetic makeup of plants. Afterward, farmers came to know about artificial mating of the plants by cross pollination, from which breeding approach emerged out. Breeding approach is basically a process by which humans use animals and plants to selectively develop particular phenotypic trait (characteristics) by choosing or selecting best males and females of animal or plant which can sexually reproduce and have offspring together. Breeding approach was more advanced than traditional approach. However, genetic gain through this technique was found to be very low, time-consuming and inefficient, especially when; the traits under consideration have low genetic variance (low heritability), traits are limited to particular sex (sex-limited traits) and when generation interval is large or traits appear late in the life. Other

Corresponding Author: Dwijesh Mishra
Email: dwij.mishra@gmail.com

limitation of this approach was that one does not know about the genetic basis of the transmission of traits from parents to their offspring consequently which causes the problem of Linkage drag (*i.e*., Transfer of genes governing undesirable trait along with the gene of interest).

Later on demerits of breeding approach was overcome by Marker Assisted Selection (MAS) based breeding. Where molecular markers associated with the trait of interest are used to select the superior plant for breeding purpose. It is a simpler method as compared to phenotypic screening used in traditional breeding, especially for the traits which require laborious screening. Through this approach, time and resources may be saved as selection can be done even at seedling stage for the important trait like grain quality which appears late in their life cycle. It also enhances reliability as there is no environmental effects play a role in the selection as in the case of traditional breeding. According to one study, 8-38% extra genetic gain can be observed when marker information's are included in Best Linear Unbiased Prediction (BLUP) methodology for the prediction of breeding value (Meuwissen *et al*., 1996). However, there are certain limitations of this approach such as Marker should be tightly linked to the trait of interest, it cannot be used in polygenic trait or Quantitative trait, where multiple minor genes play a role in governing the particular trait. But it is well known fact that most traits of economic interest in agriculture are quantitative in nature. Another problem related to this approach is that every marker needs to be statistically tested with respect to their association with the trait of interest which causes Multiple Hypothesis Testing Problem.

In order to overcome the limitations of MAS, Meuwissen et al. (2001) proposed a variant of MAS that is known as Genomic Selection (GS). It is a form of marker-assisted selection in which genetic markers covering the whole genome are used to identify quantitative trait loci (QTL) which are in linkage disequilibrium (LD) with at least one marker. Genomic Selection has been successful and the main reason behind the success is that it incorporates all markers information in the prediction model. In this approach, a prediction equation on training population containing phenotypic as well as genotypic data is generated and subsequent prediction of the breeding values of the individuals (testing population) having only genotypic data is carried out. Breeding value calculated by prediction equation is termed as Genomic Estimated Breeding Value (GEBV) and depending on the outcomes of GEBV, the selection decision is made on the breeding population.

This approach is better than the above-mentioned approaches as it offers more accurate prediction of Genomic Wide Estimated Breeding Value (GW-EBV) than MAS, consequently high accuracy in breeding values with respect to desired trait. Using this approach, there is a drastic reduction of breeding interval than traditional breeding, faster genetic gains (more than 30% reported in animals) and long-term low cost of breeding which ultimately enhances the production and productivity which ensure the food and nutritional security. However, there are certain factors such as training population size, trait heritability, influence of Genotype-Environment ($G \times E$) interaction, marker density, effective population size of breeding population, (Genetic diversity of breeding population), genetic relationship between training population and selection candidates influences the accuracy of the prediction of GEBVs. Apart from this there are certain statistical issues or challenges such as large number of markers or predictors (p) as compared to small number of observations or samples (n), multi-collinearity where markers are related with each other, presence of outliers/missing data, exist which should be taken care of.

In this paper, a brief review of the models used in the prediction of Genomic Estimated Breeding Values (GEBVs) has been presented. Starting with the very simple linear model to advance nonlinear/nonparametric models with the aim to meet various statistical challenges and issues arises during prediction of GEBVs have been briefly discussed in this paper. Apart from this, a non-linear model called Multivariate Kernelized Least Absolute Shrinkage and Selection Operator (Multivariate Kernelized LASSO) has been suggested as a case study which takes care the problem of non-linearity as well as pleiotropy present in the data of genomic selection. As a concluding remark, some potential future research prospective in this area have been highlighted.

2. Methods of Genomic Selection

Predicting GEBVs on which selection of the suitable individuals is done in genomic selection starts with simple linear model.

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{Y} = n \times 1$ vector of observations; $\boldsymbol{\beta} = p \times 1$ vector of marker effects; $\boldsymbol{\varepsilon} = n \times 1$ vector of random residual effects; $\boldsymbol{X} =$ design matrix of order $n \times p$ and $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma_e^2 \mathbf{I})$.

One major problem in linear model is that number of markers exceed the number of observations (large p and small n problem ($p >> n$)and this creates a problem in parameter estimation. Subset of the significant markers can be an alternative for dealing with large $p$and small n problem.

Meuwissen *et al.* (2001) used a modification of least squares regression for GS. Performed least squares regression analysis on each maker separately with following model

$$\boldsymbol{Y} = \boldsymbol{X}j\beta j + \boldsymbol{\varepsilon}$$

where,

$\boldsymbol{X}j = j^{\text{th}}$ column of the design matrix
βj = genetic effect of j^{th} marker

Markers with significant effects are selected by plotting the log likelihood of this model against the position of the marker. The marker with significant effects (QTL) are further used for estimation of breeding value

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

where,

$\boldsymbol{X}$ = the design matrix of order $n \times q$ $(q << p)$
$\boldsymbol{\beta}$ = vector of genetic effect of order $q \times 1$
q = number of significant markers

It is very likely that multicollinearity exist among markers (explanatory variables) and this can negatively affects the performance of variable selection methods. This problem is solved by using ridge regression (Meuwissen *et al*., 2001). Here, the goal is to derive an estimator of β with smaller variance than the least square estimator. Due to tradeoff between variance and bias of an estimator, there is a price to pay as ridge regression estimator of $\boldsymbol{\beta}$ is biased. In Ridge Regression (RR), penalty function is added to normal equation. So instead of minimizing sum of square of residuals, it minimizes:

$$(\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta}) + \lambda\boldsymbol{\beta}'\boldsymbol{\beta}$$

where λ is the penalty parameter and can be chosen by variety of ways, one solution is given by (Hoerl *et al*., 1975)

$$\lambda = \frac{ps^2}{(\widehat{\boldsymbol{\beta}})'(\widehat{\boldsymbol{\beta}})}$$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

where, p = number of markers, s^2 = estimate of error variance ($i.e.\hat{\sigma}^2$).

Ruppert *et al*. (2003) showed that ridge regression is a special case of the Best Linear Unbiased Prediction (BLUP) where mixed linear model is implemented. Restricted Maximum Likelihood Estimation (REML) is a good choice for finding a realistic value for the penalty parameter and estimating the variance component. Here objective is to minimize the function-

$$(\boldsymbol{Y} - \boldsymbol{X\beta} - \boldsymbol{Zu})'\boldsymbol{R}^{-1}(\boldsymbol{Y} - \boldsymbol{X\beta} - \boldsymbol{Zu}) + \boldsymbol{\beta}'\boldsymbol{G}^{-1}\boldsymbol{\beta}$$

where $E(\boldsymbol{u}) = \boldsymbol{0}$ and $E(\boldsymbol{e}) = \boldsymbol{0}$ and $var\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{e}\end{bmatrix} = \begin{bmatrix}\boldsymbol{G} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{R}\end{bmatrix}\sigma^2$, $\boldsymbol{G}$ and $\boldsymbol{R}$ are known positive definite matrices and σ^2 is a positive constant.

Similar to RR, Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani1996; Usai *et al*., 2009) is other variant of penalized regression which can be obtained by altering the penalty function (*i.e*. by giving linear penalty). Objective function of this variant is defined as-

$$(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda|\boldsymbol{\beta}'\boldsymbol{1}|$$

This constraint shrinks some of the marker effects and sets some of them to zero.

It may be possible that not all markers have equal variance. Therefore, variance of marker positions needs to be modeled. For this purpose, the Bayesian approach has been used. In this approach, it is assumed that there is prior distribution of marker effect. Where, inferences about model parameters are obtained on the basis of posterior distribution. Several variants of Bayes such as Bayes A, Bayes B, Bayes Cπ and Bayes Dπ were proposed for estimation of GEBVs (Meuwissen *et al*., 2001 and Habier *et al*., 2011).

The Bayes A approach applies the same prior distribution for all of the variances of the marker positions whereas Bayes B assumes that not all markers contribute to the genetic variation. In Bayes A approach, inverse chi-squared probability distribution $\chi^{-2}(\vartheta,\, S^2)$ can be used as the prior distribution. It is a conjugate prior as the posterior distribution is also an inverted chi-square distribution $\chi^{-2}(\vartheta + n_j,\, S^2 + \acute{\beta}_j\beta_j)$ where n_j is the number of haplotype effects at marker position.

The Bayes B approach has a prior density on the variance that is a mixture. It has a high probability mass at $\sigma_{\beta j}$ = 0, it can be summarized *as* $\sigma_{\beta j}$ = 0 with prob = π and $\sigma_{\beta j} \sim \chi^{-2}(\vartheta,\, S)$ with prob = (1 – π).The choice of degrees of freedom and the scale

parameters of the scaled inverse chi-square distribution can influence the outcome (Gianola *et al*., 2009).

Improved Bayesian methods were developed by Habier *et al*. (2011). Bayes Cπ and Dπ are the modification of Bayes A and Bayes B where the probability π of having a zero effect SNP is estimated.

The presence of outliers in genomic as well as phenotypic data is a common phenomenon. The presence of outliers may distort the distribution and adversely affect the accuracy of genomic prediction. Presence of outliers in genomic data increases the computational time and when the size of outliers increases, the sample size increases and consequently genetic variance decreases. Rajaratnam *et al.* (2019) recently proposed an approach for detection of influential observation based on LASSO technique. They proposed four different measures *i.e*., df-model- it measures 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: it observes changes in LASSO cross-validation path.

Df-Model: This measures the changes in model selection through LASSO when observation is discarded. To quantify this change, df-model for i^{th} observation can be defined as:

$$wdf - model(i) = \frac{\delta(i) - E\{\delta(i)\}}{\sqrt{var\{\delta(i)\}}}$$

where $\delta(i) = \sum_{j=1}^{p} \left| I\{\hat{\beta_j}^{lasso} = 0\} - I\{\hat{\beta_j}^{lasso}(i) = 0\} \right|$ is termed as model difference and it simply measures the difference in the no. of selected variable for full model vs. when i^{th} observation is dropped.

Df-lambda: It measure the changes observed in regularization parameter λ for full LASSO model *vs*. when i^{th} observation is dropped. Measuring this change is important because this parameter tells that at what extent selected LASSO model is shrinking the estimates. To quantify this change, df-lambda for i^{th} observation can be defined as

$$df - lambda(i) = \frac{\hat{\lambda} - \hat{\lambda}(i) - E\{\hat{\lambda} - \hat{\lambda}(i)\}}{\sqrt{var\{\hat{\lambda} - \hat{\lambda}(i)\}}}$$

This involves fitting LASSO, $n + 1$ times then computes difference of $\widehat{\lambda} - \hat{\lambda}(i)$. $E\{\hat{\lambda} - \hat{\lambda}(i)\}$ and $var\{\hat{\lambda} - \hat{\lambda}(i)\}$ can be simply estimated using sample mean and variance of n observed value of $\widehat{\lambda} - \hat{\lambda}(i)$, cut-off for df-lambda is justified at ± 2.

Df-Regpath: It measures the deviation in the LASSO regularization when an observation is dropped from LASSO path. If a significant deviation occurs from LASSO original path it means dropped observation could have huge impact on LASSO estimates which further may affect the conclusion and interpretation for LASSO solution. Df-Regpath for i^{th} observation could be defined as

$$df - regpath(i) = \frac{\Delta_1 \hat{\beta}^{lasso}(i) - E\{\Delta_1 \hat{\beta}^{lasso}(i)\}}{\sqrt{var\{\Delta_1 \hat{\beta}^{lasso}(i)\}}}$$

where $\Delta_1\hat{\beta}^{lasso}(i) = \int_l^u \left\|\hat{\beta}^{lasso}(s) - \hat{\beta}^{lasso}(s,i)\right\| ds$ and l, u are specified interval [l, u] defines possible λ values.$\hat{\beta}^{lasso}(s)$ represents vectors of parameter estimates obtained at λ = s using LASSO with full model, whereas$\hat{\beta}^{lasso}(s,i)$ represents vector of parameter estimates obtained at $\lambda = s$ by excluding i^{th} observation from the model.

Df-Cvpath: It measures the changes in predictive performance of LASSO when an observation is dropped from LASSO path. Quantifying this is crucial as if large 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. df-cvpath for i^{th} observation could be defined as

$$df - cvpath(i) = \frac{\Delta_\gamma(i) - E\{\Delta_\gamma(i)\}}{\sqrt{var\{\Delta_\gamma(i)\}}}$$

where $\Delta_\gamma(i) = \int_l^u |\gamma(s) - \gamma(s,\ i)|\, ds$ and $\gamma(s,\ i)$ represents the cross-validation error when i^{th} observation is dropped from path.

These measures detect 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 detect influential observations which affect model directly or indirectly, have difference in their results regarding detection of influential observations as they are used for optimizing different parameters (Budhlakoti *et al*., 2020a). To overcome these limitations, Budhlakoti *et al.* (2020b) have proposed a more robust measure for detection of influential observation by integrating above discussed measures using *p*-values based meta-analysis approach (Figure 1). It has been observed that this method significantly improves the prediction accuracy of genomic selection in the presence of outliers in the data.

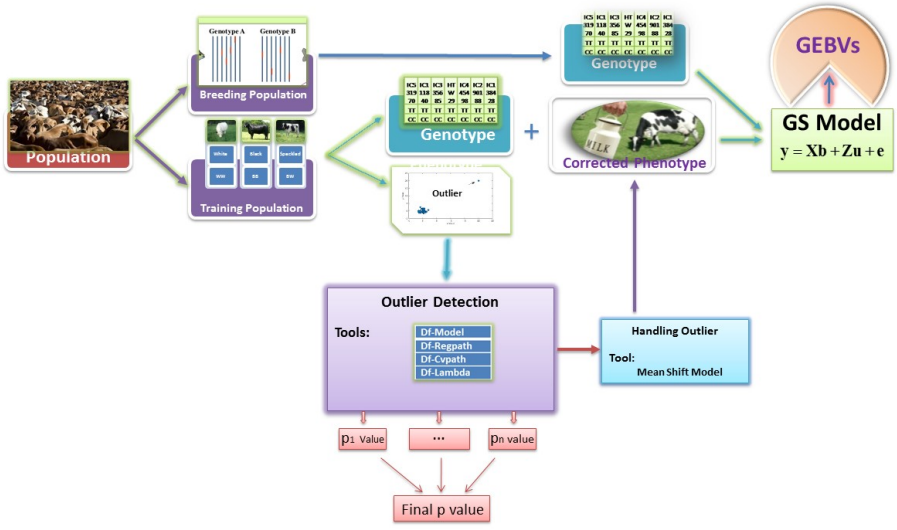

Figure 1: Workflow of the method developed by Budhlakoti *et al.* (2020) for detection of outlier in genomic selection

Genomic selection based on single trait (STGS) has been utilized successfully in recent years. But it is unable to perform well in the case of pleiotropy *i.e*., one gene links with multiple traits. A mutation in a pleiotropic gene may influence several traits simultaneously. It was found that traits with low heritability can borrow information from correlated traits and consequently achieve higher prediction accuracy. So Multi Trait Genomic Selection (MTGS) gave more accurate GEBVs than STGS for the trait with low heritability and for the trait having missing data. Jia *et al*. (2012) presented three multivariate linear models (*i.e*., GBLUP, Bayes A, and Bayes Cπ) and compared them to univariate models and a detailed comparison of various STGS and MTGS based methods also been deliberated by Budhlakoti *et al*. (2019). Moreover, the models, we generally use for GS are linear. But this assumption is generally violated. So nonlinear multi-trait-based approach may be more accurate for genomic selection.

Multivariate LASSO

This is an extension of simple LASSO model. Here the sharing involves which variables are selected, since when a variable is selected, a coefficient is fit for each response. Statistical formulation in this case is same as LASSO with some minor differences. It can be written in the form of simple statistical model as:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$$

Here all notations are as such in LASSO. Only difference is that Y is a matrix of responses instead of vector earlier. It minimizes following objective function:

$$(\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta}) + \lambda|\boldsymbol{\beta}'\mathbf{1}|$$

Kernelized Multivariate LASSO

To take advantage of higher dimensional feature spaces, we can introduce the data via nonlinear functions. For example, we can replace the inner product of the data by a kernel function. $k(\boldsymbol{X_i}, \boldsymbol{X_j}) = (\Phi(\boldsymbol{X_i})'\Phi(\boldsymbol{X_j}))$. Here we can apply so-called "kernel trick"; *i.e.* the fact that $\Phi(\boldsymbol{X_i})'\Phi(\boldsymbol{X_j}) = k(\boldsymbol{X_i}, \boldsymbol{X_j})$, we can see that $\Phi\Phi'$represents the $(n \times n)$ *Kernel Gram Matrix* $\boldsymbol{K}$ of the cross dot products between all mapped input data points $\{\Phi(\boldsymbol{X_i})\}_{i=1}^{n}$.

Some commonly used choice of kernel functions include: the Gaussian radial basis function $k(\mathbf{x}, \mathbf{z}) = exp(-\sigma\|\boldsymbol{x} - \boldsymbol{z}\|^2)$, where σ is the bandwidth parameter, the Laplace radial basis function $k(\mathbf{x}, \mathbf{z}) = exp(-\sigma\|\boldsymbol{x} - \boldsymbol{z}\|)$.

Here, we have suggested kernelized Multivariate LASSO for estimation of GEBVs. For illustration of MTGS based methods we have considered Brasica napus dataset (Kole *et al*., 2002). Dataset has 4 responses for 103 lines (individuals) genotyped for 300 markers. Lines are derived from two cultivars namely Stellar and Major. Marker genotypes are represented in 0/1, where 0 represent a Stellar allele and 1 represent Major allele. First, we applied LASSO (Multiresponse) technique for MTGS to improve the GEBVs. We have observed reasonable accuracy gain for predicted breeding value *i.e*., GEBVs for various traits under study. Then to capture nonlinearity component in populations we have also used Kernelized LASSO (Multiresponse). Very good level of accuracy has been observed for most of the traits under study.

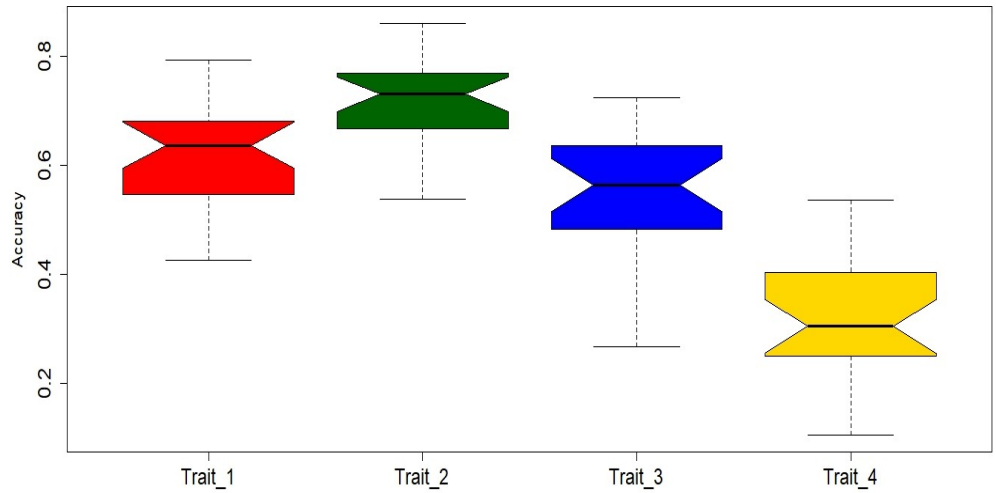

Figure 2: Performance Accuracy of the method of Multivariate Kernelized LASSO in four datasets of Brassica species

These all above discussed methods performs satisfactory well only in case of additive genetic architecture *i.e*., partitioning of genetic variance into additive, dominance, additive × additive, additive × dominance, etc. But it only holds under conditions of linkage equilibrium, random mating of male and female parents, no inbreeding, no assortative mating, no natural or artificial selection and no genotyping errors. In breeding programs, these conditions are all violated. Epistatic interaction may play a crucial role for explaining genetic variation for quantitative traits, as ignoring this kind of interaction in the model may end up with lower genomic prediction accuracy (Cooper *et al*. 2002). Gianola *et al*. (2006) first used non-parametric and semi-parametric methods for modeling complex genetic architecture, as they also include such type of higher order interaction in these models. Subsequently, several statistical methods were implemented to model both main and epistasis effects for genomic selection (Cai *et al*., 2011, Xu, 2007). Recently, some semi-parametric (Legarra *et al*., 2018) and other robust approaches (Tanaka 2020; Majumdar *et al*., 2019a; Majumdar *et al*., 2019b; Budhlakoti *et al*., 2020a; Budhlakoti *et al*., 2020b; Sehgal *et al*., 2020) have also been proposed and implemented in genomic selection.

Gianola *et al*. (2006) proposed non-parametric and semi-parametric methods to model the relationship between the phenotype and the markers that are available within the GS framework.

Nadaraya-Watson estimator

$$Y_i = g(\boldsymbol{X_i}) + e_i$$

where,
Y_i phenotypic measurement on individual i, $i = 1,2, \ldots, n$,
$\boldsymbol{X_i}$ is a $p \times 1$ vector of dummy SNP covariates observed on individual i,
$g(.)$ is some unknown function relating genotypes to phenotypes, $g(\boldsymbol{X_i}) = E(Y_i|\boldsymbol{X_i})$
where, $E(Y_i|\boldsymbol{X_i})$ is a conditional expectation of Y_i relative to $\boldsymbol{X_i}$
e_i is a residual effect for i^{th} individual and $e_i \sim (0, \sigma^2)$.

The conditional expectation can be written as

$$g(X) = \frac{\int Yp(\boldsymbol{X}, \boldsymbol{Y})dY}{p(\boldsymbol{X})}$$

Reproducing Kernal Hilbert Space

In this semi-parametric kernel mixed model approach, features of a nonparametric model are combined with a mixed model framework by *Gianola et al*. (2006). The model can be written as:

$$Y_i = \boldsymbol{w'}_i\boldsymbol{\beta} + \boldsymbol{z'}_i\boldsymbol{u} + g(\boldsymbol{X}_i) + e_i$$

where $i = 1,2, \ldots, n$ and β is a vector of fixed unknown effects (*e.g*., physical location of an individual), $\boldsymbol{u}$ is a $q \times 1$ vector which represents additive genetic effects, $\boldsymbol{w_i}$ and $\boldsymbol{z_i}$ are known incidence vectors, $g(\boldsymbol{X}_i)$ is an unknown function of the SNP data and the vector of residuals, an $\boldsymbol{e}$ is assumed to have a $N(\boldsymbol{0}, \boldsymbol{I}\sigma^2)$ distribution.

Overall summary of the methods used in genomic selection is provided in the **Figure 3:**

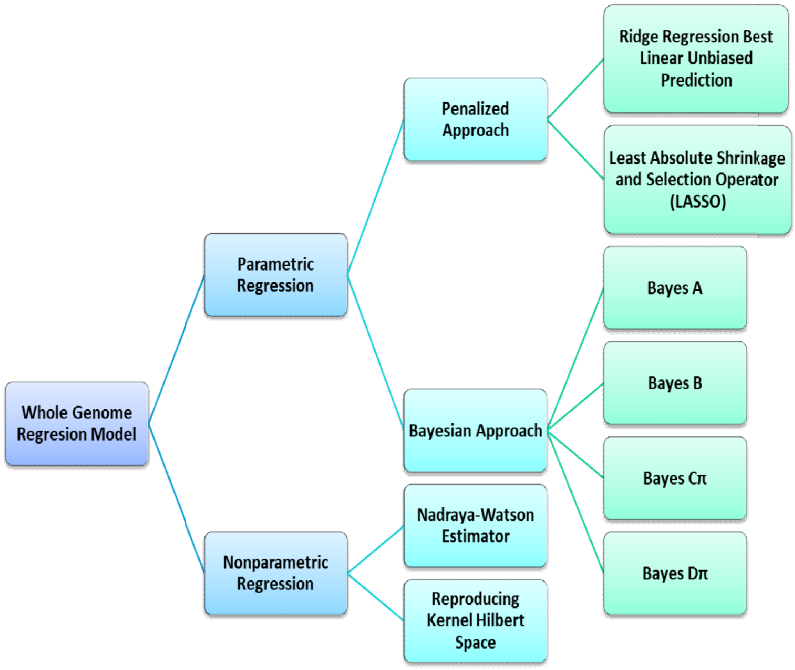

Figure 3:Overall summary of the methods used in Genomic Selection

A detailed comparison of various non-parametric methods for genomic selection at different combination of population size and heritability using simulated data was presented by Budhlakoti *et al*. (2020c). For detailed comparison and review of various genomic selection methods one can also refer to Howard *et al*. (2014).

3. Future Research Direction in This Area

Above mentioned methods of genomic selection mainly deals with genetic architecture containing additive effects, dominance effects and somewhat epistatic effects. But in real

situation, more degree of epistatic effects are present. We need to develop some more advanced statistical models or methods which could efficiently deal with epistatic effects present in the data. Moreover, since genomic selection models are based on genotypic as well as phenotypic data, therefore there is also the possibility of environmental effects and their interaction with the genetic component. Therefore, it is imperative to develop a model for genomic selection by incorporating the environmental effects and their interaction with the genotypic effects. Furthermore, genotypic data generated for genomic selection are having lot of missing data. Therefore, it is also required to develop a method which could take care of incomplete data situation. Apart from this, role of epigenetics in genomic selection can also be possible. New research initiative should be taken where epigenetic effects for genomic selection can be modelled.

References

Budhlakoti, N., Mishra, D. C., Rai, A., Lal, S. B., Chaturvedi, K. K. and Kumar, R. R. (2019). A comparative study of single-trait and multi-trait genomic selection. *Journal of Computational Biology*, **26**(**10**), 1100-1112.

Budhlakoti, N., Rai, A. and Mishra, D. C. (2020a). Effect of influential observation in genomic prediction using LASSO diagnostic. *Indian Journal of Agricultural Sciences,***90**(**6**), 1155-1159.

Budhlakoti, N., Rai, A. and Mishra, D. C. (2020b). Statistical approach for improving genomic prediction accuracy through efficient diagnostic measure of influential observation. *Scientific Reports*, **10**(**1**), 1-11.

Budhlakoti, N., Rai, A., Mishra, D. C., Jaggi, S., Kumar, M. and Rao, A. R. (2020c). Comparative study of different non-parametric genomic selection methods under diverse genetic architecture. *Indian Journal of Genetics*, **80**(**4**), 395-401.

Cai, X., Huang, A. and Xu, S. (2011). Fast empirical Bayesian LASSO for multiple quantitative trait locus mapping. *BMC Bioinformatics*, **12**(**1**), 1-13.

Gianola, D., de Los Campos, G., Hill, W. G., Manfredi, E. and Fernando, R. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics*, **183**(**1**), 347-363.

Gianola, D., Fernando, R. L. and Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics*, **173**(**3**), 1761-1776.

Habier, D., Fernando, R. L., Kizilkaya, K. and Garrick, D. J. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*, **12**(**1**), 1-12.

Hoerl, A.E., Kennard, R.W. and Baldwin, K.F. (1975). Ridge regression: some simulation. *Communications in Statistics*, **4**, 105–123.

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: Genes, Genomes, Genetics*, **4**(**6**), 1027-1046.

Jia, Y. and Jannink, J. L. (2012). Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics*, **192**(**4**), 1513-1522.

Kole, C., Thormann, C. E., Karlsson, B. H., Palta, J. P., Gaffney, P., Yandell, B. and Osborn, T. C. (2002). Comparative mapping of loci controlling winter survival and related traits in oilseed Brassica rapa and B. napus. *Molecular Breeding*, **9**(**3**), 201-210.

Legarra, A. and Reverter, A. (2018). Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. *Genetics Selection Evolution*, **50**(**1**), 1-18.

Majumdar, S. G., Mishra, D. C. and Rai, A. (2020a). Effect of genotype imputation on integrated model for genomic selection. *Journal of Crop and Weed*, **16(1)**, 133-137.

Majumdar, S. G., Rai, A. and Mishra, D. C. (2020b). Integrated framework for selection of additive and nonadditive genetic markers for genomic selection. *Journal of Computational Biology*, **27**(**6**), 845-855.

Meuwissen T. H. and Goddard M. E. (1996). The use of marker haplotypes in animal breeding schemes. *Genetics Selection Evolution*, **28**, 161-176.

Meuwissen, T. H., Hayes, B. J. and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, **157**(**4**), 1819-1829.

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**), 877-890.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. 12, Cambridge University Press.

Sehgal, D., Rosyara, U., Mondal, S., Singh, R., Poland, J. and Dreisigacker, S. (2020). Incorporating genome-wide association mapping results into genomic prediction models for grain yield and yield stability in CIMMYT spring bread wheat. *Frontiers in Plant Science*, **11**, 197.

Tanaka, E. (2020). Simple outlier detection for a multi-environmental field trial. *Biometrics*, **76(4)**, 1374-1382.

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

Usai, M. G., Goddard, M. E. and Hayes, B. J. (2009). LASSO with cross-validation for genomic selection. *Genetics Research*, **91**(**6**), 427-436.

Xu, S. (2007). An empirical Bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics*, **63**(**2**), 513-521.