

Genetic Prediction in Bovine Meat Production: Is Worth Integrating Bayesian and Machine Learning Approaches? a Comprehensive Analysis

Maria Ines Fariello¹(✉), Eileen Amstrong², and Alicia Fernandez³

¹ IMERL, Facultad de Ingenieria, UdelaR & UBi, Institut Pasteur de Montevideo, Montevideo, Uruguay

fariello@fing.edu.uy

² Facultad de Veterinaria, UdelaR, Montevideo, Uruguay

³ IIE, Facultad de Ingeniería, UdelaR, Montevideo, Uruguay

Abstract. Genomic prediction is a still growing field, as good predictions can have important economic impact in both, agronomics and health. In this article, we make a brief review and a comprehensive analysis of classical predictors used in the area. We propose a strategy to choose and ensemble of methods and to combine their results, to take advantage of the complementarity that some predictors have.

Keywords: Parametric · Non parametric · Genomic · Selection · Prediction · Fusion

1 Introduction

Beef consumers increasingly demand meat of high and consistent quality. As a consequence, research has focused on understanding muscle biology to control quality traits. In the past two decades, molecular genetics has changed dramatically animal production research. Genome sequencing has facilitated the identification of polymorphisms (here we focus on Single Nucleotide Polymorphisms: SNPs), that can be used as genetic markers in animal breeding. Genes involved in the physiological regulation of energy, body weight, triglyceride synthesis and growth are candidates that may have effects on economically important carcass and meat quality traits ([7] [6]). On the other hand, such avalanche of information has increased in a considerable way the complexity of the analysis, making obvious that the usual statistical methods may not be enough ([10] [18]). In this paper we try to predict the carcass weight from genomic information. A review of the state of the art in genetic prediction shows the interest in performance comparison between lineal regression models as Bayesian Ridge regression, Bayesian Lasso, Bayes A, B and C with non-linear models as Bayesian Regularized Neural Networks (BRNN), Reproducing Kernel Hilbert Spaces Regression (RKHS) and Support Vector Machine Regression (SVR).

In [16], the superiority of nonlinear regression methods versus linear ones is analyzed in the case of wheat genomic prediction. In [11] a comparison between Best Linear Unbiasd Prediction (BLUP) and SVR shows discrepancy between prediction accuracies obtained by cross-validation procedures and correlation ones, being more accurate BLUP when a limited set of training samples are available. In [13], a comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers is done. Fixed regression using least squares (FRLS), BLUP, bayesian regression (BayesR), partial least squares regression and SVR are compared in Australian Selection Index and protein percentage prediction. Although the selected methods have inherent differences in the underlying assumptions, they show similar performances (except FRLS, which is not recommended). In [14] methods with large conceptual differences also reached very similar predictive performances, although re-ranking of methods was observed depending on the analyzed phenotype. In [19] the effects of feature selection methods on prediction performance for different methods was observed. The authors found that feature selection and prediction algorithms should be carefully selected depending on the phenotypes. A nice review of kernel-based whole-genome prediction of complex trait is presented by [12]. They concluded that research involving analysis of raw phenotypes coupled with enviromental variables needs more attention. In recent works, like [17], the impact of predictive modes averaging is analyzed. It is proposed to combine several RKHS models with different t-kernels, but no improvements were found compared with one kernel models. Although several works compared different approaches for genomic prediction (born in breeding animal, statistics and machine learning), they use performance measures, as the prediction error, that are global statistical averages, which can hide the differences and complementarities between methods. In particular, these differences are what can make it worth a combination of methods. In pattern recognition, it is well known that the best scenario is to combine when individual methods have similar performance but bring diversity, *i.e.* different behavior in different individuals [1, 2, 8, 9].

In this paper we study the behavior of a set of known approaches for genomic prediction of carcass weight in Aberdeen Angus cattle from Uruguay. A comparative analysis of the behavior of the different methods in the sample space is presented. We propose a method to choose a subset of predictors, once their performances are computed. The proposed analysis aims to provide knowledge of the specific problematic, but also give elements for a greater understanding of the similarities and differences between approaches and to know in advance if it is worthy to use an ensemble of methods.

2 Methodology

Data Set Characterization. The database used comprise several phenotypic measures [15] from a total of 705 Aberdeen Angus animals of different age-sex categories. The animals came from ten commercial herds and were slaughtered in different slaughter houses. The database is complex, due both to the amount

of data and the diversity thereof. Apart from genotypic variables, it includes environmental effects such as age, sex and origin. To avoid dealing with fixed effects, we just considered the individuals of the most numerous herd. The trait analyzed here is carcass weight. A total of 160 SNPs were selected for genotyping from the bibliography and public genomic databases. They are located in candidate genes that take part in metabolic pathways and physiological processes related to energy expenditure, triglyceride and fatty acid synthesis, body weight and growth. After removing SNPs with minor allele frequency lower than 0.05 (to avoid bias of the data), and individuals with more than 20% of missing values or with no phenotype, there were 79 SNPs and 93 individuals left.

Prediction Methods. In genome-wide association studies the objective is to predict an individual's breeding value (here, carcass weight) from its genotype. The association between genotypes and phenotypes is modeled in a group of individuals with phenotypic and genetic information (training set). The model is then used to predict the individual phenotypes in individuals for whom only information from genetic markers is available.

The basic prediction model, that seeks to minimize the mean square error (Ordinary Least Squares (OLS)), has prediction coefficients that are unbiased estimates with variance dependent on the sample size (n), the number of coefficients prediction (p) and interdependence between the predictor variables. One way to address "the curse of dimensionality" (p large in relation to n) of OLS, which generates high variance and therefore a large mean square error (MSE), is applying regularization in the regression. This is done adding a penalty term in the optimization seeking to balance the goodness of the approximation to the complexity of the model. Ridge Regression (RR) adds an extra term to the likelihood function that reduces the regression coefficients in an amount which depends on the variance of the co-variables. The regularization introduces bias, but reduces the variance of the estimate, reducing potentially MSE estimation of the prediction coefficients. Other individual cases of regularization are Least Absolute Shrinkage and Selection Operator (LASSO) in which the penalty is the absolute value of the coefficients, instead of the squares of them (as in RR), which introduces sparsity.

In a Bayesian approach, different penalty methods can be introduced changing the priors from where the regression coefficients are sampled and the likelihood functions. The Bayesian equivalent of RR, BRR (Bayesian Ridge Regression), allows to deploy G-BLUP (BLUP using a genomic distance's matrix), which is one of the most commonly used models in genomic prediction. A set of methods that share the same likelihood function but differ in the priors, which suppose different effects of the markers is known as the Bayesian alphabet ([3]).

The problem becomes almost intractable with large p . An alternative strategy is to use semi-parametric models as proposed by Gianola [4] as RKHS (Reproducing Kernel Hilbert Space), in which the model is determined by the choice of a kernel which fixes the space in which the regression is performed, and the parameter that determines shrinkage similar to that used in RR. An alternative

method to capture additive, dominance and epistasis integrating linear and non-linear functions are complex neural networks (NN [5]). One of the distinguishing characteristics of neural networks is the flexibility to capture complex nonlinear patterns, the drawback are that increasing p with multiple neurons, increases strongly the computational requirements and tends to overfitting.

Having a set of tools as described previously, we faced the problem of defining how to use them efficiently, taking full advantage of the benefits and minimizing weaknesses. To do this we needed to define a set of measures for evaluating the performance of complementarity and/or the diversity they bring to the set. In particular, it lead us to investigate the advantages of assembly methods and how to make the assemble.

Comprehensive Analysis of Diversity of Predictors. As was shown before, different strategies have been proposed to deal with gene-trait association and genetic prediction. The studies showed that there is no method that is always superior to others in all data sets. These works make focus in MSE and they do not make a deeper comparative analysis about the diversity between the methods. They hide how the individual strategies work in the data space and if they have enough diversity between them that it could be worth embedding. Dealing with complex data sets where the traits have high dependence on environment introduce specific problems that have to be taken with care and different methods have to be used. We will discuss the relevance that different methods give to the variables, making focus in similarities and differences. We propose to study the relation between the genomic array, using its first two principal components and the error distribution in the sampling space.

Diversity Measure in Ensemble. For regression ensembles the "diversity" can be measured and quantified explicitly. The MSE can be expressed in a bias-variance-covariance decomposition for the predictors ensembles. In an assemble of methods, the predictive error depends on the bias and variance of individual predictors but also on the covariance between individuals (shown in [2]).

The default method for the assembly of regression methods is the average of the predictions of the different methods. Given that the average error is a function of the average bias, variances and covariances between methods. An improvement in performance would be expected against the individual methods. The optimal assemble choice is the one that balances the trade-off between these terms to reduce the overall MSE. Given a set of methods with similar performance in terms of individual MSE or correlation between the predicted and the real values in the training set, the ones that provide smaller covariance, *i.e.* are the most diverse, are worth to be ensemble.

Based on the above analysis, it is proposed as a criterion for the ensemble to seek "diversity" measured by the covariance between methods: (i) Select the two methods with less covariance, (ii) in an iterative way select the method with less covariance with respect to the already selected provided that the covariance

Table 1. Correlation Matrix of Predictors and the real value y: MMR: Multiple Marker Regression, RR:Ridge Regression, BC: Bayes C, BB2: Bayes B with $\pi = 10^{-4}$, RK: Regression Kernel in Hilbert Spaces, NN: Neural Networks

	y	MMR	RR	BC	BB2	RK	NN
y	1.00	-0.13	0.20	0.15	0.18	0.16	0.19
MMR	-0.13	1.00	-0.11	-0.18	-0.20	-0.16	-0.04
RR	0.20	-0.11	1.00	0.88	0.70	0.77	0.76
BC	0.15	-0.18	0.88	1.00	0.92	0.96	0.75
BB2	0.18	-0.20	0.70	0.92	1.00	0.98	0.63
RK	0.16	-0.16	0.77	0.96	0.98	1.00	0.71
NN	0.19	-0.04	0.76	0.75	0.63	0.71	1.00

is lower than a threshold, (iii) weighted average of the individual predictions is given as result.

3 Results

Relationship Between Errors and Genetic Structure. To investigate if the prediction error was related to the genomic relationship between each individual and the rest of the population, a Principal Component Analysis of the genomic matrix of the population was done. PCA is also helpful to investigate if there is a hidden substructure in the population, which could introduce confounding effects in our analysis. We suppose that if the error of a predictor is related with the genetic composition of the individuals, then individuals with the same type of errors would cluster together. Although different bayesian approaches were used, in Figure 1 only BayesC is shown because the predictors were highly correlated ($\approx 99\%$).

No obvious clusterization is observed in Figure 1. From the individual error point of view, there are some variations on the individual errors between methods, but in general the error structure remains the same, but for the multi marker regression. This predictor was negatively correlated with the real values and with the other predictors in the testing set (Table 1), so it is no longer considered in the analysis.

Combining Predictors. Although the differences between the error structure of the predictors were slight (Fig. 1), the algorithm proposed in 2 was used to investigate if there was a way of embedding predictors that predicted better than the predictors individually.

The first chosen predictors were NN and BB2, for having the smallest positive covariance. Then, the correlation between the mean of those predictors (mNN-BB2) was computed and as a result of that RK was chosen. The predictor was computed as $m2RK = (2 * mNN - BB2 + RK)/3$, to avoid underweighting the first chosen predictors. Then, the correlation between m2RK and the remaining

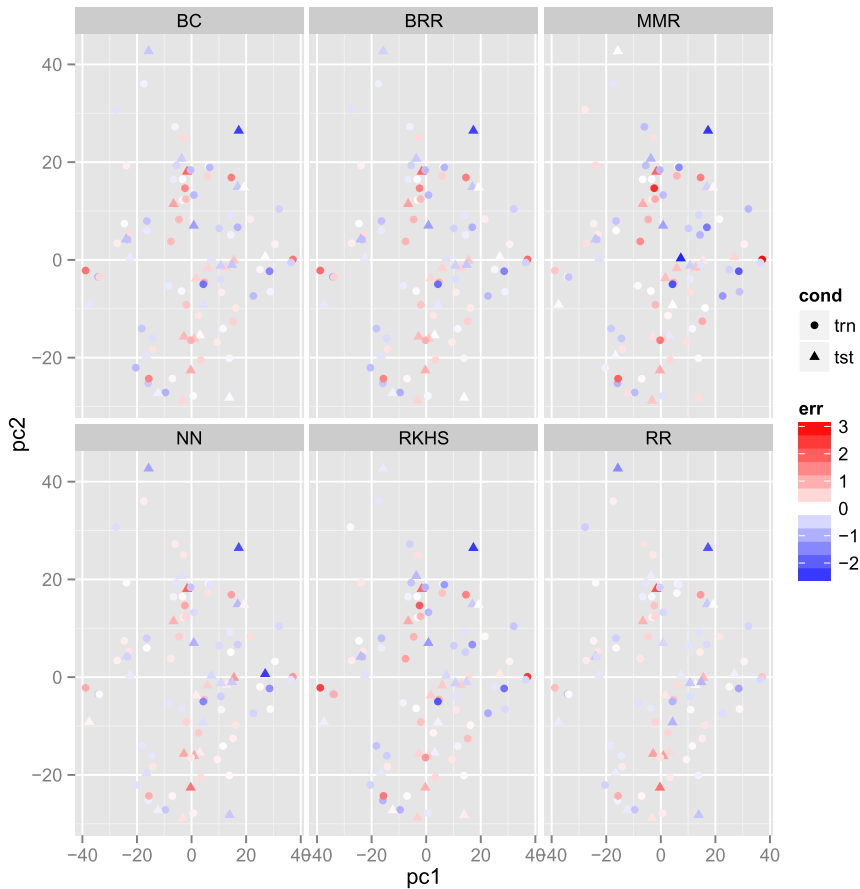


Fig. 1. Principal Component Analysis of Genomic Array. Errors in carcass weight prediction are represented with a color gradient

predictors was computed and RR was chosen. The new predictor m3RR was computed in the same way as the previous one. The correlation between the remaining predictor BC and the new one was 90%, so BC was not integrated to the predictor. The correlations of the previous steps were between 70% and 80%.

Three different combinations of the predictors were evaluated: The mean of them, the weighted mean using the correlations between each predictor (w-mean) and the

Table 2. Mean Squared Error. BRR: Bayesian Ridge Regression (= G-BLUP), BA, BB, BC and BL: Bayes A, B, C and LASSO, mean was taken over RR, BB2, RK, NN, m2RK = (mNN+BB2)/2

MMR	RR	BRR	BA	BL	BC	BB	BB2	RK	NN	mean	w-mean	m2	m2RK	m3RR
0.80	0.72	0.58	0.57	0.57	0.60	0.58	0.57	0.57	0.85	0.58	0.58	0.60	0.57	0.59

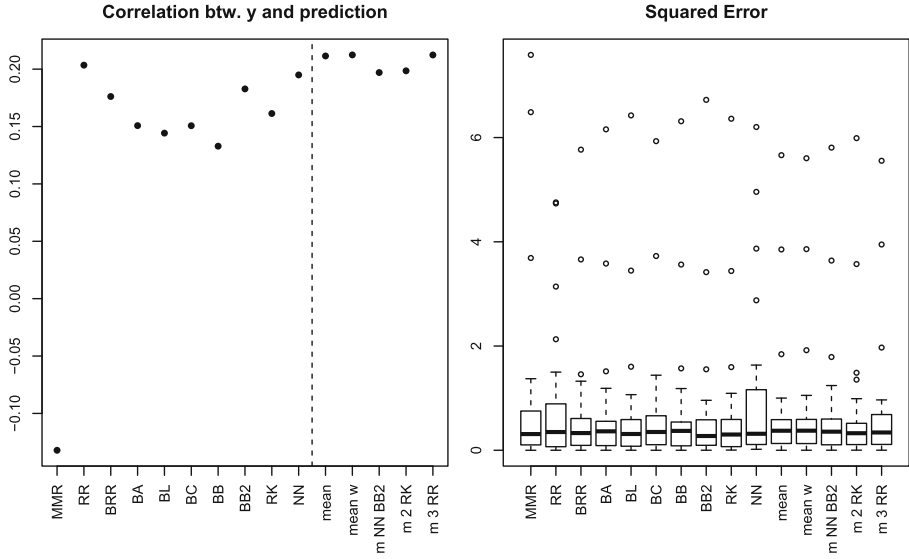


Fig. 2. Correlation between the predicted values using the different predictors and the real value (left) and distribution of the squared errors (right). Results from testing set.

real value and the predictor resulting from the algorithm (m3RR). The new estimators have almost the same MSE than the lowest observed value (0.57), and a slightly better correlation with the real value. The best correlation found between the classical predictors was the one of the Ridge Regression predictor (Fig. 2), but it has one of the worst MSE (0.80, Table 2). The new predictor has the best correlation with the phenotype and almost the lowest MSE.

4 Conclusions

A comprehensive analysis of the performances of the main methods used in genetic prediction of complex traits of high economic impact was done.

Based on the evaluation of diversity among the individuals, an ensemble strategy was proposed and evaluated. In particular, it was found that bayesian predictors have low complementarity, while BayesC (or any of the others), Ridge Regression, RKHS and Neural Networks have the highest degree of complementarity. By comparing several ways of combining the predictors, obtained by taking diversity into account, we found that the proposed criteria is consistent.

As it is not possible to know in advance which of the methods would work better, as they do not require much computation after the predictors are computed, and as the shown combinations work at least as well as the best predictor, it is worth to combine the methods.

Further reserch has to be done in order to obtain the best weights for combining these predictors, without losing interpretability of the results.

Acknowledgment. The work of Maria Ines Fariello was partially supported by ANII Posdoc research scholarship. The authors want to thanks Daniel Gianola for sharing his experience in the genetic prediction field.

References

1. Bonissone, P.P., Xue, F., Subbu, R.: Fast meta-models for local fusion of multiple predictive models. *Applied Soft Computing* **11**(2), 1529–1539 (2011)
2. Brown, G., Wyatt, J., Harris, R., Yao, X.: Diversity creation methods: a survey and categorisation. *Information Fusion* **6**(1), 5–20 (2005)
3. Gianola, D.: Priors in whole-genome regression: The bayesian alphabet returns. *Genetics* **194**(3), 573–596 (2013)
4. Gianola, D., van Kaam, J.B.C.H.M.: Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* **178**(4), 2289–2303 (2008)
5. Gianola, D., Okut, H., Weigel, K., Rosa, G.: Predicting complex quantitative traits with bayesian neural networks: a case study with jersey cows and wheat. *BMC Genetics* **12**(1), 87 (2011)
6. Golden, B., Garrick, D., Benyshek, L.: Milestones in beef cattle genetic evaluation. *J. Anim. Sci.* **87**, E3–E10 (2009)
7. Hocquette, J.F., Lehnert, S., Barendse, W., Cassar-Malek, I., Picard, B.: Recent advances in cattle functional genomics and their application to beef quality (2007)
8. Kittler, J., Alkoot, F.M.: Sum versus vote fusion in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**(1), 110–115 (2003)
9. Kuncheva, L.I.: A theoretical study on six classifier fusion strategies. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(2), 281–286 (2002)
10. Lambert, C.G., Black, L.J.: Learning from our gwas mistakes: from experimental design to scientific method. *Biostatistics* **13**(2), 195–203 (2012)
11. Maenhout, S., De Baets, B., Haesaert, G.: Prediction of maize single-cross hybrid performance: support vector machine regression versus best linear prediction. *Theoretical and Applied Genetics* **120**(2), 415–427 (2010). <http://dx.doi.org/10.1007/s00122-009-1200-5>
12. Morota, G., Gianola, D.: Kernel-based whole-genome prediction of complex traits: a review. *Frontiers in Genetics* **5** (2014)
13. Moser, G., Tier, B., Crump, R.E., Khatkar, M.S., Raadsma, H.W., et al.: A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide snp markers. *Genet. Sel. Evol.* **41**(1), 56 (2009)
14. Neves, H.H., Carnevalheiro, R., Queiroz, S.A.: A comparison of statistical methods for genomic selection in a mice population. *BMC Genetics* **13**(1), 100 (2012). <http://dx.doi.org/10.1186/1471-2156-13-100>
15. Nunes, J.L., Piquerez, M., Pujadas, L., Armstrong, E., Fernández, A., Lecumberry, F.: Beef quality parameters estimation using ultrasound and color images. *BMC Bioinformatics* **16**(Suppl. 4), S6 (2015)
16. Prez-Rodriguez, P., Gianola, D., Gonzalez-Camacho, J.M., Crossa, J., Mans, Y., Dreisigacker, S.: Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3: Genes—Genomes—Genetics* **2**(12), 1595–1605 (2012). <http://www.g3journal.org/content/2/12/1595.abstract>
17. Tusell, L., Pérez-Rodríguez, P., Forni, S., Gianola, D.: Model averaging for genome-enabled prediction with reproducing kernel hilbert spaces: a case study with pig litter size and wheat yield. *Journal of Animal Breeding and Genetics* **131**(2), 105–115 (2014)
18. Visscher, P.M., Brown, M.A., McCarthy, M.I., Yang, J.: Five years of gwas discovery. *The American Journal of Human Genetics* **90**(1), 7–24 (2012)
19. Yoon, D., Kim, Y.J., Park, T.: Phenotype prediction from genome-wide association studies: application to smoking behaviors. *BMC Systems Biology* **6**(Suppl. 2), S11 (2012). <http://www.biomedcentral.com/1752-0509/6/S2/S11>

Progress in Pattern Recognition, Image Analysis,
Computer Vision, and Applications
20th Iberoamerican Congress, CIARP 2015, Montevideo,
Uruguay, November 9-12, 2015, Proceedings
Pardo, A.; Kittler, J. (Eds.)
2015, XXI, 787 p. 242 illus. in color., Softcover
ISBN: 978-3-319-25750-1