2015 IEEE 14th International Conference on Machine Learning and Applications

# Prediction of continuous phenotypes in mouse, fly, and rice genome wide association studies with support vector regression SNPs and ridge regression classifier

Abdulrhman Aljouie and Usman Roshan
Department of Computer Science
New Jersey Institute of Technology
Newark, NJ 07102
Email: aa547@njit.edu and usman@njit.edu

***Abstract*—The ranking of SNPs and prediction of phenotypes in continuous genome wide association studies is a subject of increasing interest with applications in personalized medicine and animal and plant breeding. The ranking of SNPs in case control (discrete label) genome wide association studies has been examined in several previous studies with machine learning techniques but this is poorly explored for studies with quantitative labels. Here we study ranking of SNPs in mouse, fly, and rice continuous genome wide association studies given by the popular univariate Pearson correlation coefficient and the multivariate support vector regression and ridge regression. We perform cross-validation with the support vector regression and ridge regression models on top ranked SNPs and compute correlation coefficients between true and predicted phenotypes. Our results show that ridge regression prediction with top ranked support vector regression SNPs gives the highest accuracy. On all datasets we achieve accuracies comparable to previously published values but with fewer SNPs. Our work shows we can learn parsimonious SNP models for predicting continuous labels in genome wide studies.**

***Keywords*—Phenotype prediction, SNP selection, support vector regression, genome wide association studies, ridge regression**

## I. Introduction

The prediction of continous phenotypes has applications in breeding, farming, and medicine [1], [2]. Recent studies propose novel statistical methods to predict quantitative phenotype values in genome-wide association studies [3], [4]. Plenty of such work has been done in case control genome wide association studies in the context of disease prediction [5], [6], [7], [8], [9], [10], [11], [12], [13], [14]. There we have seen standard machine learning methods employed in clever ways to boost prediction accuracy.

In this paper we explore standard feature selection techniques [15] and two regularized risk regression methods [16] that are popular in the machine learning literature to rank SNPs and predict regression values in three genome wide association studies. Our work is different from previous feature selection studies in that we study datasets with continuous labels (regression data). In contrast previous feature selection studies focus exclusively on classification data [17], [18], [19], [20], [21], [15].

Our results show that we can predict quantitative phenotypes from a parsimonious set of SNPs instead of using tens and hundreds of thousands. We achieve comparable or higher accuracy than previously published work by rankings SNPs with the support vector regression and predicting phenotype values with the ridge regression [22].

## II. Methods

We use the univariate Pearson correlation coefficient for ranking features. We also use support vector regression and ridge regression to rank features as well as to learn a model from training data and predict regression values of validation or test data.

### A. Pearson correlation coefficient

The Pearson correlation coefficient is given by

$$\frac{\sum_i^n (x_{i,j} - x_{i,mean})(y_i - y_{mean})}{\sqrt{\sum_i^n (x_{i,j} - x_{i,mean})^2}\sqrt{\sum_i^n (y_i - y_{mean})^2}} \quad (1)$$

where $x_{i,j}$ represents the encoded value of the $j^{th}$ variant in the $i^{th}$ individual and $y_i$ is the label (+1 for case and -1 for control) of the $i^{th}$ individual. The Pearson correlation ranges between +1 and -1 where the extremes denote perfect linear correlation and 0 indicates none. We rank the features by the absolute value of the Pearson correlation.

### B. Support vector regression

We use the support vector regression (SVR) method [23] implemented in the SVM-light program [24]. SVR is a linear regression method that solves

$$\arg\min_{w,w_0} C\frac{1}{n}\sum_i \max(0, y_i - (w^T x_i + w_0) - \epsilon) + ||w||^2$$

where $x_i$ represents the $i^{th}$ individual and $y_i$ is the phenotype target value. For all experiments we use the default regularization parameter given by $C = \frac{1}{\sum_i x_i^T x_i}$ where $n$ are the number of vectors in the input training (case and control

978-1-5090-0287-0/15 $31.00 © 2015 IEEE
DOI 10.1109/ICMLA.2015.224

individuals in this study) and $x_i$ is the feature vector of the $i^{th}$ individual [24]. In other words we set $C$ to the inverse of the average squared length of feature vectors in the data.

### C. Ridge regression

Linear regression is perhaps the most popular method for solving regression problems. We use its regularized version called ridge regression [22]. This is known to alleviate problems associated with matrix inversion in linear regression and is also less prone to overfitting thanks to regularization. It finds a linear solution to the problem

$$\arg\min_{w,w_0} \frac{1}{n} \sum_i (y_i - (w^T x_i + w_0))^2 + \lambda ||w||^2$$

where $x_i$ represents the $i^{th}$ individual and $y_i$ is the phenotype target value. We use a recently proposed method to automatically set $\lambda$ [25] and the R package Ridge to run this method.

### D. Multivariate feature ranking

We rank features with the above two methods using a simple popular procedure. We first learn a model on the training data and this gives us a $w$ and $w_0$. We then rank features by the absolute value of the entries in $w$.

### E. Experimental procedure

Our experimental procedure begins with a numeric format genome wide association study (GWAS) as input. A GWAS is a matrix of single nucleotide polymorphisms (SNP) where each SNP is given by a string of two letters each taking on the values A, C, G, and T. We convert each SNP into '0', '1', and '2' to represent the number of copies of the allele with the larger alphabet value [8], [26]. In the numeric format the GWAS is given by an $n$ by $m$ matrix of characters taking on the values '0', '1', and '2' where $n$ is the number of subjects and $m$ is the number of SNPs. In Figure 1 we show a simple GWAS of four subjects and three SNPs and its numeric format.

| | | | | |
|---|---|---|---|---|
| A/C | C/T | A/T | | |
| AA | CC | AA | convert to | 0 0 0 |
| AA | CT | AA | numeric format | 0 1 0 |
| AC | TT | AT | ========➔ | 1 2 1 |
| CC | CT | AA | | 2 1 0 |

Fig. 1. Toy example of a genome-wide association study and its numeric encoded format

1) For each GWAS we create ten random splits of training and validation datasets. We do this by randomly selecting 90% of the rows for training and leave the remaining for validation. For the fly dataset, however, we select 80% of the rows instead so that we can compare our results to previously published 5-fold cross-validation.
2) For each training dataset we rank SNPs with the Pearson correlation coefficient, the support vector regression, and ridge regression.
3) For each ranking above we consider top ranked SNPs in increments. For each set of top ranked SNPs we learn a support vector regression and ridge regression model and predict regression values in the validation dataset. After prediction we compute the correlation coefficient between true and predicted values.
4) We repeat the above two steps for each of the 10 training datasets and compute an average. We then plot the average correlation coefficient for different number of top ranked SNPs and combinations of feature ranking and prediction method.

### F. Datasets

We consider three continuous label genome-wide association studies for our study. For each dataset we eliminate all SNPs with missing entries.

- **Mouse: [3]** The mouse GWAS contains 12545 SNPs from 1940 mice across 20 chromosomes and is made publicly available by Welcome Trust Centre for Human Genetics. It can be accessed from http://mus.well.ox.ac.uk/mouse/HS/.
- **Fly: [4]** The fly GWAS contains 2.5 million SNPs from 155 Drosophila Genetic Reference Panel (DGRP)-lines on the Illumina platform.
- **Rice: [27]** The rice GWAS contains 36901 SNPs from 413 rice plants (across 82 countries) across 12 chromosomes and is made publicly available by Rice Diversity Panel. It can be accessed from http://ricediversity.org/data/sets/44kgwas/.

### G. Measure of accuracy

Since we are computing regression values we measure accuracy with the correlation coefficient. This has the same formula as the Pearson correlation coefficient. A value of 1 indicates perfect correlation of predicted and true values, 0 means none, and -1 indicates a negative correlation. A value close to 1 indicates high accuracy.

## III. Results

We begin with our results on the mouse genome wide dataset where we study all six combinations of Pearson and two multivariate rankings against the two prediction methods. Following that we present results on the two remaining datasets focusing on the better of the two multivariate rankings and the univariate one.

### A. Mouse

We consider two phenotypes that were also previously studied in this dataset [3]: the percentage of CD8 cells (CD8) and mean cellular haemoglobin (MCH). In Figure 2 we make several observations. First, for a fixed classifier the multivariate rankings achieve a higher prediction accuracy than the Pearson correlation coefficient with much fewer SNPs. For example in the CD8 phenotype with the top 500 Pearson ranked SNPs the ridge regression method has a correlation coefficient of 0.63. However, with the ridge regression ranking the same classifier reaches about 0.7.

Second, we see that the ridge regression method gives higher accuracies than support vector regression. This may

be due to the fact the ridge regression method of setting the regularizer gives a better regression model than the support vector regression default value. Upto the top 900 ranked SNPs both methods give similar correlation coefficients but with all SNPs in the model the ridge regression attains a much higher value.

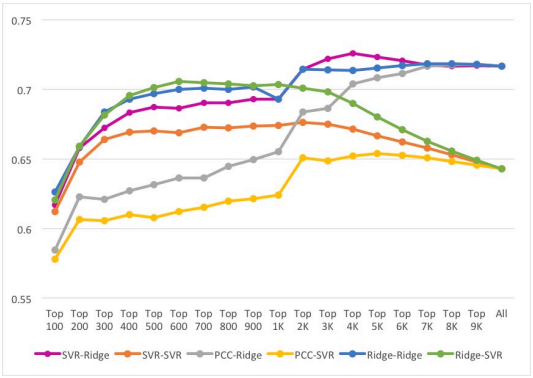


Fig. 2. Mean correlation coefficients of predicting mouse CD8% phenotype shown as a function of top ranked SNPs. Each curve legend contains the SNP ranking method followed by the phenotype prediction method.

In the mean haemoglobin phenotype we see a more pronounced difference between the Pearson ranking and the two multivariate ones (see Figure 3). Both multivariate rankings achieve high (and similar) accuracies very early on in the SNP rankings. Even with all SNPs the difference in accuracy between ridge and support vector regression is about 0.1.

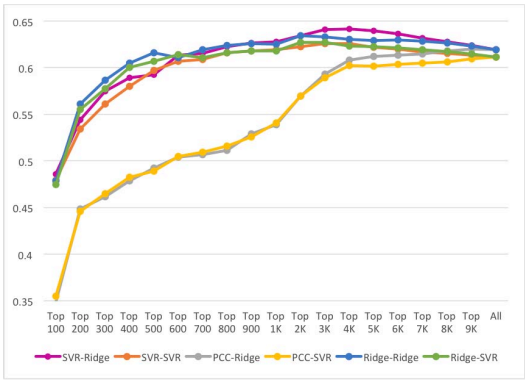


Fig. 3. Mean correlation coefficients of predicting mouse MCH phenotype shown as a function of top ranked SNPs. See Figure 2 caption for more details.

Compared to previously published values under 90:10 cross validation studies (published in [3]) our correlation coefficients are similar for CD8 and higher for MCH. For CD8 the original study of the paper reported an accuracy of 0.73. We reach the same accuracy with the ridge regression applied to the support vector regression ranking of SNPs. However, our peak is with 4000 SNPs whereas the original study does not list the number of SNPs. (We assume all SNPs were used for their model.) The original paper also reports an accuracy of 0.61 for MCH whereas we reach a peak accuracy of 0.64 with the ridge regression applied to top 3000 support vector regression ranked SNPs (the same combination that was optimal for CD8).

Of the two multivariate rankings we see that support vector regression works better for both phenotypes. Therefore we omit the ridge regression ranking going forward.

### B. Fly

The fly dataset contains 2.5 million SNPs and this presents a challenge for multivariate classifiers. For such high dimension multivariate classifiers give a poor ranking and prediction with genome wide SNP data [8]. Therefore we consider just the top 200,000 SNPs in the Pearson ranking and re-rank them with the support vector regression. In Figure 4 we see that upto the top 3000 SNPs the multivariate ranking gives higher accuracies. The ridge regression, however, yields a higher accuracy than the support vector regression.

In the original study of this dataset the authors use additional non-genomic information to predict phenotype. With just SNP data we reach an accuracy of 0.33 whereas the original study report 0.23.

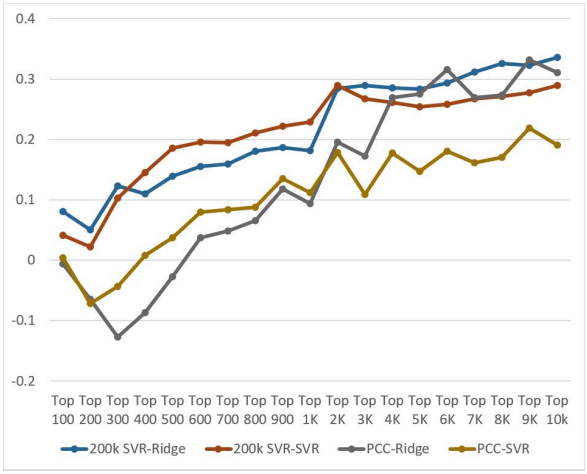


Fig. 4. Mean correlation coefficients of predicting fly startle response shown as a function of top ranked SNPs. See Figure 2 caption for more details.

### C. Rice

We study several phenotypes in the rice dataset [27]. We see similar trends that we observed in the above datasets (see Figure 5). Within the top 500 SNPs the support vector regression ranking gives higher prediction accuracies than the Pearson one. For phenotype prediction the ridge regression gives higher accuracies than support vector regression. The blue curve that denotes the ridge regression applied to the support vector regression ranking of SNPs usually attains its highest accuracy before it crosses 10,000 SNPs. The number of SNPs where the peak prediction accuracy is reached and the peak prediction accuracy differ from one phenotype to the next.

In the original study of this dataset the authors do not study phenotype prediction. The original study aims to discover and study (significant) genetic variants in the data.

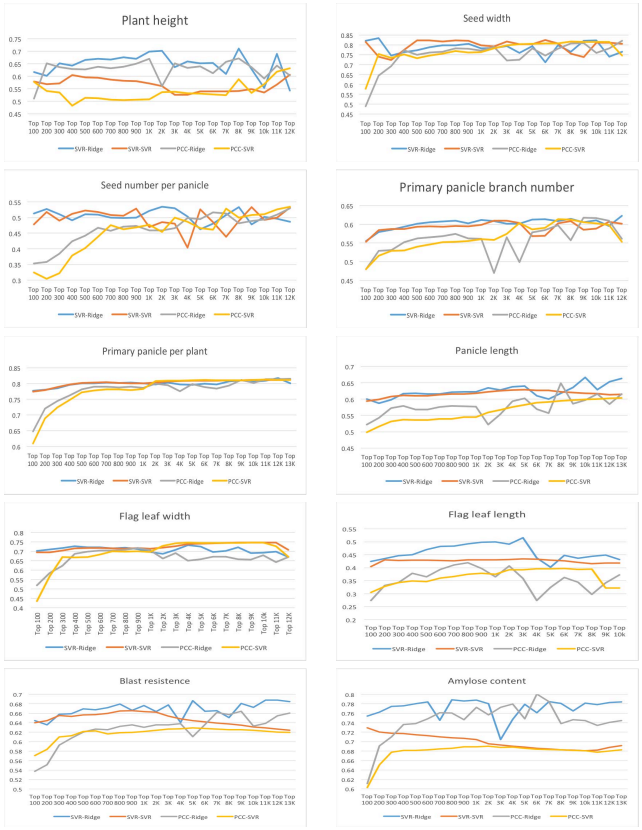


Fig. 5. Mean correlation coefficients of predicting various rice phenotypes shown as a function of top ranked SNPs. See Figure 2 caption for more details.

## IV. Discussion

In previous work we have studied SNP selection with univariate and multivariate methods in large real and simulated case control genome wide association studies with at least 30,000 SNPs [8]. There we found that applying a multivariate classifier to all SNPs yielded a poor ranking compared to the univariate chi-square test. We found it was better (for case control prediction) to first rank SNPs with a univariate method and then select the top few to re-rank with a multivariate method like the support vector machine.

In this study we don't necessarily make this observation. For example in the mouse data we first started with this method. We obtained the top 1000 and top 10,000 Pearson ranked SNPs and re-ranked them with the support vector regression and ridge regression methods. However, when we ranked all SNPs with support vector regression and ridge regression we obtained a high prediction correlation coefficient than compared to re-ranking the top 1000 and top 10,000 Pearson ranked ones.

We conjecture this may be happening because in this study we are dealing with regression data as opposed to classification where we would have discrete labels. In fact while feature selection for classification data is widely studied [17], [18], [19], [20], [21], [15] there are hardly any studies that look at this problem for regression data. Thus our study provides some insight into feature selection in regression data.

The mouse dataset provides several other phenotypes that we do not show in this study. In those phenotypes we make similar observations as for the two shown in the paper: the support vector regression ranking followed by ridge regression gives the highest accuracy. We study CD8 and MCH since these were chosen in the original study and we can compare our accuracies to theirs [3].

One avenue of future work would be to study kernel and L1 norm regression methods [22]. The former would determine non-linear regression solutions. The latter, however, is non-convex and likely to have longer runtime solutions than the methods we have used. Another avenue of future work is the selection of the optimal penalty coefficient in the support vector regression both for feature selection and phenotype prediction.

## V. Conclusion

Our results show that we can learn parsimonious SNP models for predicting continuous phenotypes using the support vector regression for ranking SNPs and ridge regression for prediction. With this combination we reach comparable or higher prediction values than previously reported and we do so with fewer SNPs than previously used. Thus our method may be useful for obtaining SNP models to perform accurate phenotype prediction in genome wide studies.

## Acknowledgment

We thank Advance Research and Computing Services at NJIT for providing and supporting computing services necessary to complete this research.

## References

[1] H. Zhang, Z. Wang, S. Wang, and H. Li, "Progress of genome wide association study in domestic animals," *J Anim Sci Biotechnol*, vol. 3, no. 1, p. 26, 2012.

[2] C. Riedelsheimer, F. Technow, and A. E. Melchinger, "Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines," *BMC genomics*, vol. 13, no. 1, p. 452, 2012.

[3] S. H. Lee, J. H. van der Werf, B. J. Hayes, M. E. Goddard, and P. M. Visscher, "Predicting unobserved phenotypes for complex traits from whole-genome snp data," *PLoS Genet*, vol. 4, no. 10, p. e1000231, 2008.

[4] U. Ober, J. F. Ayroles, E. A. Stone, S. Richards, D. Zhu, R. A. Gibbs, C. Stricker, D. Gianola, M. Schlather, T. F. Mackay *et al.*, "Using whole-genome sequence data to predict quantitative trait phenotypes in drosophila melanogaster," *PLoS genetics*, vol. 8, no. 5, p. e1002685, 2012.

[5] G. Abraham, A. Kowalczyk, J. Zobel, and M. Inouye, "Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease," *Genetic Epidemiology*, vol. 37, no. 2, pp. 184–195, 2013. [Online]. Available: http://dx.doi.org/10.1002/gepi.21698

[6] N. Chatterjee, B. Wheeler, J. Sampson, P. Hartge, S. J. Chanock, and J.-H. Park, "Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies," *Nature Genetics*, vol. 45, p. 400405, 2013.

[7] J. Kruppa, A. Ziegler, and I. Knig, "Risk estimation and risk prediction using machine-learning methods," *Human Genetics*, vol. 131, no. 10, pp. 1639–1654, 2012. [Online]. Available: http://dx.doi.org/10.1007/s00439-012-1194-y

[8] U. Roshan, S. Chikkagoudar, Z. Wei, K. Wang, and H. Hakonarson, "Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest," *Nucleic Acids Research*, vol. 39, no. 9, p. e62, 2011.

[9] M. Sandhu, A. Wood, and E. Young, "Genomic risk prediction," *The Lancet*, vol. 376, pp. 1366–1367, 2010.

[10] C. Kooperberg, M. LeBlanc, and V. Obenchain, "Risk prediction using genome-wide association studies," *Genetic Epidemiology*, vol. 34, no. 7, pp. 643–652, 2010.

[11] D. M. Evans, P. M. Visscher, and N. R. Wray, "Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk," *Human Molecular Genetics*, vol. 18, no. 18, pp. 3525–3531, 2009.

[12] A. C. J. Janssens and C. M. van Duijn, "Genome-based prediction of common diseases: advances and prospects," *Human Molecular Genetics*, vol. 17, no. R2, pp. R166–R173, 2008.

[13] N. R. Wray, M. E. Goddard, and P. M. Visscher, "Prediction of individual genetic risk of complex disease," *Current Opinion in Genetics and Development*, vol. 18, pp. 257–263, 2008.

[14] ——, "Prediction of individual genetic risk to disease from genome-wide association studies," *Genome Research*, vol. 17, pp. 1520–1528, 2007.

[15] I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror, "Result analysis of the nips 2003 feature selection challenge," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. Cambridge, MA: MIT Press, 2005, pp. 545–552.

[16] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[17] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *Journal Of The Royal Statistical Society Series B*, vol. 70, no. 5, pp. 849–911, 2008.

[18] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[19] A. Statnikov, D. Hardin, and C. Aliferis, "Using svm weight-based methods to identify causally relevant and non-causally relevant variables," in *Proceedings of Neural Information Processing Systems (NIPS) Workshop on Causality and Feature Selection*, 2006.

[20] Y.-W. Chen and C.-J. Lin, "Combining svms with various feature selection strategies," in *Feature Extraction*, ser. Studies in Fuzziness and Soft Computing, I. Guyon, M. Nikravesh, S. Gunn, and L. Zadeh, Eds. Springer Berlin / Heidelberg, 2006, vol. 207, pp. 315–324.

[21] D. Hardin, I. Tsamardinos, and C. F. Aliferis, "A theoretical characterization of linear svm-based feature selection," in *ICML '04: Proceedings of the twenty-first international conference on Machine learning*. New York, NY, USA: ACM, 2004, p. 48.

[22] E. Alpaydin, *Machine Learning*. MIT Press, 2004.

[23] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[24] T. Joachims, "Making large-scale svm learning practical," in *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. MIT Press, 1999.

[25] E. Cule and M. De Iorio, "Ridge regression in prediction problems: Automatic choice of the ridge parameter," *Genetic Epidemiology*, vol. 37, no. 7, pp. 704–714, 2013. [Online]. Available: http://dx.doi.org/10.1002/gepi.21750

[26] P. Paschou, E. Ziv, E. G. Burchard, S. Choudhry, W. Rodriguez-Cintron, M. W. Mahoney, and P. Drineas, "Pca-correlated snps for structure identification in worldwide human populations," *PLoS Genet*, vol. 3, no. 9, p. e160, 09 2007. [Online]. Available: http://dx.plos.org/10.1371%2Fjournal.pgen.0030160

[27] K. Zhao, C.-W. Tung, G. C. Eizenga, M. H. Wright, M. L. Ali, A. H. Price, G. J. Norton, M. R. Islam, A. Reynolds, J. Mezey *et al.*, "Genome-wide association mapping reveals a rich genetic architecture of complex traits in oryza sativa," *Nature communications*, vol. 2, p. 467, 2011.