Brown University

Master's Project Report

Linear Methods for SNP Selection

Author:
Alexander Stuart Gillmor

Advisor:
Sorin Istrail

A report submitted in partial fulfilment of the requirements
for the degree of Master of Science

in the

Istrail Lab
Department of Computer Science

May 2012

Acknowledgements

I'd like to thank and acknowledge Sorin Istrail for his guidance, knowledge, and interest; Austin Huang for his leadership and support; Derek Aguiar and Ryan Tarpine for their thoughts and input.

Introduction

A common study design for genetic mapping for a trait or disease is genome-wide association studies (GWAS) that seeks to identify a number of individuals carrying a trait or disease and compare those individuals with individuals whom are not known to carry that disease or trait. These GWAS genotype a large number of single nucleotide polymorphisms (SNP) from across the entirety of the genome and study the statistical association between those SNPs and the trait or disease.

Due to the vast quantity of SNPs, up to 10 million in the genome and often 500 thousand to 1 million in GWAS consisting of tens of thousand individuals, single association tests must undergo power corrections to adjust for many possible hypothesis so only the strongest of signals come through. However, though GWAS have been successful in identifying hundreds of genetic variants associated with complex human disease and traits most of these variants explain at best a small amount of risk and this leads to the question of how the remaining missing heritability can be explained.

The concept of epistasis or gene-gene interaction has long been offered as an explanation for deviations from Mendelian ratios. The nature of these disease suggest that such interactions are ubiquitous. However traditional parametric statistical methods such as linear and logistic regression that consider the interactions among multiple polymorphisms quickly devolve into extremely complex problems as they encounter the so called curse of dimensionality. The aim of this project was to investigate the strength of penalized regression techniques that may offer a relatively fast technique at identifying some of these interactions.

Linear Regression Models

Ordinary Least Squares

$$\hat{\beta} = \arg\min_{\beta} \textstyle\sum_{i=1}^{N}(y_i - \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

The most common linear regression model is to be given p predictors X and predict a response Y given a linear combinations of coefficients β. The ordinary least squares (OLS) estimates are obtained by a closed form solution that minimizes the residual sum of squares at the cost of inverting the matrix. Interpretation of the coefficients β from an OLS solution are difficult to interpret due to the p number of non-zero terms. It is important to note that OLS produces coefficient estimates that have the smallest variance among unbiased estimates and therefore in many cases does not do very well in prediction. A variety of penalization techniques have been proposed to introduce bias and in some cases improve upon OLS prediction estimates. The two goals of such penalizations are to increase prediction accuracy and offer a better chance of interpretation by both removing and shrinking coefficients. In regards to interpretation, finding a smaller subset that explains a big picture can reduce the complexity of large problems and in regards to prediction, some of the bias is lost to increase the variance of the predicted values and that may improve the overall prediction. Key methods to be discussed in this paper are the following:

Ridge Regression

$$\hat{\beta} = \arg\min_{\beta} \textstyle\sum_{i=1}^{N}(y_i - \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j)^2 \text{ s.t } \sum_{j=1}^{p} \beta_j^2 \leq \lambda$$

Ridge regression was an early technique [1] that built upon OLS by enforcing a L2 quadratic penalty on the regression coefficients. Applying the L2 penalty leads to many small coefficients that are selected continuously and has the benefit of cancellings out the effect that colinear coefficients can exhibit with each other and producing many small coefficients.

LASSO Regression

$$\hat{\beta} = \arg\min_{\beta} \textstyle\sum_{i=1}^{N}(y_i - \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j)^2 \text{ s.t } \sum_{j=1}^{p} |\beta_j| \leq \lambda$$

The LASSO technique was proposed by Tibshirani [2]. Lasso regression enforces an L1 absolute penalty on the regression coefficient. Due the nature of the L1 penalty, the LASSO does variable selection automatically by penalizing many of the coefficients to zero. The selection of the coefficients is continuous and the number of coefficients that enter the model are controlled by a parameter that dictates the absolute weight of the coefficients, and this parameter is set outside during model selection. It is important to note that there is no closed form solution for LASSO coefficient solutions and the best techniques for generating LASSO solutions are generally approximations and can be very computationally expensive.

Naive Elastic Net Regression

$$\hat{\beta} = \arg\min_{\beta} \textstyle\sum_{i=1}^{N}(y_i - \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

$$\text{s.t } (1-\alpha)\textstyle\sum_{j=1}^{p}|\beta_j| + \alpha\sum_{j=1}^{p}\beta_j^2 \leq \lambda \text{ for } \alpha \in [0,1]$$

Naive Elastic Net[3] regression is a combination of the L1 and L2 penalties in the objective function. The L1 penalty performs variable selection and the L2 encourages a grouping effect. Conceptually it can viewed as a combination of LASSO and Ridge regression. The objective function can impose a shrinkage that will introduce additional bias while not reducing variance.

Least Angle Regression

Least Angle Regression (LAR)[4] is a regression algorithm that is based on L1 norm optimization suitable for high-dimensional data containing many irrelevant and many correlated variables. All the coefficients are initialized at 0 and estimated coefficient parameters are estimated one by one and increased in a direction equiangular to the given parameters correlation with the residual. LAR has the benefit of solving the system of equations for the number of predictors that has entered the model at any given point while maintaining that path. The obvious advantage here is that the systems of equations to be solved is only as large as coefficients that have entered the solution, so LAR can produce models to explain a large amount of data in a relatively small set of time. LASSO solutions can be obtained via a small modification of the LAR algorithm.

Tuning

Each model has one or more parameters that have to be tuned for fitting and prediction represented in the given examples as λ. There are a variety of criterion that can be used to tune the model. The most straightforward is cross-validation in which a partition is made of the data is made into many subsets and a parameter is chosen based on how well the prediction of the training data is made on the subset. The tuning parameters are chosen to minimize an error such as the root mean squared error (RMSE). Other measures work to minimize another measure of relative goodness of fit such as the Akaike information criterion as well as the Bayesian information criterion. All of these parameters seek to either regularize, in the context of controlling the amount attributed to each coefficient, or to reduce the number of terms that enter each model.

Software

Experiments and data exploration was done with the 0.9 implementation of SciKit-learn [5]. SciKit learn exposes many machine learning algorithms and built upon the scientific computing library of SciPy and the numerical library of NumPy[6]. SciKit learn has many different linear models with a selection of criterion material to experiment with including many discussed above.

Generated Data

To test the ability of regression methods to recover information such as highly correlated and "causal" SNPs a variety of simple models were generated that simulated two way and three way epistatic interactions. If the associated SNPs are labeled SNP1, SNP2, and SNP3 the alleles are represented by (A,a), (B,b) and (C,c), respectively. The models are described as follows:

Two Way Generated models

$$
\begin{aligned}
Model_1 &= -3.198 + 1.23\ [\text{SNP1} = \text{AA}] - 0.5\ [\text{SNP2} = \text{BB}] + 2.12\ [\text{SNP1} = \text{AA}\ \&\ \text{SNP2} = \text{Bb}] \\
Model_2 &= -8.9 + 7.3\ [\text{SNP1} = \text{Aa}\ \&\ \text{SNP2} = \text{BB}] \\
&\qquad + 7.3\ [\text{SNP1} = \text{AA}\ \&\ \text{SNP2} = \text{Bb}] + 7.3\ [\text{SNP1} = \text{AA}\ \&\ \text{SNP2} = \text{BB}] \\
Model_3 &= -6.9 + 3.5\ [\text{SNP1} = \text{AA}] + 3.4\ [\text{SNP2} = \text{BB}] \\
Model_4 &= -2.197 + 0.88\ [\text{SNP1} = \text{AA}] + 0.88\ [\text{SNP2} = \text{BB}]
\end{aligned}
$$

Three Way Generated models

$$
\begin{aligned}
Model_1 &= -9.1 + 3\ [\text{SNP1} = \text{AA}] + 3\ [\text{SNP2} = \text{BB}] + 3\ [\text{SNP3} = \text{CC}] \\
Model_2 &= -15 + 4\ [\text{SNP1} = \text{AA}] + 4\ [\text{SNP3} = \text{CC}] - 8\ [\text{SNP1} = \text{AA}\ \&\ \text{SNP2} = \text{BB}] \\
&\qquad + 5\ [\text{SNP1} = \text{Aa}\ \&\ \text{SNP2} = \text{BB}\ \&\ \text{SNP3} = \text{CC}] \\
&\qquad + 6\ [\text{SNP1} = \text{AA}\ \&\ \text{SNP2} = \text{BB}\ \&\ \text{SNP3} = \text{CC}] \\
Model_3 &= -10 + 3\ [\text{SNP1} = \text{AA}\ \&\ \text{SNP2} = \text{BB}] + \\
&\qquad 3\ [\text{SNP1} = \text{AA}\ \&\ \text{SNP3} = \text{CC}] + 3\ [\text{SNP2} = \text{BB}\ \&\ \text{SNP3} = \text{CC}] \\
Model_4 &= -5.8 + 3\ [\text{SNP2} = \text{Bb}] + 3\ [\text{SNP1} = \text{AA}]\ \&\ \text{SNP2} = \text{BB}\ \&\ \text{SNP3} = \text{CC}
\end{aligned}
$$

The models were generated arbitrarily to simulate a range of human disease along the methods of similar studies with generated data [7]. For our purposes it is safe to say that the chosen SNPs are causal since we know them a priori, in non-generated data the best would be due be to call recovered SNPs correlated.

A seed population of 2,000 individuals from the HapMap[8] r.24 phased data Central European (CEU) population containing SNP from the first 150,000 base pairs was generated using hapgen2 [9], a region that contained 120 SNPs. The 6th chromosome was chosen because of its polymorphic nature and involvement with many autoimmune disease. For simplicity data was encoded with homozygous as 0 and 2, and the heterozygous allele was encoded with 1.

From this population Binomial trials were run using the given model as $p = e^{Model_x}$ until 400 cases were selected and then 400 controls were then be chosen from the seed population not affected by disease.

Three SNPs positions were chosen from the 120 SNPs generated with relative frequencies as follows:

SNP 75 : 0.4115 and .1350 for the major and minor homozygous alleles. .4535 for the heterozygous allele
SNP 22 : 0.6245 and .0405 for the major and minor homozygous alleles. .3350 for the heterozygous allele
SNP 107 : 0.789 and 0.0135 for the major and minor homozygous alleles. 0.1975 for the heterozygous allele.

SNPs 75 and 22 were used as SNP1 and SNP2 in the given two way models and SNPs 75, 22 and 107 were used as SNP1, SNP2 and SNP3 in the given three way models.

Generated Experimental Method

Some similar regression studies [10] have set a fixed number of SNPs to select via LARS and then used the majority of SNPs selected from an extensive series of simulations to produce features to build prediction models from. However selection of SNPs like this is not robust as it may fully capture colinearity that exists among features and does not account for the fact that features that enter the model can also leave. The generated dataset was much smaller than previous studies and since the data was generated the SNPs that we call causal were known a priori. For this purpose we ran 100 random permutations of the data in each model through criterion selection to tune the model

Two Way Results							
$Model_1$		$Model_2$		$Model_3$		$Model_4$	
SNP 75	0.74	**SNP 75**	0.74	**SNP 75**	0.83	**SNP 75**	0.75
SNP 22	0.61	**SNP 22**	0.66	**SNP 22**	0.79	SNP 15	0.68
SNP 76	0.37	SNP 101	0.32	SNP 2	0.45	**SNP 22**	0.61
SNP 2	0.37	SNP 39	0.31	SNP 85	0.44	SNP 87	0.54
SNP 20	0.29	SNP 93	0.13	SNP 68	0.39	SNP 101	0.49
SNP 63	0.25	SNP 85	0.11	SNP 111	0.38	SNP 107	0.42

Table 1: Top frequencies of SNP selection for 100 model runs

Three Way Results							
$Model_1$		$Model_2$		$Model_3$		$Model_4$	
SNP 75	0.66	**SNP 22**	0.66	**SNP 75**	0.72	**SNP 107**	0.54
SNP 22	0.65	**SNP 107**	0.55	**SNP 22**	0.69	**SNP 75**	0.51
SNP 107	0.50	**SNP 75**	0.52	SNP 15	0.51	**SNP 22**	0.46
SNP 2	0.28	SNP 108	0.17	**SNP 107**	0.49	SNP 53	0.27
SNP 85	0.24	SNP 100	0.01	SNP 66	0.33	SNP 46	0.27
SNP 5	0.24	SNP 85	0.01	SNP 2	0.30	SNP 82	0.25

Table 2: Top frequencies of SNP selection for 100 model runs

for feature selection. The coefficients selected from each random initialization were then compared across all the runs.

Each model was run through various LAR models with a different model selection criterion. The Lasso LAR algorithm was used in these experiments. The AIC and BIC were not effective in effectively reducing the set of SNPs in our simulated data. 5-fold cross validation minimizing the RMSE was relatively successful at picking smaller models with decent parameters.

Generated Experimental Results

The results for all the runs are given but the trend is that the generated causal SNPs were selected by a much higher majority than most other given SNPs. This hints at LAR as being a meaningful technique to recover important information for correlated loci for complex human diseases.

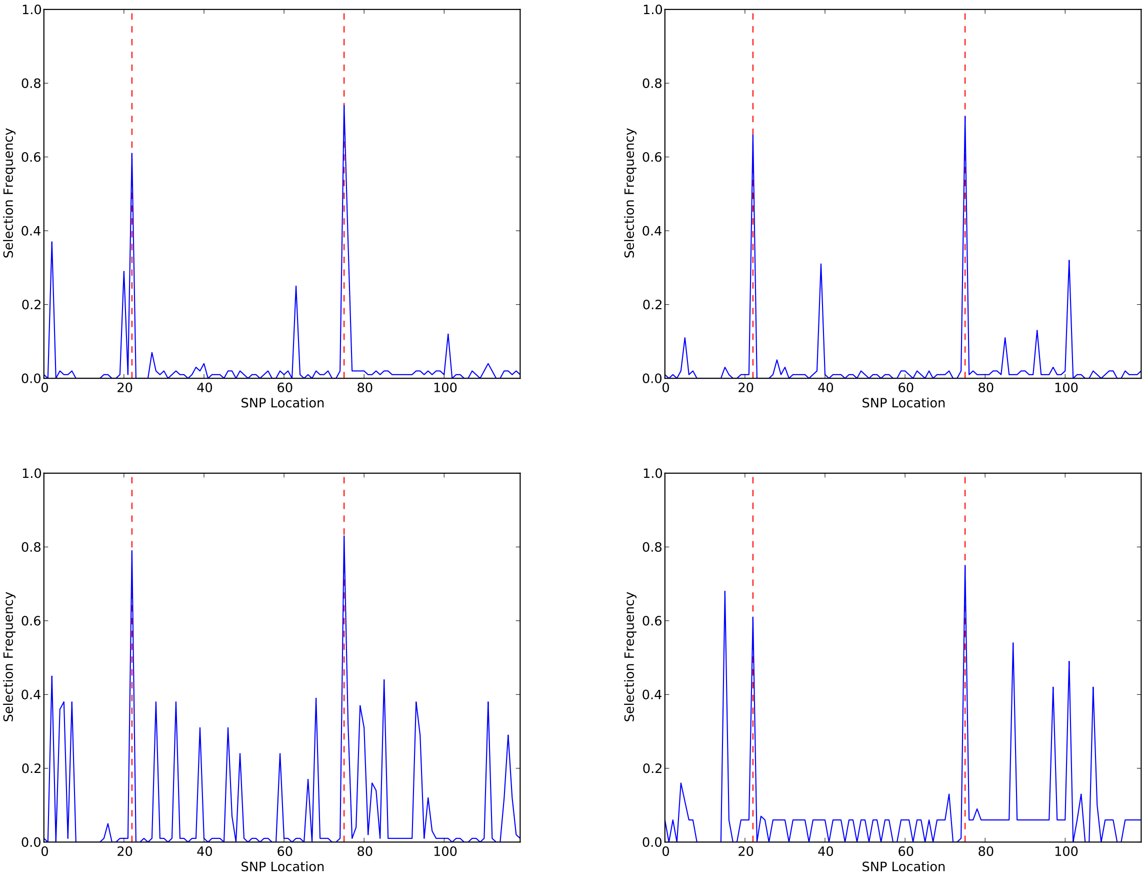

Figure 1: Two Way Model Frequency Graphs for Generated Epistatic Data. Frequency in blue. 'Causal' SNP in red.

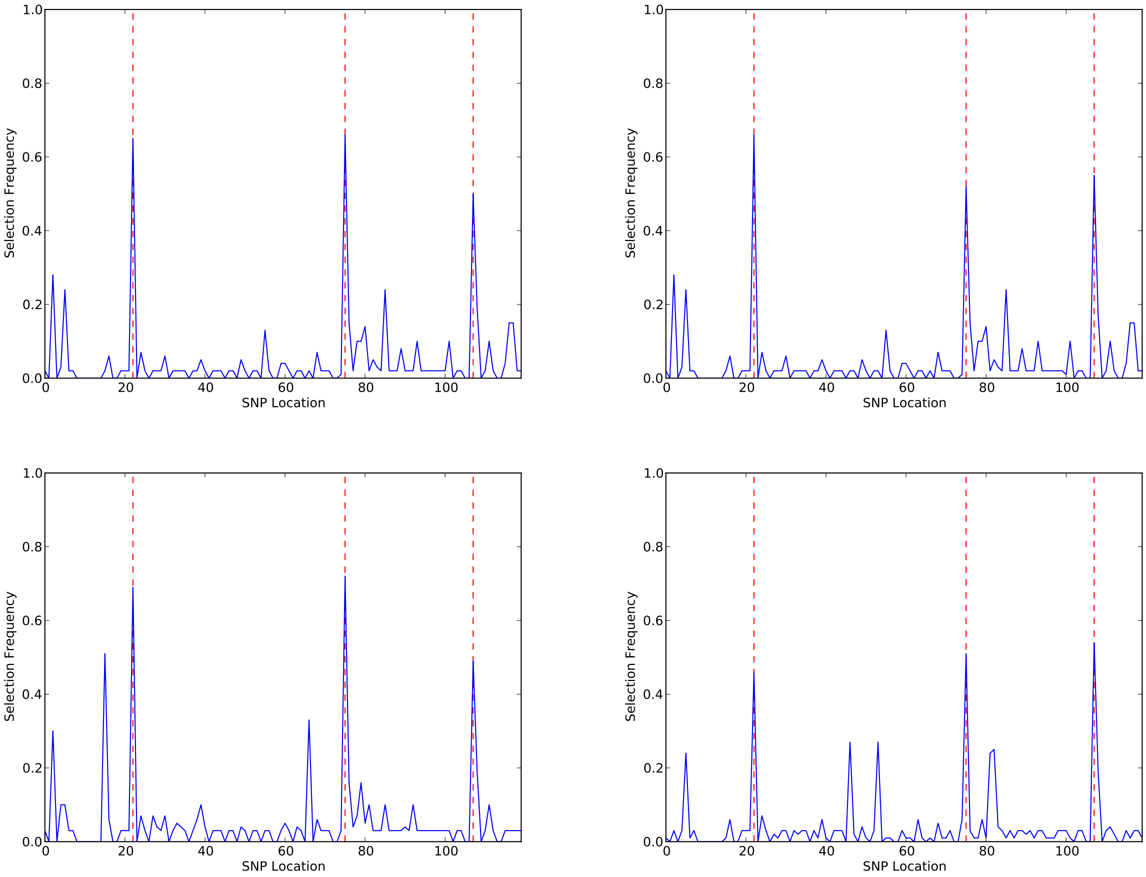

Figure 2: Three Way Model Frequency Graphs for Generated Epistatic Data. Frequency in blue. 'Causal' SNP in red.

Data Exploration

Two real world data sources were made available for this research. The International Multiple Sclerosis Genetics Consortium data and multiple sclerosis data from the Wellcome Trust Case Control Consortium were available to run linear techniques on. In both of these datasets the data was so large that running large scale runs containing multiple permutations was infeasible on a desktop machine so a less computationally intensive approach was used.

International Multiple Sclerosis Genetics Consortium

The data of the International Multiple Sclerosis Genetics Consortium (IMSGC) [11] is 931 trios consisting of 2 unaffected parents and 1 affected child. The data comprised 334,723 SNPs for the 2793 individuals and was an extensive GWAS study looking for markers associated with multiple sclerosis. Associated SNPs were found via transmission disequillibrium testing of the family trios and Cochran-Mantel-Haenszel testing of case controls with significant p-values.

For our purposes the data was encoded as described previously, with homozygous alleles encoded as 0 and 2 and the heterozygous allele encoded as a 1. The extent of the data was too large to load into memory in the test machine (with 8GB of RAM). In the published GWAS study from this data the 6th chromosome and, in particular, the major histocompatability complex (MHC) was a region of where many loci in high association with disease were located, so Lasso LAR was run on both the MHC and then the 6th chromosome as a whole. In both instances Lasso LAR was trained on the entire data set; the MHC with 5 fold cross validation and the 6th chromosome with 7-fold cross validation. The features selected from this data are listed in Appendix A1. Lasso LAR on the MHC was not able to recover any of the SNPs in high association from the study but did identify one particular SNP, rs9270986, that was shown to have a highly significant residual association when the data was controlled for HLA-DR locus; the highest association found in their study. Lasso LAR on the 6th chromosome did not replicate this or the study's results.

Wellcome Trust Case Control Consortium

The data of the Wellcome Trust Case Control Consortium [12] comprised of 14,790 SNPs in 1,476 control individuals and 994 case individuals. In these cases the tuning parameter was selected via Cross Validation, AIC, and BIC on 2,070 case control individuals with the remaining 400 individuals (200 case, 200 control) held out as test data with Lasso

LAR and Naive Elastic Net Regression. SNPs with an unknown frequency greater than 5% were removed from the dataset and then unknown SNPs were assigned to the heterozygous encoding. The processing occured on a chromosome by chromosome basis. Generally speaking the resulting models often are only accurate at predicting true negatives. Still, for each chromosome we see that some models preform better than others.

Conclusion

We have shown through generated data that LAR can be effective in recovering epistatic SNP interactions. LAR and other linear methods are interesting techniques for possibly capturing some of the information in epistatic interaction. Running these regressions on commonly available consumer hardware implies the strength of their computational power. Future work should investigate running such regressions in grid, cluster, or cloud computing to fully take advantage of their power. The LAR path has been generalized such that the LASSO and Elastic Net have built upon it and there may be other definitions that would help to maintain its power while producing more meaningful epistatic features.

Since in all of these cases the techniques were run on whole GWAS matrices they can potentially handle as many interaction terms as there are in the matrix. It is important to note for the generated data the interaction terms are not explicitly defined and yet recovered. Further computational models of epistasis should be formalized and investigated with a these linear techniques in concert with standard corrected p-value association tests.

Acknowledgements

Thanks to the International Multiple Sclerosis Genetics Consortium and Wellcome Trust Case Control Consortium for making their data available.

MHC Results		
Chromosome	Location	SNPid
6	29782176	rs3116788
6	29812379	rs1736913
6	29841721	rs1737055
6	31528479	rs3131622
6	32308125	rs3134926
6	32419357	rs3129900
6	32682038	**rs9270986**
6	32712350	rs9272346
6	32766288	rs9469220

Table 3: 5-fold CV Lasso LAR results

Chromosome 6 Results		
Chromosome	Location	SNPid
6	2620192	rs12205879
6	38792163	rs4714188
6	53703465	rs4639324
6	77988470	rs1342638
6	96853615	rs4839826
6	117489335	rs10484309
6	137774624	rs9484046
6	139452276	rs6570309
6	151706569	rs3900024
6	153333549	rs503366

Table 4: 7-fold CV Lasso LAR results

Chromosome		Lasso LAR			EN
		7f C.V.	AIC	BIC	5f C.V.
1	Precision	0.956522	0.829787	0.950000	0.648649
	Accuracy	0.552500	0.577500	0.545000	0.555000
2	Precision	0.532110	0.641026	0.888889	0.527778
	Accuracy	0.517500	0.527500	0.517500	0.505000
3	Precision	DIV0	1.000000	1.000000	DIV0
	Accuracy	0.500000	0.540000	0.540000	0.500000
4	Precision	DIV0	0.500000	1.000000	0.500000
	Accuracy	0.500000	0.500000	0.619324	0.500000
5	Precision	0.591667	0.894737	0.894737	0.692308
	Accuracy	0.555000	0.537500	0.537500	0.537500
6	Precision	0.777778	0.785047	0.781250	0.736000
	Accuracy	0.650000	0.652500	0.590000	0.647500
7	Precision	DIV0	1.000000	1.000000	0.555556
	Accuracy	0.500000	0.530000	0.530000	0.502500
8	Precision	0.800000	0.625000	DIV0	DIV0
	Accuracy	0.507500	0.505000	0.500000	0.500000
9	Precision	DIV0	0.791667	0.782609	0.700000
	Accuracy	0.500000	0.535000	0.532500	0.540000
10	Precision	DIV0	0.000000	DIV0	DIV0
	Accuracy	0.500000	0.497500	0.500000	0.500000
11	Precision	DIV0	1.000000	1.000000	DIV0
	Accuracy	0.500000	0.520000	0.520000	0.500000
12	Precision	DIV0	0.863636	DIV0	DIV0
	Accuracy	0.500000	0.540000	0.500000	0.500000
13	Precision	1.000000	1.000000	1.000000	DIV0
	Accuracy	0.520000	0.522500	0.520000	0.500000
14	Precision	1.000000	0.933333	1.000000	DIV0
	Accuracy	0.527500	0.532500	0.527500	0.500000
15	Precision	0.685714	0.588235	1.000000	0.533333
	Accuracy	0.532500	0.530000	0.507500	0.510000
16	Precision	1.000000	1.000000	1.000000	0.723404
	Accuracy	0.542500	0.542500	0.540000	0.552500
17	Precision	DIV0	0.586207	1.000000	0.833333
	Accuracy	0.500000	0.525000	0.525000	0.510000
18	Precision	0.888889	1.000000	DIV0	DIV0
	Accuracy	0.517500	0.517500	0.500000	0.500000
19	Precision	0.896552	0.780488	0.916667	0.652174
	Accuracy	0.557500	0.557500	0.550000	0.552500
20	Precision	0.670588	1.000000	1.000000	DIV0
	Accuracy	0.572500	0.542500	0.537500	0.500000
21	Precision	DIV0	DIV0	DIV0	DIV0
	Accuracy	0.500000	0.500000	0.500000	0.500000
22	Precision	0.490446	0.785714	0.785714	DIV0
	Accuracy	0.492500	0.520000	0.520000	0.500000

TABLE 5: Chromosome by Chromosome Accuracy and Precision

Chromosome	Size in SNPs	Lasso LAR			EN
		7f C.V.	AIC	BIC	5f C.V.
1	1320	10.658926	2.185073	2.061518	736.142465
2	765	6.346181	1.430657	1.065680	342.702147
3	740	6.048923	1.274461	1.041007	360.659427
4	564	4.308561	1.093926	0.898759	174.509831
5	665	6.146835	1.246230	1.054225	293.963403
6	1666	15.603409	3.496535	3.339984	1857.113945
7	498	4.318206	1.057292	0.946202	115.846669
8	361	3.849544	1.022752	0.933357	92.704458
9	499	4.328411	1.075688	0.953451	182.459633
10	556	4.689394	1.245128	1.015039	195.345757
11	795	6.719915	1.689862	1.467723	399.033024
12	582	4.626239	1.285364	0.993746	228.760478
13	291	2.912949	0.898435	0.869864	72.166622
14	465	4.222388	1.035889	0.909310	177.195516
15	456	4.077359	0.979585	0.888176	147.397505
16	505	4.439754	1.110380	0.815451	172.783174
17	608	5.085032	1.204284	1.096623	238.189752
18	232	1.772682	0.271745	0.229249	46.339692
19	889	7.154027	1.654671	1.619452	490.897624
20	391	3.721791	1.206640	1.142858	111.530976
21	148	0.777694	0.107960	0.122645	13.008152
22	329	4.216304	0.949633	0.850607	69.904153

Table 6: Chromosome by Chromosome Runtime in Seconds

Times generated on a quad-core 3.4 GHz AMD Phenom II X4 Processor with 8GB of DDR2 1066Mhz RAM running Ubuntu 10.10. The Laso LARS with cross validation made use of all four cores via multiprocessing. All other tests were with a single processor.

Bibliography

[1] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

[2] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288, 1996.

[3] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

[4] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Duchesnay E. Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[6] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. URL `http://www.scipy.org/`.

[7] Hua He, William Oetting, Marcia Brott, and Saonli Basu. Power of multifactor dimensionality reduction and penalized logistic regression for detecting gene-gene interaction in a case-control study. *BMC Medical Genetics*, 10(1):127, 2009. ISSN 1471-2350. doi: 10.1186/1471-2350-10-127. URL `http://www.biomedcentral.com/1471-2350/10/127`.

[8] The HapMap Consortium. The international hapmap project. 426:789–796, 2003. URL `http://www.nature.com/nature/journal/v426/n6968/abs/nature02168.html`.

[9] Zhan Su, Jonathan Marchini, and Peter Donnelly. Hapgen2: simulation of multiple disease snps. *Bioinformatics*, 27(16):2304–2305, 2011. URL `http://dblp.uni-trier.de/db/journals/bioinformatics/bioinformatics27.html#SuMD11`.

[10] Dairong Wang, Henry P. Parkman, Michael R. Jacobs, Anurag K. Mishra, Evgeny Krynetskiy, and Zoran Obradovic. DNA microarray SNP associations with clinical efficacy and side effects of domperidone treatment for gastroparesis. *Journal of Biomedical Informatics*, 2011. ISSN 15320464. doi: 10.1016/j.jbi.2011.11.013. URL http://dx.doi.org/10.1016/j.jbi.2011.11.013.

[11] The International Multiple Sclerosis Genetics Consortium. Risk alleles for multiple sclerosis identified by a genomewide study. *The New England Journal of Medicine*, page 073493, July 2007. URL http://content.nejm.org/cgi/content/full/NEJMoa073493?query=TOC.

[12] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447 (7145):661–678, June 2007. URL http://dx.doi.org/10.1038/nature05911.