Interpretive Risk Assessment on GWA Data with Sparse Linear Regression

Ingrid Brænne, Kai Labusch, Thomas Martinetz and Amir Madany Mamlouk

Institute for Neuro and Bioinformatics, University of Luebeck
braenne@inb.uni-luebeck.de

Abstract. Genome-wide association (GWA) studies provide large amounts of high-dimensional data. GWA studies aim to identify variables, i.e., single nucleotide polymorphisms (SNP) that increase the risk for a given phenotype and have been successful in identifying susceptibility loci for several complex diseases. A remaining challenge is however to predict the individual risk based on the genetic pattern. Counting the number of unfavorable alleles is a standard approach to estimate the risk of a disease. However this approach limits the risk prediction by only allowing for a subset of predefined SNPs. Recent studies that apply SVM-learning have been successful in improving the risk prediction for Type I and II diabetes. However, a drawback of the SVM is the poor interpretability of the classifier. The aim is thus to classify based on only a small number of SNPs in order to also allow for a genetic interpretability of the resulting classifier. In this work we propose an algorithm that can do exactly this. We use an approximation method for sparse linear regression problems that has been recently proposed and can be applied to large data sets in order to search for the best sparse risk predicting pattern among the complete set of SNPs.

1 Introduction

There are three general aims of genome-wide association (GWA) studies: identifying genetic loci or patterns associated with common complex multifactorial diseases; understanding the complex genetic mechanisms underlying the disease; and predicting the individual risk of the disease based on the genetic patterns. In the past decade, GWA studies have been successfully employed to identify genetic loci (SNPs) associated to common complex diseases such as diabetes [13], myocardial infarction [11, 4], and Crohn's disease [10]. However, so far these findings have only limited impact on risk assessment and clinical treatment [6, 5]. A reason for the lack of feasible risk prediction by means of genetic variants can be explained by the fact that a disease effect may only come about through the interaction of multiple loci. Studies that focus on single locus effects alone are thus not likely to reveal the more complex genetic mechanisms underlying multifactorial traits [13, 12, 7]. To understand the underlying genetics of a disease, it might be feasible to identify sparse patterns that influence the risk. Such patterns can be helpful not only to assess the risk, but also to detect

possible genetic interactions. However, it is not straightforward to identify multiple SNPs that together increase the risk. Testing all possible combinations is not possible due to the typically enormous amount of SNPs in a GWA study.

The standard method for computing a genotype score (GS) that is used in order to predict the individual risk, is to count the number of unfavorable alleles (those associated to the disease) [3, 2, 1]. However, the drawback of this approach is rather obvious: if we only account for SNPs that are directly associated to the disease we will not allow for interactions between SNPs without an association. Validated susceptibility loci only explain a small proportion of the genetic risk and thus by only accounting for these it is not likely to gain a significant increase in risk prediction [6, 13]. Thus, a major improvement in risk prediction can only be achieved by training a multivariate classifier taking all SNPs into account. Since we expect the SNPs to have different influence on the risk, we need to apply weights to the SNPs in the classifier which again is not straightforward.

A classical tool for classification accounting for all features is the support vector machine (SVM) [15, 16]. The advantage of the SVM is that it is applicable to very large datasets and has already been applied successfully on GWA data [13, 17, 18]. However, the disadvantage of the SVM when it comes to GWA data is that the learned classifier is based on all SNPs, which makes it difficult to interpret the resulting classifier in a biological context. Thus, since we aim to classify based only on a small number of SNPs, we need to apply an appropriate SNP selection in advance and train the SVM only on these. A standard approach is to select the SNPs based on single significance values (p-values). However, this leads to the same issues as described for the GS approach except that a weighting of the SNPs is done by the SVM. Thus, if we aim to predict the risk of an individual as well as understanding the complex genetic mechanisms underlying the disease, the SVM might not be the appropriate choice.

In this work, we propose a novel method for GWA analysis that searches for a sparse risk predictor on the complete data. Predicting the phenotype from the genotype is approached in the framework of sparse linear regression, i.e., our method determines a set of weights that generate the phenotype as a sparse linear combination of the genotype.

2 Data & Methods

2.1 Data

We simulated case/control datasets using the PLINK software [8, 9] with the `SIMULATE` option. We simulated a dataset of 5000 cases and 5000 controls. 100 SNPs of a total of 10100 SNPs where associated to the disease phenotype. Common complex diseases have typically low effect size [6], hence we simulated the effect size i.e. odds ratio (OR) of 1.3 and 1.6 for heterozygous and a multiplicative risk of 2.6 and 3.2 respectively for the homozygous. As previously described most of the identified genetic variants have only limited impact on risk assessment and clinical treatment. This missing heritability might be caused

by the fact that the main contribution result from variants with low minor allele frequency (MAF) and such variants are difficult to detect [6]. Thus, we simulated datasets with MAFs ranging from 0.05 to 0.1 and 0.1 to 0.2. Varying the described parameters we gain a total of 4 different datasets as shown in Table 2.1.

	MAF_{min}	MAF_{max}	OR
dataset type 1	0.05	0.1	1.3
dataset type 2	0.1	0.2	1.3
dataset type 3	0.05	0.1	1.6
dataset type 4	0.1	0.2	1.6

Table 1. Simulated datasets obtained with the PLINK software.

We obtain a genotype matrix G with $G = (\mathbf{g}_1, ..., \mathbf{g}_L)^T$,$\mathbf{g}_i \in \mathbb{R}^d$ where L is the number of individuals (10000) and d the number of SNPs (10100). Furthermore, we have phenotype labels $\mathbf{p} = (p_1, .., p_L)$, $\mathbf{p} \in \mathbb{R}^L$, $p_i \in \{-1, 1\}$. For the cases $p_i = 1$ holds whereas for the controls we have $p_i = -1$. We divided the dataset into two sets of equal size, one set for training and the other one for testing.

2.2 Genotype Score

A genotype score (GS) is calculated on the basis of the number of risk alleles (those associated with the disease phenotype) that are carried by each individual for a predefined subset of SNPs S. The subset S is selected according to their p-values estimated with the Chi-square statistics. The number of selected SNPs, i.e. $|S| = k$, is varied from 1 to 20. For the genotype data $G_{ij} \in \{0, 1, 2\}$ holds. The encoding is performed such that G_{ij} corresponds to the number of risk alleles that are carried by individual i at SNP location j. The homozygous genotype for the risk allele is coded with 2, heterozygous with 1 and homozygous for the non risk allele with 0. Hence, the GS for an individual i of the test set is computed by

$$R_i = \sum_{j \in S} G_{ij} \, . \tag{1}$$

2.3 Support Vector Machine

A support vector machine (SVM) determines the hyperplane that separates two given classes with maximum margin [19]. It has been applied to a broad range of classification problems and is among the methods that are used as benchmark in many cases. In order to measure the classification performance that is obtained using the set of SNPs S (selected as for the GS approach), we train a Gaussian-kernel SVM on the genotype data of the selected SNPs of the

training set. The hyper-parameters, i.e., kernel width and softness of the margin are adjusted by 10-fold cross-validation on the training set.

2.4 Sparse Linear Regression

In contrast to the GS and SVM approaches, in the sparse linear regression (SLR) approach the set of selected SNPs is not obtained from the p-values of the chi-square statistics but the selection of a predefined number of SNPs is performed automatically by the method. We consider the following optimization problem

$$\mathbf{w}_{SP} = \arg\min_{\mathbf{w}} \|\mathbf{p} - G\mathbf{w}\| \quad \text{subject to} \|\mathbf{w}\|_0 = k \; . \tag{2}$$

We are looking for a weight vector $\mathbf{w}$ that approximates the phenotype vector $\mathbf{p}$ as a linear combination of the SNPs G where the number of non-zero entries of the weight vector is equal to k[1]. It has been shown that (2) is a NP-hard combinatorial problem. A number of approximation methods such as Optimized Orthogonal Matching Pursuit (OOMP) [20] or Basis Pursuit [21] have been proposed that provide close to optimal solutions in benign cases [22]. The bag of pursuits method (BOP) can also be applied to this optimization problem. It is derived from the OOMP and performs a tree-like search that employs a set of optimized orthogonal matching pursuits [14]. The larger the number of pursuits used in the search is, the closer BOP approximates $\mathbf{w}_{SP}$. In this work the number of pursuits was set to 100. In contrast to BP, it does not lead to a quadratic optimization problem which might become computational very demanding since nowadays very large genotype data is available ($\approx 10^4$ individuals, $\approx 10^6$ SNPs). Let $\mathbf{w}_{BOP}$ be the approximation of $\mathbf{w}_{SP}$ that has been determined using the BOP method on the training data. The decision value of a given test individual $\mathbf{g}_i$ is obtained as $d_i = \mathbf{w}_{BOP}^T \mathbf{g}_i$. Again, the number of selected SNPs, i.e., k, is varied from 1 to 20.

3 Results

We trained and tested the GS, SVM, and SLR with varying numbers of selected SNPs. We evaluated the performance of the three algorithms by means of the receiver operator characteristic (ROC) obtained on the test set. The area under this curve (AUC) is a commonly used approach to evaluate the performance of a binary classifier. We generated a total of 5 random datasets for each choice of the simulation parameters, i.e., odds ratio (OR) and minor allele frequency (MAF) as described in section 2.1. Then, for a varying number of selected SNPs, i.e., $k = 1, \ldots, 20$, we evaluated the classification performance of the three approaches by means of their mean AUC for each of the data types respectively.

[1] $\|\mathbf{w}\|_0$ is the number of non-zeros in $\mathbf{w}$

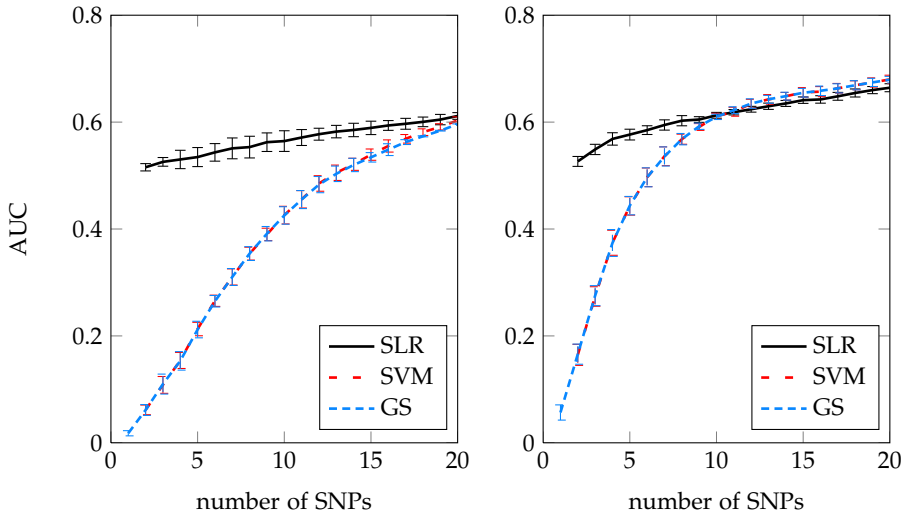

Fig. 1. OR 1.3, left: MAF 0.05-0.1, right: MAF 0.1-0.2

3.1 Dataset 1 and 2: Odds Ratio 1.3

We first compared the three algorithms on datasets with an OR of 1.3. In order to explore the effects of a higher OR on the performance of the algorithms, we then tested the three approaches on datasets with an OR of 1.6.

MAF: 0.05-0.1: As shown in Figure 1(left) the SLR approach clearly outperforms the traditional genotype score and the SVM for small numbers of selected SNPs. The performance of all three approaches improves with increasing numbers of selected SNPs and is almost the same for large numbers of selected SNPs. The performance of the GS and the SVM is very similar due to the selection of the SNPs based on the p-values that is performed for these methods. The SVM only marginally improves the performance for larger numbers of SNPs compared to the GS approach. In particular for small numbers of selected SNPs almost the same performance as for the GS is obtained.
MAF: 0.1-0.2: Figure 1(right) shows the performance of the methods on datasets with a MAF between 0.1 and 0.2. The results are qualitatively the same. However, if the number of selected SNPs is very large, the SVM and GS achieve better results than the SLR method.

3.2 Dataset 3 and 4: Odds Ratio 1.6

The performance of all three approaches on datasets with an OR of 1.6 improves compared to datasets having an OR of 1.3. This is not surprising since a higher OR implies that each single significant SNP is a stronger classifier.

MAF: 0.05-0.1: As for the results that were obtained from datasets with an OR of 1.3, the performance of the SLR method is best for small amounts of SNPs (see Figure 2,left). However, in contrast to the previous results, the GS and SVM approaches close up to the SLR and even become slightly better for larger numbers of SNPs.
MAF:0.1-0.2: Compared to the results on the dataset with a MAF of 0.05 -0.1, the performance only improves marginally as shown in Figure 2(right). In contrast to the results that were obtained on datasets with an OR of 1.3 the SLR method is not beaten by the two other approaches but rather equally good for large numbers of SNPs.

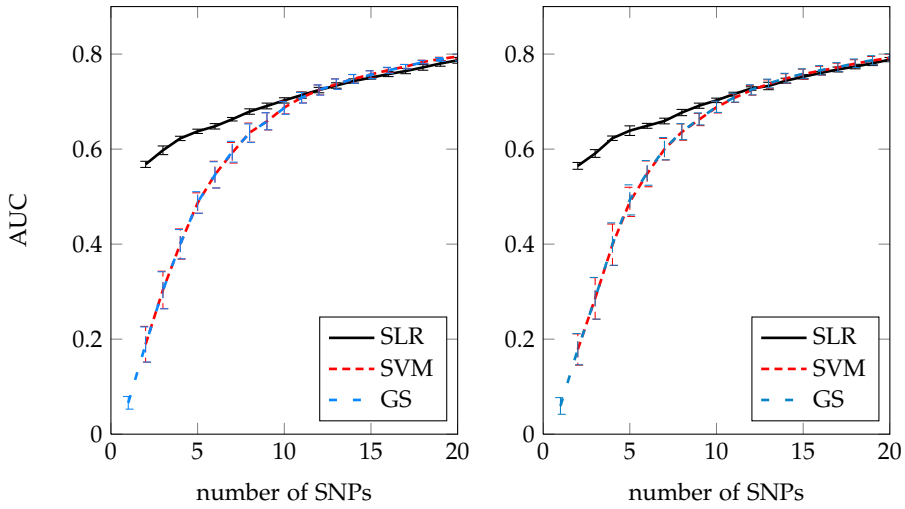

Fig. 2. OR 1.6, left: MAF 0.05-0.1, right: MAF 0.1-0.2

4 Conclusion

The aim of GWA studies is not only to perform risk predictions based on the genetic patterns but also to identify a small set of susceptibility loci in order to understand the genetic mechanisms underlying the disease. The set of susceptibility loci should be as small as possible in order to enable an analysis of the biological mechanisms that correspond to the loci that have been selected. The standard genotype score (GS) that counts the number of unfavorable alleles limits the risk prediction to be based only on SNPs associated to the disease, i.e., that are significant according to their p-value. Since we do not expect that a small number of SNPs that have been selected according to their p-values

can explain more than a small proportion of the genetic risk this approach is not sufficient. The risk prediction can be greatly improved by employing more powerful methods such as SVM-learning. However, for better interpretability of the classifier to understand the genetic mechanisms of the disease, we want to classify using only a small number of SNPs. Therefore, we again have to select SNPs on the basis of prior knowledge. In this work we ranked the SNPs by the p-values and trained the SVM on the best ranked SNPs which is a standard approach in GWA data analysis but often leads to weak performance if the number of selected SNPs is small.

In this paper, we approached the selection problem in the framework of sparse linear regression (SLR). The advantage of the SLR approach is that it does not need any a priori assumptions and thus does not have any limitations due to a preselection step. We applied the bag of pursuits method to the SLR problem which is possible even on huge data sets. We compared the three methods GS, SVM and SLR on GWA data that has been simulated using the PLINK software. If the set of selected SNPs is small, the SLR approach clearly outperforms SVM and classical GS approaches. Even though the performance of the other methods comes close to the performance of SLR if the number of SNPs is large enough, the presented results suggest that SLR should be the method of choice, in particular if the aim is not only to assess the risk of a disease but also to better understand the genetic mechanisms underlying the disease.

Acknowledgments. This work was supported by the Graduate School for Computing in Medicine and Life Sciences funded by Germany's Excellence Initiative [DFG GSC 235/1]

References

1. Sekar Kathiresan, Olle Melander, Dragi Anevski, Candace Guiducci, et al. Polymorphisms Associated with Cholesterol and Risk of Cardiovascular Events. *N Engl J Med*, 358(12):1240–1249, 2008.
2. Jianhua Zhao, Jonathan P. Bradfield, Mingyao Li, Kai Wang, et al. The role of obesity-associated loci identified in genome wide association studies in the determination of pediatric BMI. *Obesity (Silver Spring, Md.)*, 17(12):2254–2257, December 2009. PMID: 19478790 PMCID: 2860782.
3. Jianhua Zhao, Mingyao Li, Jonathan P Bradfield, Haitao Zhang, et al. The role of height-associated loci identified in genome wide association studies in the determination of pediatric stature. 11:96–96. PMID: 20546612 PMCID: 2894790.
4. Jeanette Erdmann, Anika Großhennig, Peter S. Braund, Inke R. König, et al. New susceptibility locus for coronary artery disease on chromosome 3q22.3. *Nat Genet*, 41(3):280–282, Mar 2009.
5. John P.A. Ioannidis. Prediction of Cardiovascular Disease Outcomes and Established Cardiovascular Risk Factors by Genome-Wide Association Markers. *Circ Cardiovasc Genet*, 2(1):7–15, February 2009.
6. Teri A. Manolio, Francis S. Collins, Nancy J. Cox, David B. Goldstein, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, October 2009.

7. J. H. Moore. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human heredity*, 56(1-3):73–82, 2003.
8. Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3):559–575, September 2007. PMID: 17701901.
9. Shaun Purcell. http://pngu.mgh.harvard.edu/purcell/plink/
10. J. V. Raelson, R. D. Little, A. Ruether, H. Fournier, B. Paquin, P. Van Eerdewegh, W. E. Bradley, et al. Genome-wide association study for Crohn's disease in the Quebec Founder Population identifies multiple validated disease loci. *Proc Natl Acad Sci U S A*, 104(37):14747–14752, 2007.
11. Nilesh J. Samani, Jeanette Erdmann, Alistair S. Hall, , Christian Hengstenberg, et al. Genomewide Association Analysis of Coronary Artery Disease. *N Engl J Med*, 357(5):443–453, 2007.
12. Naomi R Wray, Michael E Goddard, and Peter M Visscher. Prediction of individual genetic risk of complex disease. *Current Opinion in Genetics and Development*, 18(73):257–263, 2008.
13. Zhi Wei, Kai Wang, Hui-Qi Q. Qu, Haitao Zhang, et al. From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS genetics*, 5(10):e1000678+, October 2009.
14. Kai Labusch and Thomas Martinetz. Learning Sparse Codes for Image Reconstruction. In Michel Verleysen, editor, *Proceedings of the 18th European Symposium on Artificial Neural Networks*, pages 241–246. d-side, 2010.
15. Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, 2003.
16. Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46:389–422, 2002.
17. Y. Yoon, J. Song, S. H. Hong, and J. Q. Kim. Analysis of multiple single nucleotide polymorphisms of candidate genes related to coronary heart disease susceptibility by using support vector machines. *Clin Chem Lab Med*, 41(4):529–534, 2003.
18. H. J. Ban, J. Y. Heo, K. S. Oh, and K.J. Park. Identification of Type 2 Diabetes-associated combination of SNPs using Support Vector Machine. *BMC Genet.*, 11(1):26, 2010.
19. V. N. Vapnik. Statistical Learning Theory *Wiley*,1998.
20. L. Rebollo-Neira and D. Lowe. Optimized orthogonal matching pursuit approach *IEEE Signal Processing Letters*, 9(4):137-140, 2002.
21. S. S. Chen, D. L. Donoho and M. A. Saunders. Atomic Decomposition by Basis Pursuit. *SIAM Journal on Scientific Computing*, 20(1):33-61, 1998.
22. J. A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231-2242, 2004.