Constructing Genetic Risk Scores through Statistical Learning

Michael Lau[1,2], Tamara Schikowski[2], Holger Schwender[1]

[1]Mathematical Institute, Heinrich Heine University, Düsseldorf, Germany
[2]IUF – Leibniz Research Institute for Environmental Medicine, Düsseldorf, Germany

Genetic Risk Scores

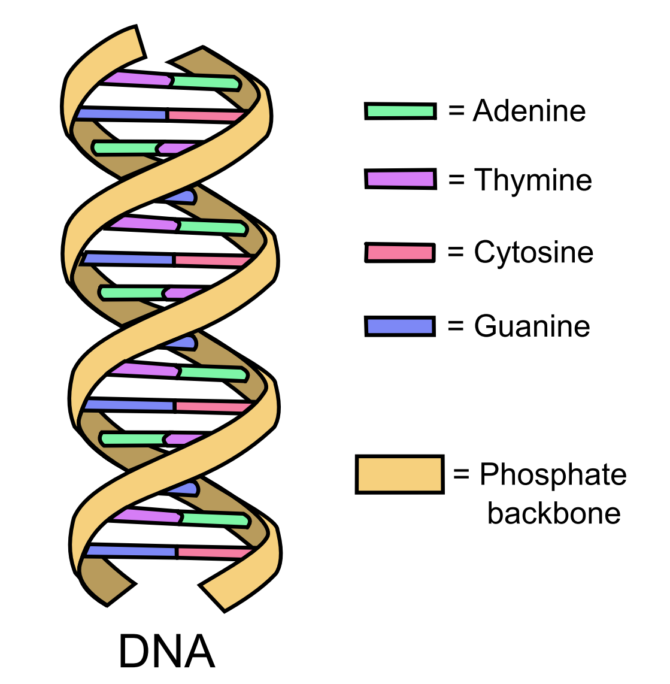

Figure 1: The structure of human DNA

One main goal of genetic epidemiology is finding the underlying mechanisms of disease developments with regard to genetic features. Consequently, the unique DNA fingerprints consisting of sequences of nucleotide bases are jointly analyzed with disease outcomes and environmental factors that can also play a role in developing certain diseases.

GRS (genetic risk scores) summarize genetic features of individuals in a single statistic with regard to a certain disease (e.g., type II diabetes) and a specific subset of the DNA (e.g., the TCF7L2 gene). Their main purpose is to

- derive which genetic loci influence the disease development in which interplay and to
- construct highly predictive models for disease prevention in precision medicine.

In practice, GRS are constructed using a generalized linear model (GLM)

$$g(\mathbb{E}[Y \mid \boldsymbol{X} = x]) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,$$

in which Y may be a binary disease status or a quantitative outcome (e.g., blood pressure), $\boldsymbol{X}$ is a p-dimensional vector of genetic features, usually SNPs (single nucleotide polymorphisms) that encode base-pair substitutions at certain loci, and g is a link function. The weights β_i are in practice obtained by gathering the effect sizes of single SNPs in independent association studies, i.e., the weights usually result from univariate regression models.

Statistical Learning Approach for Constructing GRS

The problem of constructing GRS can also be stated more generally. A function $\text{GRS} : \mathcal{X} \to \mathcal{Y}$ is to be found that resembles the true regressor $\mathbb{E}[Y \mid \boldsymbol{X} = x]$ as close as possible. Now, given epidemiological data, this problem can be addressed using statistical learning procedures that are capable of extracting knowledge using training data sets with small samples sizes as it usually is the case for data sets from epidemiological studies.

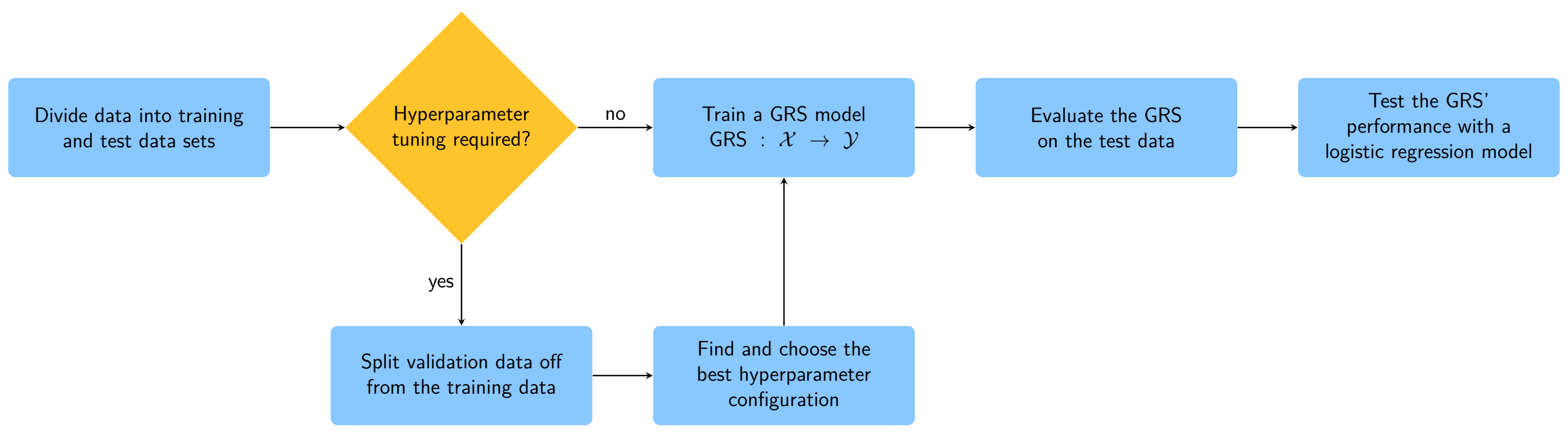
Divide data into training and test data sets
Hyperparameter tuning required?
no
yes
Train a GRS model GRS : $\mathcal{X} \to \mathcal{Y}$
Evaluate the GRS on the test data
Test the GRS' performance with a logistic regression model
Split validation data off from the training data
Find and choose the best hyperparameter configuration

Figure 2: Workflow of constructing and evaluating GRS using statistical learning

Tree-Based Statistical Learning Methods I

A reference procedure commonly used for constructing GRS is the elastic net creating sparse linear models.

However, linear models cannot take interaction effects between features into account unless specifying which loci might interact prior to fitting the model. Since this precise prior knowledge is usually not available, methods that can identify interactions on their own might be preferable.

The tree-based statistical learning methods random forests, an ensemble of randomized decision trees, and logic regression are able to achieve this.

Tree-Based Statistical Learning Methods II

Since the feature space of GRS consists of discrete variables, random forests and logic regression are both theoretically able to cover each possible prediction scenario. For stabilizing single logic regression models, we also evaluated logic bagging – an application of bagging to logic regression.

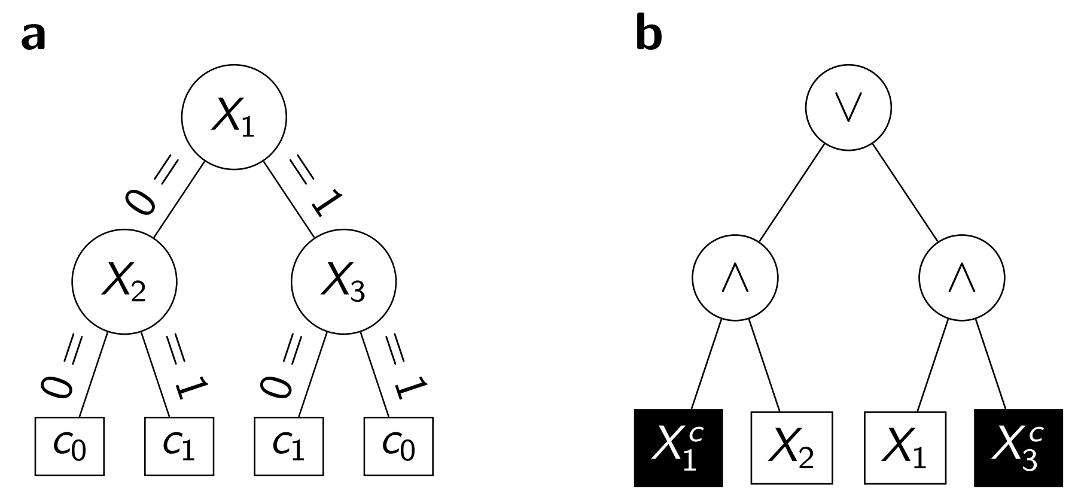

Figure 3: Exemplary tree models describing the same prediction function. **a** Classification tree. **b** Logic tree.

Results

Random forests, random forests VIM (random forests with a prior variable selection), logic regression, logic bagging, and elastic net were evaluated in an extensive simulation study consisting of 92 data scenarios in total. Figure 4 shows an excerpt from this evaluation for varying interaction effects. For increasing effect sizes, the tree-based procedures lead to increased predictive performances compared to elastic net. Even for solely marginal genetic effects, the tree-based methods tend to outperform elastic net. See [1] for details.

For confirming our results on real data, we evaluated a data set from a cohort study (the SALIA study conducted at the IUF) and constructed GRS for rheumatoid arthritis. We followed two approaches,

- a gene-based construction approach where influential genes from a literature research were selected and all available SNPs in these genes were used to construct the GRS and
- a genome-wide approach being the complement to the first approach by excluding genetic variants located in these genes.

Again, the tree-based methods seem to yield superior predictive performances, highlighting that the ensemble methods random forests and logic bagging lead to the highest results overall.

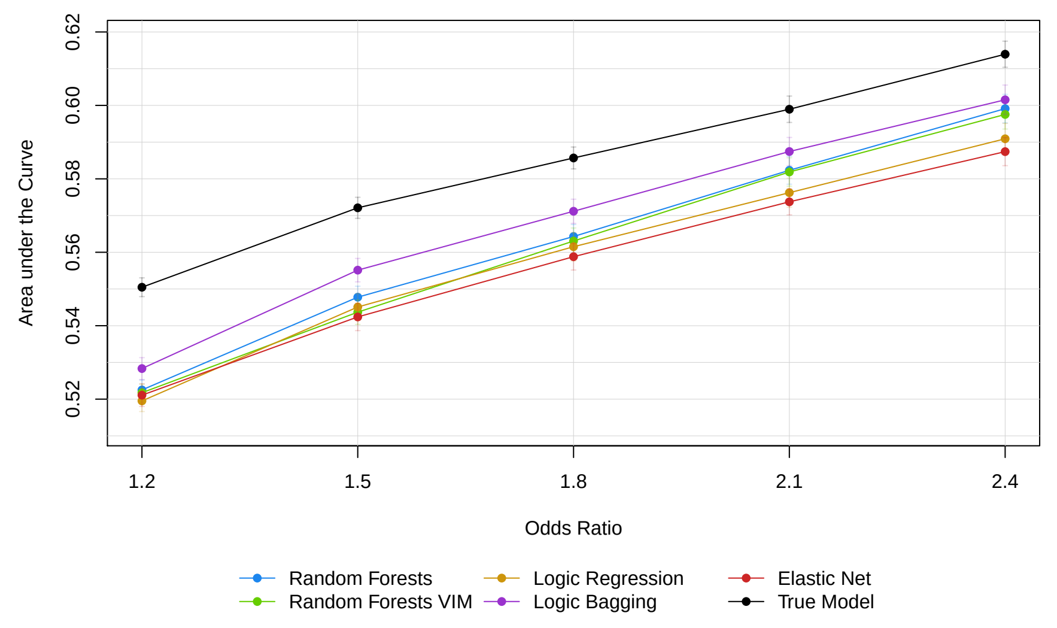

Figure 4: Comparison of the predictive performances of the regarded statistical learning procedures as part of a simulation study

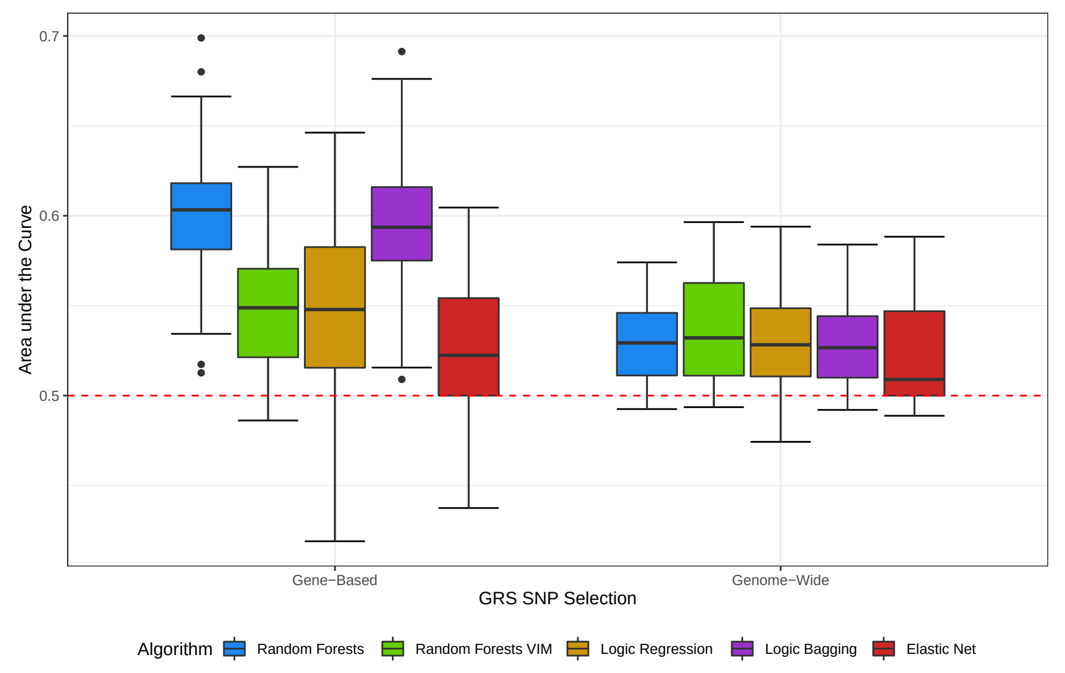

Figure 5: Predictive performances of GRS constructed for rheumatoid arthritis as part of a real data application

Logic Decision Trees

It became clear that random forests and logic bagging can create highly predictive GRS. However, these ensemble models are black boxes that can no longer be easily interpreted. Furthermore, SNP interactions involving negligible marginal effects and interactions with continuous environmental variables might be missed. Hence, we developed a procedure that is aimed at both interpretability and predictive ability, **logicDT (Logic Decision Trees)**.

A global search using simulated annealing is carried out for identifying the ideal set of Boolean conjunctions of input variables which are then used as splitting variables in the decision trees. Environmental covariables (such as the exposure to NO_2) are used to fit continuous regression models in the leaves for fully exploiting the data structure.

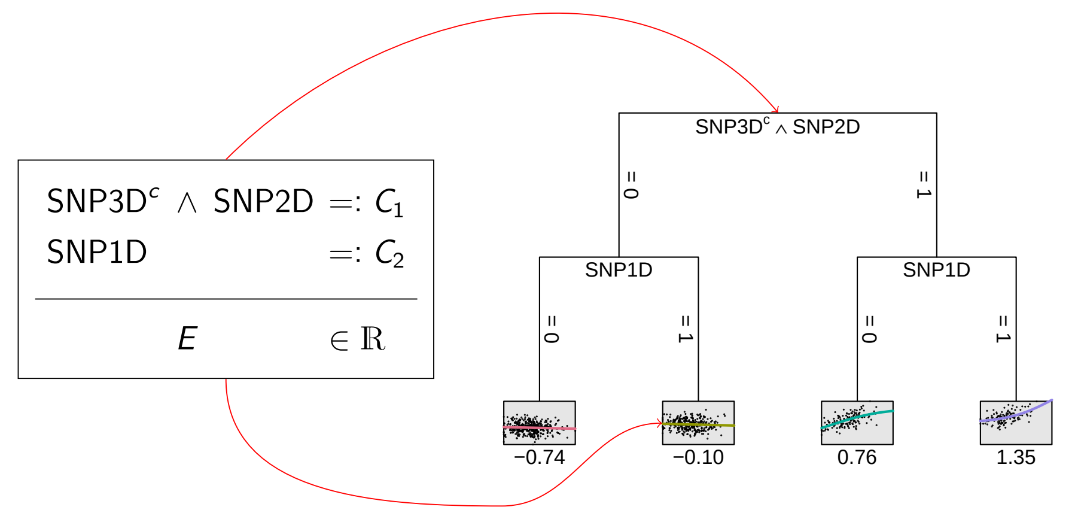

Figure 6: An exemplary logic decision tree using the identified set of conjunctions as splitting terms and the continuous covariable for fitting regression models in the leaves

References

[1] M. Lau, C. Wigmann, S. Kress, T. Schikowski, and H. Schwender. Evaluation of tree-based statistical learning methods for constructing genetic risk scores. *BMC Bioinformatics*, 23:97, 2022.

HeiCAD Lightning Talks 2022 June 24, 2022 michael.lau@hhu.de