Statistical Applications in Genetics and Molecular Biology

Volume 11, *Issue* 4 2012 *Article* 13

Cluster-Localized Sparse Logistic Regression for SNP Data

Harald Binder, *Institute of Medical Biostatistics, Epidemiology and Informatics, University Medical Center Johannes Gutenberg University Mainz*
Tina Müller, *Global Drug Discovery Statistics, Bayer Pharma AG*
Holger Schwender, *Faculty of Statistics, TU Dortmund University*
Klaus Golka, *Department of Toxicology, IfADo - Leibniz Research Centre for Working Environment and Human Factors*
Michael Steffens, *Institute of Medical Biostatistics, Epidemiology and Informatics, University Medical Center Johannes Gutenberg University Mainz*
Jan G. Hengstler, *Department of Toxicology, IfADo - Leibniz Research Centre for Working Environment and Human Factors*
Katja Ickstadt, *Faculty of Statistics, TU Dortmund*

Recommended Citation:
Binder, Harald; Müller, Tina; Schwender, Holger; Golka, Klaus; Steffens, Michael; Hengstler, Jan G.; Ickstadt, Katja; and Schumacher, Martin (2012) "Cluster-Localized Sparse Logistic Regression for SNP Data," *Statistical Applications in Genetics and Molecular Biology*: Vol. 11: Iss. 4, Article 13.

Martin Schumacher, *Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg*

Cluster-Localized Sparse Logistic Regression for SNP Data

Harald Binder, Tina Müller, Holger Schwender, Klaus Golka, Michael Steffens, Jan G. Hengstler, Katja Ickstadt, and Martin Schumacher

Abstract

The task of analyzing high-dimensional single nucleotide polymorphism (SNP) data in a case-control design using multivariable techniques has only recently been tackled. While many available approaches investigate only main effects in a high-dimensional setting, we propose a more flexible technique, cluster-localized regression (CLR), based on localized logistic regression models, that allows different SNPs to have an effect for different groups of individuals. Separate multivariable regression models are fitted for the different groups of individuals by incorporating weights into componentwise boosting, which provides simultaneous variable selection, hence sparse fits. For model fitting, these groups of individuals are identified using a clustering approach, where each group may be defined via different SNPs. This allows for representing complex interaction patterns, such as compositional epistasis, that might not be detected by a single main effects model. In a simulation study, the CLR approach results in improved prediction performance, compared to the main effects approach, and identification of important SNPs in several scenarios. Improved prediction performance is also obtained for an application example considering urinary bladder cancer. Some of the identified SNPs are predictive for all individuals, while others are only relevant for a specific group. Together with the sets of SNPs that define the groups, potential interaction patterns are uncovered.

KEYWORDS: single nucleotide polymorphisms, weighted regression, clustering

Author Notes: Harald Binder and Martin Schumacher (DFG Forschergruppe FOR 534) and Holger Schwender (SCHW 1508/3-1) gratefully acknowledge support from Deutsche Forschungsgemeinschaft. In addition, Holger Schwender and Katja Ickstadt gratefully acknowledge the support by the Deutsche Forschungsgemeinschaft, Research Training Group ``Statistical Modelling'' of the Department of Statistics, TU Dortmund. We thank Arif B. Ekici (University Hospital Erlangen, Germany) for performing the preprocessing of the data for the application example. We would like to acknowledge the provision of population controls from PopGen (S. Schreiber, M. Krawczak, H. v. Eberstein; University of Kiel) and from KORA-gen (H. E. Wichmann; Helmholtz Center Munich), which were genotyped within the GWAS Add-on project of the NGFN-2.

1 Introduction

Single nucleotide polymorphism (SNP) data promise compelling insight into the molecular basis of diseases. Such data are often analyzed in a case-control setting, in which two groups consisting of unrelated individuals are considered, one with affected individuals, and a second with individuals not showing the disease. In genome-wide association studies, measurements for a huge number of SNPs are available for each individual. Typically, the SNPs are analyzed individually by univariate statistical tests, comparing the two groups with respect to each SNP (see, e.g., Ziegler et al., 2008).

Several SNPs can be investigated simultaneously by considering the case-control status as a binary response and incorporating SNP information as covariates into a logistic regression model. For example, Wu et al. (2009) adapted the lasso for fitting such models with genome-wide data, allowing for an automatic selection of a small number of SNPs that are predictive with respect to the case-control status. Such regression modeling approaches have the additional advantage that predictions can also be obtained for new individuals. The predicted probabilities from such models might indicate the degree of an individual's susceptibility to the investigated disease, thus potentially contributing towards personalized medicine. Silver et al. (2012) proposed a method based on the group lasso (Yuan and Lin, 2006) for detecting pathways associated with a quantitative trait. Ayers and Cordell (2010) compared the lasso and other regularized regression methods with standard test procedures and forward stepwise regression in their application to genotype data, and showed that the former approaches outperform single marker analysis.

Recently, interactions between genes have received renewed attention. Specifically, *compositional epistasis* is considered important, which means that the contribution of two or more genes is needed for having an effect on the phenotype, e.g., the case-control status (Phillips, 2008). Depending on the specific interaction pattern, classical statistical techniques for dealing with interactions might be problematic (VanderWeele, 2010). Test-based approaches have been extended to account for different types of interaction patterns (see, e.g., Schwender et al., 2011). However, due to the large number of SNPs, even two-way interactions require sophisticated algorithms for a genome-wide search (Steffens et al., 2010), and already three-way interactions require some kind of heuristic (see, e.g., Kayano et al., 2009). The BEAM approach (Zhang and Liu, 2007), e.g., partitions SNPs into three groups, where one group comprises potentially interacting SNPs, allowing for subsequent tests for interactions. Wu et al. (2009) suggest to use the lasso for selecting a small number of predictive SNPs and then explore interactions between these. However, this method only works if main effects of SNPs with interactions are large enough to be selected by the lasso in the first step. Similarly, Yang et al. (2010) propose a procedure based on the group lasso for detecting main effects and epistatic interactions, which, however, also requires a prefiltering of the SNPs when applied to genome-wide data. Logic regression provides a direct approach for regression models with terms for a combined effect of SNPs (Ruczinski et al., 2003, Schwender and Ickstadt, 2008), but it is so far restricted to about 1000 covariates. Similarly, the MECPM approach (Miller et al., 2009), which learns the empirical class probability under constraints on SNP interaction structure, is difficult to implement for a huge number of SNPs.

Two other approaches for directly searching for interactions, random forests and multifactor dimensionality reduction, have only recently been adapted for genome-wide data (Schwarz et al., 2010, Greene et al., 2010). However, the performance of multifactor dimensionality reduction relative to other multivariable approaches has been found to be problematic in low-dimensional settings (Park and Hastie, 2008). Winham et al. (2011) provide a more detailed comparison to lasso-based approaches. Random forests, while known to result in good prediction performance, does not provide one model for prediction for new individuals, but uses an ensemble of a large number of trees. This makes interpretation and extraction of important interactions difficult.

In the following, we propose an alternative approach, cluster-localized regression (CLR), for relaxing the assumption of a pure main effects model without having to explicitly search for interactions. The CLR approach is inspired by localized logistic regression (Loader, 1999, Tutz and Binder, 2005), which has successfully been applied to a small number of SNPs (Müller et al., 2010), but could so far not directly be used in a genome-wide setting.

For motivating our new proposal, imagine two groups of individuals mainly defined by one specific SNP. For example, the frequency of the minor allele might be much higher in one of the two groups than in the other. Such a difference might, of course occur, when considering different populations. Here, we assume, however, that the individuals are from the same population, and neither the two groups nor the SNPs are known. Within each of the two groups, the risk for being a case might be predicted by a second SNP, where the effects are different between the groups. For example, in the group defined by high minor allele frequency of the first SNP, presence of the minor allele for the second SNP might be connected to increased risk while not being connected to the risk in the other group. Such a pattern is equivalent to an interaction, as the risk with respect to all individuals depends on the combination of two SNPs.

Naturally, a group may be defined by many SNPs, and also many SNPs might be important for risk prediction within that group. Such group definitions automatically take the genomic background into account when searching for risk-related SNPs. Consideration of the genomic background has recently also been suggested as a promising approach for dealing with rare variants (Bansal et al., 2010). Statistical techniques for localizing models to such groups of individuals can adapt main effects models for representing structure that has so far mostly been addressed by explicit use of interaction terms. Besides a description of SNP effects in terms of local models, which might be natural in some settings, localization techniques can also be considered for identifying interactions when this is needed.

We adapt the clustering approach of Friedman and Meulman (2004) for identifying groups of individuals, such that they are mainly defined via a small number of SNPs, where different SNPs might be important for different groups. For identifying important SNPs within each group, logistic regression models are fitted. We adapt a componentwise likelihood-based boosting approach (Tutz and Binder, 2007), which is similar to the lasso, for fitting a sparse logistic regression model locally for each group, i.e., for selecting a small number of predictive SNPs for each group. The combination of SNPs important for defining a group and SNPs found to be important in the regression model for this group then provides candidate interactions for further validation.

When employing componentwise boosting for estimating regression models, the number of covariates with estimated non-zero effect sizes is typically much smaller than the number of individuals, as often only one of several correlated covariates is incorporated. In our situation, thus typically only one SNP from a block of SNPs in strong linkage disequilibrium (LD) is selected. A number of less than, say, 100 SNPs would be useful for keeping subsequent validation steps manageable. However, as recent results indicate that a larger number might provide better prediction performance (Evans et al., 2009), we will select model complexity, i.e., the number of groups and the number of SNPs with non-zero effects, automatically based on the data.

The clustering approach for obtaining groups of individuals is described in Section 2. This includes criteria for assigning new individuals to groups and for identifying the SNPs most relevant for the definition of a group. In Section 3, the details of the componentwise boosting approach are given. Practical aspects of implementation, tuning parameter selection, and computational demand, are considered in Section 4. In Section 5, the CLR approach is explored in a simulation study, in which a main effects approach, similar to the one of Wu et al. (2009), serves as a reference. The CLR approach is furthermore illustrated in an application to urinary bladder cancer data in Section 6. Based on clustering results and fitted regression models, we show how different groups of individuals can be identified, and individual multivariable regression models can be fitted for these groups. In a second step, we indicate how a list of candidate interactions could be extracted and further investigated. Concluding remarks are provided in Section 7.

2 Identifying groups of individuals

For fitting cluster-localized regression models, each of $i = 1, \dots, n$ individuals has to be assigned to a group. To obtain predictions for a new individual, furthermore a rule is needed for assigning the new individual to a specific group (and then using the regression model from this group for prediction).

In the following, we describe a clustering approach for empirically obtaining a definition of groups $\mathscr{I}_k \subset \{1, \dots, n\}, k = 1, \dots, K$, i.e., $|\mathscr{I}_k|$ individuals are assigned to the kth group based on their genotypes. We consider the genotypes of q SNPs as given by the numbers $z_{il} \in \{0, 1, 2\}$ of minor alleles, $l = 1, \dots, q$. Thus, 0 indicates the homozygous reference genotype, 1 heterozygosity, and 2 the homozygous variant.

In addition to the clustering approach, a rule for group assignment is provided, based on the clustering results. This rule also provides the basis for a measure for indicating SNPs particularly important for group assignment, i.e. potentially interesting from a biological perspective.

2.1 Clustering with group-specific attribute weights

For obtaining a manageable characterization of each group, the groups should be defined based on as few SNPs as possible. To obtain such group definitions, we use the clustering approach of Friedman and Meulman (2004), which assigns attribute weights to SNPs, where these weights can differ between groups. In addition, this approach has the advantage that the number K of groups does not have to be prespecified, i.e., results for different Ks can be obtained from a single run of the approach. Specifically, the approach of Friedman and Meulman (2004) incorporates attribute weights for building a distance matrix, based on which standard hierarchical clustering approaches can be employed, readily providing solutions for different numbers of groups.

There exist other approaches allowing for clustering with different attribute weights in different groups. For example, Chan et al. (2004) present an extension of the k-means approach for determining attribute weights at the same time as group membership in an iterative approach. This has subsequently been adapted for clustering of SNP data (Ng et al., 2006, Liu et al., 2010). However, these approaches require the number K of groups to be specified prior to clustering. In addition, preliminary experiments indicated that the approach of Friedman and Meulman (2004) provides better results when used as a basis for localized regression.

We provide a brief description of the approach of Friedman and Meulman (2004) for our application in the following. Clustering is performed using the SNP values $z_i = (z_{i1}, \ldots, z_{iq})'$, where each individual i has separate weights $w_i = (w_{i1}, \ldots, w_{iq})'$ for all SNPs. These are used for defining the distance between two individuals with indices i_1 and i_2 as

$$\begin{aligned} D_{i_1 i_2} = \max \Bigg(& -\xi \log \left(\sum_{l=1}^{q} w_{i_1 l} \exp(-d_{i_1 i_2 l}/\xi) \right), \\ & -\xi \log \left(\sum_{l=1}^{q} w_{i_2 l} \exp(-d_{i_1 i_2 l}/\xi) \right) \Bigg), \end{aligned} \tag{1}$$

where $d_{i_1 i_2 l} = \delta_{i_1 i_2 l}/s_l$ with $\delta_{i_1 i_2 l} = I(z_{i_1 l} \neq z_{i_2 l})$ and $s_l = 1/n^2 \sum_i \sum_{i'} \delta_{ii'l}$, $I(\cdot)$ being the indicator function, taking value 1 if its argument is true and 0 otherwise. The value of the parameter ξ is increased in the course of an iterative clustering algorithm, resulting in

$$\lim_{\xi \to \infty} D_{i_1 i_2} = \max \left(\sum_{l=1}^{q} w_{i_1 l} d_{i_1 i_2 l}, \sum_{l=1}^{q} w_{i_2 l} d_{i_1 i_2 l} \right).$$

This iterative algorithm for determining weights w_{il} and distances $D_{i_1 i_2}$ is as follows:

1. Initialize $w_{il} = 1/q, i = 1, \ldots, n, l = 1, \ldots, q$, and $\xi = \phi$.
2. Repeat until the weights w_{il} stabilize:
   (a) Determine distances from (1).
   (b) Determine the N nearest neighbors $\mathscr{I}_{NN,i} \subset \{1, \ldots, n\}$ for each individual $i = 1, \ldots, n$.
   (c) Update the weights via

   $$w_{il} = \frac{\exp(-s_{il}/\phi)}{\sum_{l'=1}^{q} \exp(-s_{il'}/\phi)},$$

   where

   $$s_{il} = \frac{1}{N} \sum_{i' \in \mathscr{I}_{NN,i}} d_{ii'l}.$$

   (d) Update ξ by setting it to $\xi = \xi + \alpha\phi$.

For the parameters ϕ, α, and N, we use values suggested by Friedman and Meulman (2004), specifically $\phi = 0.2$, $\alpha = 0.1$, and $N = \sqrt{n}$.

The distances (1), corresponding to the resulting weights w_{il}, are fed into a standard hierarchical clustering algorithm. This results in a tree that can be cut at various levels for obtaining different numbers K of groups. We use the Ward method for agglomeration in hierarchical clustering, e.g., as implemented in the function `hclust` of the statistical environment R (R Development Core Team, 2011). In contrast to other criteria, this results in groups of similar size, which provides a similar amount of information when fitting the local model for each group.

2.2 Assigning new individuals

For obtaining predictions for a new individual with SNP vector $z^* = (z_1^*, \ldots, z_l^*)'$ from cluster-localized regression, the individual is assigned to the group k^* that is closest to it in terms of the mean distance to members of this group, i.e.,

$$k^* = \underset{k}{\operatorname{argmin}} \frac{1}{|\mathscr{I}_k|} \sum_{i \in \mathscr{I}_k} \sum_{l=1}^{q} w_{il} I(z_l^* \neq z_{il})/s_l. \tag{2}$$

For interpretation, it might be useful to have a measure that indicates SNPs important for defining a group. One such measure is obtained by considering the relevance of a SNP for group assignment. When using the mean distance (2) for group assignment, the corresponding contribution of the SNP l for assigning a new individual to group k^*, i.e., its assignment contribution, is given by

$$\mathrm{AC}_{k^*,l} = \left(\frac{1}{|\mathscr{I}_{k^*}|} \sum_{i \in \mathscr{I}_{k^*}} \frac{w_{il} I(z_l^* \neq z_{il})}{s_l} \right) - \left(\frac{1}{n - |\mathscr{I}_{k^*}|} \sum_{i \notin \mathscr{I}_{k^*}} \frac{w_{il} I(z_l^* \neq z_{il})}{s_l} \right). \tag{3}$$

By averaging this quantity over a set of new individuals that have been assigned to group k^*, the importance of specific SNPs with respect to this group can be assessed. While SNP importance might, in principle, also be seen from (3) for a single individual, averaging over several individuals will reduce the influence of random noise and attenuate SNPs that are more generally important.

3 Cluster-localized regression (CLR) approach

The clustering approach from Section 2 provides a basis for fitting cluster-localized regression models in each of the resulting groups. Generally, case-control SNP data can be analyzed via multivariable regression models by considering a binary response $y_i \in \{0, 1\}, i = 1, \ldots, n,$ (where $y_i = 1$ if that individual is a case and $y_i =$

0 otherwise). Values of $p = 2q$ covariates $x_{ij}, i = 1, \dots, n, j = 1, \dots, p$, are then obtained from the genotypes $z_{il}, l = 1, \dots, q$, of the SNPs as follows:

$$x_{ij} = \begin{cases} I(z_{i\,(j+1)/2} \geq 1) & \text{for } j \in \{1, 3, 5, \dots, p-1\} \\ I(z_{i\,j/2} = 2) & \text{for } j \in \{2, 4, 6, \dots, p\} \end{cases}, \tag{4}$$

so that the x_{ij} with odd j code for a dominant, and the x_{ij} with even j for a recessive effect of the corresponding SNPs.

For investigating the connection between the response y_i and the covariate vector $x_i = (x_{i1}, \dots, x_{ip})'$ for individual i, we use the logistic regression model

$$P(Y_i = 1|x_i) = h(\eta_i) = h(\beta_0 + x_i'\beta) = \frac{1}{1 + \exp(-(\beta_0 + x_i'\beta))}, \tag{5}$$

where β_0 is an intercept term and the parameters in the vector $\beta = (\beta_1, \dots, \beta_p)'$ quantify the influence of covariates coding for the SNPs. The effect of each SNP is represented by two dummy variables, i.e., two β_j. If estimation of the latter is performed in a regularized way, such that many estimated values are equal to zero, model (5) will allow for settings in which there might be a dominant, recessive, or both types of effect. Alternatively, under the assumption of an additive effect of the SNP, the variables z_{il} could also be directly incorporated as a continuous covariate taking the number of minor alleles as values.

3.1 Localized logistic regression

While the parameter vector β in model (5) could be estimated globally for all individuals by using regularized techniques (see, e.g., Wu et al., 2009), we want to obtain separate estimates $\hat{\beta}^{(k)}$ for K disjoint groups $\mathscr{I}_k \subset \{1, \dots, n\}, k = 1, \dots, K$, of individuals, i.e., *local* estimates for each group of individuals. The regression models, each fitted for a group, can then be used for predicting the response for new individuals. While taken literally, the predicted probability for an individual is the probability of being a "case" in the original case-control design, it could also be interpreted as indicating susceptibility with respect to the investigated disease.

When fitting a localized version of model (5) for group $\mathscr{I}_k$, some information should potentially be considered from individuals not in that particular group. This will, for example, increase stability of estimates when the number of SNPs is huge compared to the number of individuals. Therefore, the weighted log-likelihood is used for estimation,

$$l^{(k)}(\beta) = \sum_{i=1}^{n} (y_i \log h(\eta_i) + (1 - y_i) \log(1 - h(\eta_i))) W^{(k)}(x_i). \tag{6}$$

The weights $W^{(k)}(x_i)$ have constant value for $i \in \mathscr{I}_k$. For individuals not in group $\mathscr{I}_k$, lower weights should be assigned, which might be based on the distances to the group members. However, this would require some distance function. For reducing the arbitrariness in defining such a function and for keeping the proposed approach simple, we assign equal weights to all individuals not in the respective considered group. Correspondingly, the weights are given by

$$W^{(k)}(x_i) = \frac{W + (1-W)I(i \in \mathscr{I}_k)}{nW + (1-W)|\mathscr{I}_k|},$$

where $W \in [0,1]$ is a tuning parameter that determines the weight of an individual not in group k, relative to an individual from that group, with value $W = 1$ meaning equal weight.

3.2 Weighted componentwise boosting

In genome-wide association studies, the number of SNPs is much larger than the number of individuals, i.e., direct maximization of the weighted log-likelihood (6) is no longer possible. Tutz and Binder (2005) use a penalized weighted likelihood, in which a penalty term, comprised by the sum of the squared elements of β, is subtracted from (6). Such ridge-type shrinkage moves estimates closer to zero (Hoerl and Kennard, 1970), allowing for estimation in a high-dimensional setting. However, this does not provide variable selection.

In the present application, we want to combine estimation with simultaneous variable selection, i.e., many elements of $\hat{\beta}^{(k)}$ should become zero. Therefore, we adapt componentwise likelihood-based boosting (Tutz and Binder, 2007, Binder and Schumacher, 2008), which provides sparse estimates similar to the lasso (Efron et al., 2004) for the weighted log-likelihood (6). In componentwise boosting, estimates are built up in a large number of small boosting steps, where in each step only one element of the estimated parameter vector is updated.

A componentwise boosting algorithm for estimating $\beta^{(k)}$ in model (5), with weights $W^{(k)}(x_i)$ for a group of individuals $\mathscr{I}_k$, is given in the following. For ease of notation, the additional superscript (k), indicating the specific group for which estimation is performed, is omitted for all quantities:

1. Initialize $\hat{\beta}^{(0)} = (0,\dots,0)'$, $\hat{\eta}^{(0)} = (\hat{\eta}_1^{(0)},\dots,\hat{\eta}_n^{(0)}) = (0,\dots,0)'$.
2. For boosting steps $m = 1,\dots,M$:
   (a) Obtain an estimate $\hat{\beta}_0^{(m)}$ from the intercept model

$$\eta_i = \hat{\eta}_i^{(m-1)} + \beta_0^{(m)}$$

by performing one Newton-Raphson step with respect to the weighted log-likelihood (6), where the offset term $\hat{\eta}^{(m-1)}$ is kept fixed.

(b) Consider candidate models

$$\eta_i = \hat{\eta}_i^{(m-1)} + \hat{\beta}_0^{(m)} + \gamma_j^{(m)} x_{ij}, \quad j = 1, \ldots, p, \tag{7}$$

where estimates $\hat{\gamma}_j^{(m)}$ are obtained from one Newton-Raphson step with respect to the penalized weighted log-likelihood

$$l_{pen}(\gamma) = l(\gamma) - \frac{\lambda}{2}\gamma, \tag{8}$$

with penalty parameter λ.

(c) Update

$$\hat{\beta}_j^{(m)} = \begin{cases} \hat{\beta}_j^{(m-1)} + \hat{\gamma}_j^{(m)} & \text{for } j = j^* \\ \hat{\beta}_j^{(m-1)} & \text{otherwise} \end{cases}$$

and $\hat{\eta}_i^{(m)} = x_i' \hat{\beta}^{(m)}, i = 1, \ldots, n$, where j^* is the index of the covariate that improves the fit the most.

Using one Newton-Raphson step with respect to the penalized weighted likelihood (8), the estimates $\hat{\gamma}_j^{(m)}$ in the candidate models (7) are

$$\hat{\gamma}_j^{(m)} = \frac{S_j^{(m)}}{F_j^{(m)}},$$

where

$$S_j^{(m)} = \sum_{i=1}^{n} W^{(k)}(x_i) x_{ij} \left(y_i - \frac{1}{1 + \exp(-(\hat{\eta}_i^{(m-1)} + \hat{\beta}_0^{(m)}))} \right)$$

is the score function evaluated at zero, and

$$F_j^{(m)} = \sum_{i=1}^{n} W^{(k)}(x_i) x_{ij}^2 \frac{\exp(\hat{\eta}_i^{(m-1)} + \hat{\beta}_0^{(m)})}{(1 + \exp(\hat{\eta}_i^{(m-1)} + \hat{\beta}_0^{(m)}))^2} + \lambda$$

is the scalar value of the Fisher matrix evaluated at zero. This simple form is obtained, since h in model (5) is the response function corresponding to the canonical link in a binary response generalized linear model (for more details see, for example, Fahrmeir and Tutz, 2001).

The estimates $\hat{\gamma}_j^{(m)}$ could be plugged into (8) for determining the index j^* that provides the most improvement in boosting step m. However, for a reduced computational demand, we use the (penalized) score statistic

$$T_j^{(m)} = \frac{(S_j^{(m)})^2}{F_j^{(m)}},$$

(cf. Binder and Schumacher, 2008). The covariate maximizing this expression will often also be the one maximizing the likelihood.

4 Implementation

4.1 Selecting the tuning parameters

When combining the clustering approach with the localized boosting approach from Section 3, three main tuning parameters have to be taken into account: the number K of groups, the relative weight W for individuals from other groups, and the number M of boosting steps.

The number M of boosting steps is the main tuning parameter of the boosting approach that drives the complexity of the fitted models. For example, it determines the maximum number of covariates in the fitted model, as the parameter estimate for a covariate that has not been selected in any boosting step will remain zero, i.e., the covariate is excluded from the model. Therefore, the number of boosting steps has to be chosen carefully for optimizing the prediction performance. Ideally, a different number of boosting steps should be allowed for each group, as the groups might require models of different complexity. However, to simplify the selection of tuning parameters, we here use the same number M for all groups. Although the penalty parameter λ in (8) is of minor importance, it should be chosen such that the updates in the boosting steps are not too large. As a rule of thumb, the parameter estimates for most of the covariates included in the model should be built up in several boosting steps. While the value of λ that achieves this depends on the number of individuals in standard componentwise boosting, this dependency is removed in the weighted version with $\sum_i W^{(k)}(x_i) = 1$.

For the parameters of the clustering algorithm, ϕ, α, and N, we use the fixed values suggested by Friedman and Meulman (2004), as indicated in Section 2.

The tuning parameters K, W, and M can be selected according to prediction performance, as estimated, for example, by 10-fold cross-validation. All model building steps are performed in each corresponding training set. This includes clustering and fitting of models for various numbers K of groups, using the proposed weighted componentwise boosting approach from Section 3 for various values of the relative weight W. Given that only a limited amount of data is available for model fitting, the number of SNPs that can be incorporated without being detrimental to prediction performance will be limited. Therefore, cross-validation avoids overfitting.

Prediction performance in the test sets is evaluated via the Brier score (Brier, 1950), i.e., the mean squared distance between the true status and the predicted probability $h(\hat{\eta}_i)$. The Brier score has mostly been used for judging risk prediction for future events (see, e.g., Gerds et al., 2008). Since, it allows for more detailed judgement as simple misclassification rates, we consider it here also in a case-control setting. Using a grid search, the optimal combination of the number K of clusters, the relative weight W, and the number M of boosting steps is chosen as the combination for which the average Brier score on the test sets is minimal. In the rarely occurring situation of clusters containing only cases or controls (in which thus the group-defining SNPs would coincide with disease-defining SNPs), zero boosting steps would be used for this cluster, i.e. an intercept model would be fitted. In our applications, however, the clusters always contained a mixture of cases and controls.

4.2 Computational demand

The computational demand of the CLR approach is driven by its two components, the clustering algorithm and the componentwise boosting algorithm.

When using cross-validation for tuning parameter selection, clustering has to be performed in each training data set, i.e., ten times for 10-fold cross-validation. Similarly, the boosting approach has to be applied for different tuning parameter values. Due to the nature of the algorithm, results for any smaller number of boosting steps are automatically obtained when evaluating a maximum number of steps, e.g., 500, for finding an optimal number M of steps. This has to be performed for each candidate weight W. Furthermore, when considering up to K groups, boosting has to be performed $K(K+1)/2$ times for each candidate weight. To decrease computational demand, the search grid for W and K has to be rather coarse. For example, five different values of W and values K from 1 to 5 require 750 boosting runs for 10-fold cross-validation.

Such a large number of boosting runs might not seem feasible, but the algorithm is rather fast. In the following, we consider exemplary runs on a single compute core of an Intel Xeon 2.93GHz processor. For $n = 400$ individuals and $p = 1000$ SNPs, $M = 500$ boosting steps took about 6 seconds. For $p = 2000$ SNPs, about 9 seconds were needed, and about 35 seconds for $p = 10000$ SNPs. Computing time is roughly linear in the number of covariates, as would be expected from the construction of the algorithm. For $n = 800$ individuals (and $p = 1000$ SNPs), computation took about 10 seconds, and about 36 seconds for $n = 4000$ individuals, i.e., computing time also is roughly linear in the number of observations. For example, a computing time of less than 150 minutes would be expected for one million SNPs and $n = 1000$ individuals. This would imply a computing time of about 76 days for cross-validation with five values of W and values K from 1 to 5. However, the latter can easily be parallelized, i.e., the actual time can be cut to under one day on a compute cluster with 100 cores.

One iteration of the clustering approach took about 1.3 seconds in an exemplary run for $n = 400$ individuals and $p = 1000$ SNPs. For $p = 2000$ SNPs, about 3 seconds were needed, and about 14 seconds for $p = 10000$ SNPs, i.e., computational demand is, again, roughly linear in the number of SNPs. For $n = 800$ individuals (and $p = 1000$ SNPs), about 8 seconds were needed, and about 313 seconds for $n = 4000$ individuals. Hence, computing time is roughly quadratic in the number of individuals. This means that for $n = 1000$ individuals and one million SNPs, about 6 hours are to be expected for one iteration. The exact number of required iterations depends on the underlying structure, but about 10 iterations are sufficient in our experience. Unfortunately, the iterations are difficult to parallelize, i.e., a straightforward speedup can only be obtained by performing the cross-validation runs in parallel, resulting in a maximum improvement factor of 10. In the above example, this still means a runtime of several days. Besides the recommendations of Friedman and Meulman (2004) for the specification of the tuning parameters, this was another reason not to attempt to optimize the tuning parameters of the clustering approach, as this would no longer be feasible. For the (one) set of suggested parameters, however, computations are still more than feasible.

5 Simulation study

5.1 Design

For investigating the performance of the CLR approach, we consider case-control simulation designs with complex interaction patterns in which the effects of SNPs depend on the effects of other SNPs. SNPs (with values 0, 1, and 2) are generated independently from a binomial distribution considering a pre-specified minor allele frequency. For each simulation scenario described in the following, 50 repetitions are performed, where the performance of the proposed procedure is compared to the ones of global models, as fitted by componentwise boosting, and per-cluster intercept models. The tuning parameters for all of the approaches (the number M of boosting steps for the global models, the number K of clusters for the per-cluster intercept models, and K, M, and W for the CLR approach) are selected by 10-fold cross-validation.

In the most basic scenario, there are two subgroups of individuals in which the minor allele frequencies of 60 SNPs differ substantially. One subgroup has a minor allele frequency of 0.3, the other a minor allele frequency of 0.1. For all other SNPs, the minor allele frequency for both subgroups is set to 0.1. Any individual that exhibits one or two minor alleles in at least 20 of the 60 SNPs is considered to have a high disease risk. The binary response $y_i \in \{0,1\}$ that indicates whether an individual is affected is generated with a penetrance of 0.9 (or 0.5 in a more difficult scenario) for high risk individuals, and a penetrance of 0.1 for the other individuals. After having generated an equal number of individuals for both subgroups, a random subset from all affected individuals in both groups as well as an equal number of unaffected individuals are drawn. We consider scenarios with a total of $n \in \{200,400\}$ individuals and a total of $q \in \{1000,2000\}$ SNPs.

While this basic design already introduces interaction patterns implicitly, as a certain number of SNPs with at least one minor allele is needed for being considered as an individual with high risk, we also introduce interactions explicitly. For this, a second group of SNPs, in addition to the 60 SNPs, is taken into account, comprising of only three SNPs (with minor allele frequency of 0.1, regardless of the subgroup). An individual then is only considered to be high risk if it has at least 20 SNPs with one or two minor alleles in the first group of 60 SNPs and at least one SNP with one or two minor alleles in the second group.

In an even more challenging design, not only one group of 60 SNPs (potentially in combination with a second group of three SNPs) is considered, but a further group of 60 SNPs is introduced (potentially also with a separated corresponding group of three SNPs). An individual will then have high disease risk if the above conditions are met for one of the two 60 SNP groups (and the corresponding three SNP group).

5.2 Results

Table 1 shows the misclassification rates of the CLR approach, global models and per-cluster intercept models in their application to the simulation scenarios. The per-cluster intercept models are outperformed by both the global models and the CLR approach in most of the scenarios. The only exceptions are the two scenarios with two (pairs of) groups and low penetrance, i.e., the most difficult scenarios.

Table 1: Mean misclassification rate (25% and 75% quantiles in parentheses) of per-cluster intercept models, global models, and the CLR models for simulation scenarios with different numbers *n* of observations and numbers *p* of SNPs. With ("yes") or without ("no") explicit interactions ("ei"), different numbers of (pairs of) groups ("gr"), and different penetrances ("pt").

n	*p*	ei	gr	pt	per-cluster intercept	global	cluster-localized
200	1000	no	1	0.5	45.1 (41.2, 50.0)	39.0 (36.8, 40.9)	39.5 (37.0, 41.3)
		yes	1	0.9	45.0 (40.9, 50.0)	34.5 (32.1, 35.9)	34.9 (31.9, 37.4)
400	1000	no	1	0.5	36.6 (34.7, 37.9)	35.7 (34.4, 36.6)	35.6 (34.1, 36.3)
			2	0.5	44.4 (42.5, 45.7)	45.4 (44.2, 46.7)	45.4 (43.8, 46.8)
		yes	1	0.5	41.8 (40.4, 43.0)	41.2 (39.8, 42.7)	40.8 (39.1, 42.3)
				0.9	36.8 (34.6, 38.1)	29.4 (27.8, 30.7)	27.7 (26.5, 29.4)
			2	0.5	44.9 (42.9, 46.2)	46.2 (44.5, 47.5)	45.7 (43.7, 47.0)
				0.9	40.8 (39.5, 42.1)	35.0 (33.6, 36.2)	35.1 (33.2, 36.7)
	2000	yes	1	0.5	48.1 (45.8, 50.0)	42.1 (40.1, 44.0)	42.2 (40.5, 43.8)
				0.9	46.3 (43.6, 50.0)	29.6 (28.2, 31.3)	30.1 (28.8, 31.9)

While the CLR approach performs similar to the global model in one of these two scenarios, it performs somewhat better in the second scenario.

For $n = 200$ individuals, the CLR approach is slightly outperformed by the global model, i.e. there seems to be a minimum level of information required for reliably identifying clusters and fitting local models. For $n = 400$ individuals, the CLR approach and the global model show comparable performances, where the former approach outperforms the global model in most of the scenarios with $p = 1000$. The global model is slightly better than the CLR approach when analyzing $p = 2000$ SNPs, which might be due to the more difficult task of identifying important SNPs from a larger overall number of SNPs. An advantage for the CLR approach can be seen in the scenarios with explicit interactions and one pair of SNP groups, where this advantage increases with increasing penetrance.

Besides prediction performance, the CLR approach is compared to global models concerning identification of important SNPs. In the present simulation scenarios, each SNP is considered important that is a member of the (pairs of) groups contributing to the high risk definition. Table 2 shows sensitivity and 1$-$specificity for these SNPs, when considering all non-zero coefficients of the global or the localized models as selected, when considering assignment contributions (average over (3)) beyond a certain cutoff, or the union of SNPs selected by localized models and according assignment contributions. The cutoff for the (average) assignment contribution is arbitrary, and therefore specified such that 50 SNPs are selected in each repetition.

Table 2: Mean sensitivity/1−specificity with respect to SNP identification by global models and cluster-localized models (according to fitted models, assignment contribution, or both) for simulation scenarios with different numbers *n* of observations and numbers *p* of SNPs. With ("yes") or without ("no") explicit interactions ("ei"), different numbers of (pairs of) groups ("gr"), and different penetrances ("pt").

n	*p*	ei	gr	pt	global	cluster-localized		
						model	assignment	combined
200	1000	no	1	0.5	0.217 / 0.014	0.284 / 0.029	0.042 / 0.050	0.318 / 0.079
		yes	1	0.9	0.247 / 0.017	0.304 / 0.034	0.042 / 0.051	0.334 / 0.083
400	1000	no	1	0.5	0.340 / 0.020	0.372 / 0.035	0.163 / 0.043	0.484 / 0.076
			2	0.5	0.079 / 0.012	0.112 / 0.025	0.091 / 0.044	0.194 / 0.068
		yes	1	0.5	0.204 / 0.020	0.229 / 0.037	0.189 / 0.041	0.377 / 0.076
				0.9	0.384 / 0.030	0.450 / 0.057	0.223 / 0.038	0.573 / 0.093
			2	0.5	0.082 / 0.014	0.145 / 0.032	0.086 / 0.045	0.222 / 0.076
				0.9	0.248 / 0.029	0.342 / 0.069	0.090 / 0.044	0.404 / 0.110
	2000	yes	1	0.5	0.169 / 0.011	0.220 / 0.020	0.039 / 0.025	0.260 / 0.046
				0.9	0.358 / 0.016	0.426 / 0.029	0.019 / 0.025	0.439 / 0.054

While the sensitivity for SNPs selected by the cluster-localized models is in general larger than the one of the global model, specificity is larger for the latter. The worse specificity for the cluster-localized models might have been expected, as a larger number of models is fitted. However, the increase in sensitivity is notable, as in the worst case the same SNPs could have been selected in each cluster, i.e. there would have been no increase in sensitivity. Therefore, it seems that the clusters were chosen in such a way that different important SNPs have been be selected in each cluster.

The absolute level of sensitivity/1−specificity for selection based on assignment contribution should not be interpreted, as it depends on an arbitrary cutoff. For the chosen cutoff, there are several scenarios in which selection according to assignment contribution is strictly dominated by selection according to the localized models, i.e. sensitivity as well as specificity is worse, e.g., $n = 400$, $p = 2000$, one (pair of) group(s), penetrance of 0.5). This might be due to the fact that the localized models were fitted by an approach that is explicitly designed for variable selection, while the assignment contributions are derived from the clustering approach that uses SNPs weights, but does not perform SNP selection.

When combining localized models and assignment contribution, i.e., considering SNPs selected by the former or beyond the cutoff for the latter, specificity considerably decreases. The values for 1−specificity are close to the sum of the values for the two separate selection schemes, which is due to the introduction of two separate sets of noise SNPs into the list of selected SNPs. By contrast, the combined selection shows a considerable increase in sensitivity when compared to the separate selections based on the localized models and the assignment contribution. Therefore, it seems that, besides some overlap, selection according to assignment contribution might be valuable, as it indicates sets of SNPs that could not be identified by the localized models.

6 Application example

6.1 Data description and preprocessing

For illustrating the CLR approach in a real example, we consider a case-control study concerned with urinary bladder cancer. The case cohort, which is described in detail by Golka et al. (2009), consists of 308 German patients genotyped using the Affymetrix Genome-Wide Human SNP Array 5.0. The genotypes are called by employing CRLMM (Corrected Robust Linear Model with Maximum-likelihood based distances; Carvalho et al., 2010), and by paying, in particular, attention to possible batch effects. The signal-to-noise ratio computed by CRLMM for each patient and the CRLMM quality score for each SNP are used to exclude 14 patients with a signal-to-noise ratio less than 5 and 13,679 SNPs with a quality score of less than 0.7, respectively, from further analyses. Furthermore, all SNPs with a call rate smaller than 95%, and all monomorphic SNPs are removed from the analysis, leading to a total of 294 patients and 392,582 SNPs. The control cohort, consisting of 936 German population controls from the biobanks KORA-gen (Wichmann et al., 2005) and PopGen (Krawczak et al., 2006), has been genotyped with the Affymetrix Genome-Wide Human SNP Array 6.0, and preprocessed by applying Birdseed, the standard Affymetrix algorithm for this chip type. From this cohort, the 620,711 SNPs are selected that exhibit a call rate larger than 95%, have a minor allele frequency larger than 1%, are in Hardy-Weinberg equilibrium, and show no considerable difference in the allele frequencies between the control populations (for details, see Steffens et al., 2010). These SNPs are matched with the 392,582 SNPs from the case cohort to identify 317,909 SNPs available for both cases and controls. Missing values are cohort-wise replaced using weighted k nearest neighbors (Schwender, 2012). To prevent that SNPs with genotyping errors dominate the results of our analysis, we repeatedly fit global logistic regression models by componentwise boosting, manually check the signal-intensity plots for the selected SNPs whether they suggest that the calling of the genotypes has not worked properly, and remove suspicious SNPs, until all selected SNPs are unsuspicious. For developing models on the resulting $q = 317741$ SNPs, $n = 1000$ individuals are randomly chosen as training set (comprising 226 cases and 774 controls), the remaining 230 individuals (68 cases and 162 controls) are considered for evaluation.

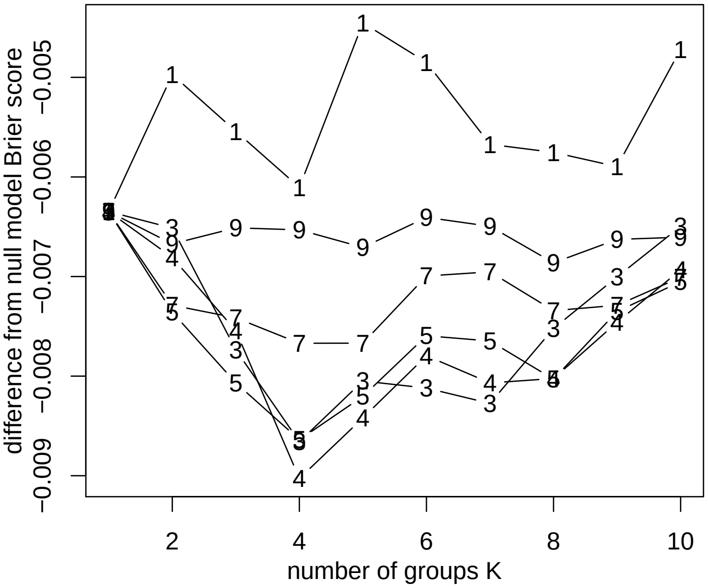

Figure 1: Brier score difference to the null model, i.e., Brier score of each model subtracted from that of an intercept model, as estimated from 10-fold cross-validation, for different numbers K of groups and values W of the weight parameter (as indicated by $W \cdot 10$ in the curve symbols).

6.2 Localized regression

For selecting the number K of groups, the relative weight W for individuals from other groups, and the number M of boosting steps, we use 10-fold cross-validation, as described in Section 4.1 Specifically, we consider up to $K = 10$ groups, and W is evaluated on a coarse grid.

Figure 1 shows the difference between the estimated Brier scores of the cluster-localized regression model and the intercept model for different numbers K of groups and values of W. By considering the Brier score difference, the inseparability component of the Brier score is removed, which is a property of the classification task at hand, leaving only the imprecision component. Thus, Figure 1 shows the gain in precision (see, e.g., Binder and Graf, 2009, for more details). The maximum gain in precision is obtained at $K = 4$ and $W = 0.4$. While values of $W \leq 0.1$ or $W \geq 0.9$ even result in inferior performance, compared to a global model ($K = 1$), all other values result in an improvement. $K = 4$ groups is optimal for most of the latter values, supporting the assumption of four distinct groups. The optimal number of boosting steps, corresponding to $K = 4, W = 0.4$, is $M = 69$. For

Table 3: Prediction performance of cluster-localized models ($K = 4$) compared to global models ($K = 1$) when considering intercept models (no covar) both globally ($K = 1$) and per-cluster ($K = 4$) and sparse risk prediction models, fitted by boosting, in their application to the urinary bladder cancer test data. For the misclassification rate (Misclass) and the area under the ROC curve (AUC) the average over all individuals in the test data is presented. For the Brier score, the average over all individuals (all) as well as the average in each of the four groups is shown.

Model		Criterion						
		Misclass	AUC	Brier score				
				All	Group			
					1	2	3	4
No covar	$K = 1$	0.296	-	0.213	0.214	0.287	0.234	0.126
	$K = 4$	0.296	0.655	0.198	0.217	0.246	0.234	0.118
Boosting	$K = 1$	0.296	0.644	0.203	0.207	0.269	0.231	0.117
	$K = 4$	0.291	0.726	0.196	0.209	0.254	0.223	0.113

the global model ($K = 1$), 10-fold cross-validation selects $M = 51$ boosting steps. Therefore, it seems that more information can be extracted when fitting cluster-localized models.

Using these tuning parameter values, we build models based on the training data and evaluate the resulting models by applying them to the 230 individuals from the test set. Besides the intercept model and the global model, per-cluster intercept models are considered as competitors.

Table 3 shows the prediction performance for the 230 individuals from the test set, evaluated via the Brier score, misclassification rates, and the area under the ROC curve (AUC). For calculating misclassification rates, a cutoff of 0.5 was used for obtaining predicted class memberships from predicted probabilities. For obtaining the AUC, this cutoff was varied, resulting in various levels of sensitivity and specificity, i.e., ROC curves.

Improved performance, in terms of the (overall) Brier score and the AUC, is obtained when moving from a global intercept model (corresponding to an AUC of 0.5) to separate intercept models for each group, as well as to a global logistic regression model incorporating SNP information. Therefore, clustering as well as sparse risk prediction models seem to be able to extract SNP information with respect to case-control status. When combining both by the CLR approach, prediction performance further increases. In terms of of misclassification rate, only the latter approach improves over the global intercept model.

For characterizing the groups, identified by the proposed CLR approach, we separately consider the Brier score for each of the $K = 4$ groups in Table 3. In

Group 1, both logistic regression approaches improve over the intercept-only models, i.e., prediction in this group particularly benefits from sparse logistic regression models. As this group by far comprises the most individuals in the training data (480 individuals), the SNPs important for this group probably dominate the global model. This might explain why the regression model for this group cannot improve prediction performance over the global model. For Group 2 (comprising 162 individuals), the global regression model outperforms the global intercept model, but is inferior to the group-specific intercept models. The CLR approach provides some improvement, but still is inferior to the intercept models. It seems that no small set of SNPs can be found that could improve prediction performance for this group. For Groups 3 and 4 (comprising 164 and 194 individuals in the training data, respectively), the global regression model modestly improves over the intercept models. For both groups, the CLR approach further improves performance. A particularly large gain is seen for Group 3 (from 0.231 for the global model to 0.223 for the local model). Therefore, we will focus on this group for further investigating the fitted models in the following section.

6.3 Extracting interactions

Figure 2 shows the contribution of all 317741 SNPs towards assigning individuals to Group 3 for the cluster-localized models, determined by computing (3) for the 230 individuals in the test set. There is considerable variability, making it difficult to determine important genomic regions. Nonetheless, there seem to be a few genomic regions that are more important for assigning individuals to Group 3, e.g., a small region on chromosome 7 with several large assignment contribution values.

Alternatively, Group 3 can be characterized by the SNPs having the same genotype for all individuals from the training set that belong to this group. As shown in Figure 2, the positions of these 47 SNPs are not equally distributed over the genome, which could potentially indicate some systematics. However, there is little agreement with the locations that might be considered important based on the assignment contribution. For example, none of the zero-variability SNPs is on chromosome 7, while there are six on chromosome 10. Therefore, if both quantities are believed to be informative, they seem to carry different information.

While 32 SNPs received non-zero estimated coefficients for at least one of the two corresponding covariates in the global model, there are 49 non-zero effect SNPs in the model for Group 3 (positions indicated in Figure 2 by crosses), where 19 of these SNPs are in both models. Thus, there are 30 non-zero SNPs in Group 3 not contained in the global model. Most of these (28 SNPs) are also not contained in any of the models for the other groups. The improved prediction performance in

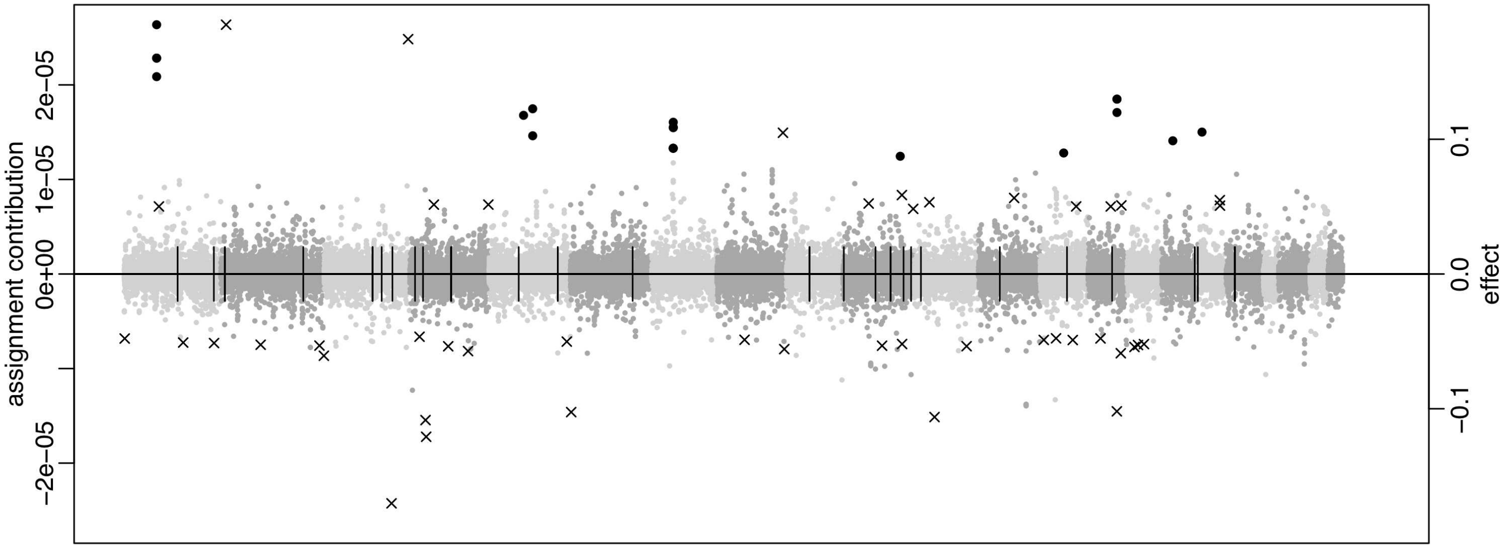

Figure 2: Contribution of SNPs towards assigning individuals from the test set to Group 3 for prediction (left axis, dots) and estimated coefficients in cluster-localized models (right axis, black crosses), for SNP covariates in the urinary bladder cancer data, ordered horizontally according to genomic location. Different chromosomes, which are sorted in increasing order, are indicated by alternating shades of grey. SNPs are indicated by black dots if their assignment contribution is above the threshold used for extracting candidate interactions ($1.2 \cdot 10^{-5}$). SNPs for which all individuals in the training set that belong to Group 3 have the same genotype, i.e., zero-variability SNPs, are indicated by vertical lines. SNPs with missing chromosome position annotation are not shown.

Group 3 (see Table 3) indicates that these SNPs might be truly important, but only for this certain group of individuals.

There are only four SNPs with non-zero effect in all groups. When the absolute sizes of coefficient estimates differ between groups, this corresponds to a typical statistical interaction in the presence of main effects. However, due to regularized estimation, the estimated coefficients will be highly biased towards zero. It is therefore difficult to distinguish between a pure main effect present in several groups and a combination of main and interaction effects, if a SNP receives a non-zero estimate in several groups. Therefore, we focus in the following on the 30 SNPs having an estimated non-zero effect only in Group 3 for identifying candidate interactions.

Due to linkage disequilibrium (LD), the different SNPs with non-zero effects in the different groups might be closely related. For investigating whether the CLR approach really identifies SNPs that only have an effect in Group 3 or whether a SNP in strong LD with one of these SNPs shows an effect in one of the other groups, we calculate r^2 as a measure of LD based on the 774 controls for each pair of SNPs, where the first SNP only has an effect in Group 3 and the second SNP only in one of the other groups. For these pairs, the maximum LD value is $r^2 = 0.467$, 0.013, and 0.038, for comparison of Group 3 to Group 1, Group 2, and Group 4, respectively. This indicates that a few of the non-zero effects in Group 3 might at most modestly overlap with those of Group 1, but that they are truly different from the non-zero effects in the other two groups. Considering the LD for pairs of SNPs that only have an effect within Group 3, the maximum value is $r^2 = 0.014$, i.e., componentwise boosting provides a sparse solution with non-zero effects for SNPs that are not in LD.

For identifying SNPs that are important for group assignment, we use a cutoff on the aggregated quantity (3) (dots in Figure 2), where the choice of such a cutoff is arbitrary by nature. In the present example, we consider SNPs with values larger or equal to $1.2 \cdot 10^{-5}$, resulting in 21 SNPs for Group 3. The maximum of the pairwise LD for these SNPs is $r^2 = 0.990$ (third quartile: 0.003), indicating that a few SNPs in strong LD are selected. As an alternative for identifying SNPs important for assignments to Group 3, we consider those SNPs for which all individuals, belonging to this group in the training set, have the same genotype, i.e., zero-variability SNPs. This results in 47 SNPs with maximum pairwise LD of $r^2 = 0.995$ (third quantile: 0.002), i.e., an overlap similar to the 21 SNPs above. Note that these are not SNPs with general zero-variability, but just zero-variability in this group.

The CLR approach does not distinguish between interaction patterns with several pairwise interactions and more complex interaction patterns. In the following, we concentrate on extracting pairwise interactions, and validate these in the test data. Having ascertained that at least the SNPs with a non-zero effect only in Group 3 are truly different between groups, we consider interactions between those SNPs and the SNPs that are important for group assignment, i.e., we investigate $30 \times 21 = 630$ interactions and $30 \times 47 = 1410$ interaction based on quantity (3) and zero-variability SNPs, respectively. For the SNPs with a non-zero effect, the binary indicator variables (4) are used that receive non-zero estimates in the regression model for Group 3.

Each of these SNP pairs, consisting of one SNP with a non-zero effect, and another SNP important for the group assignment, is tested by using a likelihood ratio test, similar to the one proposed by Cordell (2002) for testing epistatic interactions. Thus, a main effects model is fitted, in which the former SNP is incorporated via the indicator variable with non-zero effect, and the other SNP z_{il} via two dummy variables $x_{il}^{(a)}$ and $x_{il}^{(d)}$, coding for additive and dominant effects. Specifically, $x_{il}^{(a)} = 1$ and $x_{il}^{(d)} = -0.5$ for $z_{il} = 0$, $x_{il}^{(a)} = 0$ and $x_{il}^{(d)} = 0.5$ for $z_{il} = 1$, and $x_{il}^{(a)} = -1$ and $x_{il}^{(d)} = -0.5$ for $z_{il} = 2$. The log-likelihood of this main effects model is then compared with the log-likelihood of a model that additionally contains the interaction terms between the indicator variable and all dummy variables, thus, resulting in a likelihood ratio test.

The likelihood ratio tests are performed using the 230 individuals from the test set to investigate whether the interactions, identified by the CLR approach applied to the 1000 individuals from the training set, can be recovered. To address multiple testing, the false discovery rate is controlled by the approach of Benjamini and Hochberg (1995). While the individual likelihood ratio tests will be dependent, this approach is still valid, as positive regression dependency can be assumed (Benjamini and Yekutieli, 2001).

As the number of individuals in the test set is rather small, there is little power to recover the interactions. Therefore, a rather liberal level for the false discovery rate might be justified. Starting from the 630 candidate interactions based on the aggregated quantity (3), there are five interactions identified at an FDR of 0.127. All other interactions are above 0.5. When considering the 1410 candidate interactions based on the zero-variability SNPs, there are three interactions identified at FDR=0.118, and five further interactions with an FDR of about 0.2. All other FDRs are above 0.3. There is no overlap between the five interactions identified using (3) and the eight interactions identified using zero-variability SNPs. The total set of 13 candidate interactions would have also been selected, if the two sets of interaction p-values would have been considered jointly at a FDR of 0.2. The maximum pairwise LD in any of these 13 candidate interactions is $r^2 = 0.002$, i.e., these are truly interactions and not just artifacts due to LD.

The 13 interactions encompass 18 unique SNPs. For checking whether these SNPs would also have been found based on their marginal effects, we consider χ^2-tests based on the 1000 individuals. For the 13 SNPs that are important for group assignment, all resulting (unadjusted) *p*-values are larger than 0.10, i.e., none of these would have been considered significant. For the five SNPs with non-zero effects, the *p*-values are all smaller than 0.006. After an adjustment for multiple testing, none of these SNPs would have been considered important based on marginal effects. However, based on the results from cluster-localized regression, all these SNPs should be considered for further validation of interactions.

7 Discussion

When analyzing SNP data from case-control studies, regularized multivariable regression models allow for considering all SNPs simultaneously. Regularized techniques, such as the lasso or componentwise boosting, provide sparse model fits, i.e., select a small set of SNPs that can be used to predict the risk for new individuals. However, most techniques that can deal with a huge number of SNPs do not consider interactions, and approaches that can deal with interactions in a multivariable model are often limited to a smaller number of SNPs.

We addressed the above problems by introducing a cluster-localized regression (CLR) approach for fitting risk prediction models. Groups of patients are automatically identified by a clustering approach that assigns different weights to different SNPs in each group. This allows for characterizing groups via the SNPs with the largest weights when assigning new individuals to this group. For each group, weighted componentwise likelihood-based boosting is used for fitting a sparse logistic regression model, which, hence, provides a selection of SNPs predictive for specific groups.

For evaluating the CLR approach, we considered several simulation scenarios in which the case-control status depends on a combination of one or more groups of SNPs. An advantage in terms of prediction performance over a main effects approach, similar to the one proposed by Wu et al. (2009), or per-cluster intercept models was seen in a subset of the scenarios. Improved sensitivity was obtained with respect to identifying important SNPs, however at the cost of specificity. Both the set of SNPs selected for the per-cluster models and SNPs selected according to assignment contribution helped in SNP identification.

The CLR approach was, furthermore, illustrated in its application to case-control data from a urinary bladder cancer study. The localized models for each group provided improved prediction performance over a global model. This lends credibility to the group structure uncovered by the CLR approach. Specifically, the group in which the most improvement in prediction performance was seen will be considered for further validation steps, as it might form the basis for individualized risk prediction.

Based on the group structure, also a list of candidate interactions could be identified, using the SNPs with non-zero effects only in one of the groups and SNPs important for assigning individuals or having zero-variability in this group. Pairwise investigation of these SNPs showed that they are not in linkage disequilibrium, i.e., they are candidates for real interactions. Some of these interactions could be recovered in a test data set, indicating validity of the identified structure. The interactions that could be recovered in the test data would not have been found based on marginal testing. Therefore, the CLR approach is more capable in identifying interactions than two-step approaches based on marginal effect screening in a first step.

Naturally, the CLR approach will also not be able to detect all effects, given a limited number of individuals. Specifically, only strong effects can be detected by componentwise boosting approach, while intermediate and weak effects will probably be missed in a sparse regression model. Still, the identified SNPs might provide an interesting, albeit incomplete glimpse of underlying biological processes.

While pairwise interactions could be extracted, the CLR approach does not distinguish between pairs of interacting SNPs and more complex interaction patterns comprising several SNPs. Ideally, each separate interaction pattern would receive its own group. However, often only a small number of groups will be chosen, due to small sample size, effectively collapsing several distinct interaction patterns into one group. This problem might be tackled either by increasing sample sizes, or by more sophisticated post-processing for disentangling the interaction patterns. Suitable approaches that can identify complex patterns from a moderate number of candidate SNPs are, e.g., logicFS (Schwender and Ickstadt, 2008), BEAM (Zhang and Liu, 2007), or MECPM (Miller et al., 2009).

We considered two quantities for determining the importance of specific SNPs for group assignment, but both exhibited considerable variability. While some genomic locations might have strong contributions, others might have been overlooked. Furthermore, similar to the lasso, the covariates selected by componentwise boosting are known to be subject to considerable variability. Therefore, uncertainty in the identification of interactions cannot be neglected. Some advanced tools for quantifying the uncertainty in this process would be useful, e.g., based on logic regression testing procedures and importance measures (Schwender et al., 2011). Variability of the results could potentially be decreased by incorporating information on genomic location into clustering as well as into the boosting algorithm. For example, information on haplotype blocks could be considered similar to pathway information (Binder and Schumacher, 2009).

However, the CLR approach should not be judged solely on its success in identifying potential interactions. As already indicated, having risk prediction models that are local to specific groups might be useful in its own right. The identified groups might, e.g., correspond to distinct patient groups requiring a focus on different molecular processes, as indicated by the SNPs selected by the boosting approach. Availability of such a direct interpretation is in contrast to approaches such as random forests that might result in good prediction performance, but do not provide single fitted models that can be interpreted with respect to an individual patient. Thus being able to adapt risk prediction models, based on the group a patient belongs to, might even provide a valuable starting point for personalized medicine.

References

Ayers, K. L. and Cordell, H. J. (2010). SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genetic Epidemiology*, 34(8):879–891.

Bansal, V., Libiger, O., Torkamani, A., and Schork, N. J. (2010). Statistical analysis strategies for association studies involving rare variants. *Nature Reviews Genetics*, 11(11):773–785.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1):289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188.

Binder, H. and Graf, E. (2009). Brier scores. In Kattan, M. W., editor, *Encyclopedia of Medical Decision Making*, pages 101–104. SAGE Publications.

Binder, H. and Schumacher, M. (2008). Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics*, 9:14.

Binder, H. and Schumacher, M. (2009). Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. *BMC Bioinformatics*, 10:18.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Carvalho, B. S., Louis, T. A., and Irizarry, R. A. (2010). Quantifying uncertainty in genotype calls. *Bioinformatics*, 26(2):242.

Chan, E. Y., Ching, W. K., Ng, M. K., and Huang, J. Z. (2004). An optimization algorithm for clustering using weighted dissimilarity measures. *Pattern Recognition*, 37(5):943–952.

Cordell, H. J. (2002). Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463–2468.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

Evans, D. M., Visscher, P. M., and Wray, N. R. (2009). Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Human Molecular Genetics*, 18(18):3525–3531.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, New York, 2nd edition.

Friedman, J. H. and Meulman, J. J. (2004). Clustering objects on subsets of attributes. *Journal of the Royal Statistical Society B*, 66(4):815–849.

Gerds, T. A., Cai, T., and Schumacher, M. (2008). The performance of risk prediction models. *Biometrical Journal*, 50(4):457–479.

Golka, K., Hermes, M., Selinski, S., Blaszkewicz, M., Bolt, H. M., Roth, G., Dietrich, H., Prager, H.-M., Ickstadt, K., and Hengstler, J. G. (2009). Susceptibility to urinary bladder cancer: Relevance of rs9642880[T], GSTM1 0/0 and occupational exposure. *Pharmacogenetics and Genomics*, 19(11):903–906.

Greene, C. S., Sinnott-Armstrong, N. A., Himmelstein, D. S., Park, P. J., Moore, J. H., and Harris, B. T. (2010). Multifactor dimensionality reduction for graphics processing units enables genome-wide testing of epistasis in sporadic ALS. *Bioinformatics*, 26(5):694–695.

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

Kayano, M., Takigawa, I., Shiga, M., Tsuda, K., and Mamitsuka, H. (2009). Efficiently finding genome-wide three-way gene interactions from transcript- and genotype-data. *Bioinformatics*, 25(21):2735–2743.

Krawczak, M., Nikolaus, S., von Eberstein, H., Croucher, P. J., El Mokhtari, N. E., and Schreiber, S. (2006). PopGen: Population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. *Community Genetics*, 9:55–61.

Liu, Y., Li, M., Cheung, Y. M., Sham, P. C., and Ng, M. K. (2010). SKM-SNP: SNP markers detection method. *Journal of Biomedical Informatics*, 43(2):233–239.

Loader, C. (1999). *Local Regression and Likelihood*. Springer, New York.

Miller, D. J., Zhang, Y., Yu, G., Liu, Y., Chen, L., Langefeld, C. D., Herrington, D., and Wang, Y. (2009). An algorithm for learning maximum entropy probability models of disease risk that efficiently searches and sparingly encodes multilocus genomic interactions. *Bioinformatics*, 25(19):2478–2485.

Müller, T., Schiffner, J., Schwender, H., Szepannek, G., Weihs, C., and Ickstadt, K. (2010). Local analysis of SNP data. In Locarek-Junge, H. and Weihs, C., editors, *Classification as a Tool for Research*, Berlin. Springer.

Ng, M. K., Li, M. J., Ao, S. I., Sham, P. C., Cheung, Y.-M. . M., and Huang, J. Z. (2006). Clustering of SNP data with application to genomics. In *ICDMW '06: Proceedings of the Sixth IEEE International Conference on Data Mining - Workshops*, pages 158–162, Washington, DC, USA. IEEE Computer Society.

Park, M. Y. and Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics*, 9(1):30–50.

Phillips, P. C. (2008). Epistasis – the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9(11):855–867.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ruczinski, I., Kooperberg, C., and LeBlanc, M. (2003). Logic regression. *Journal of Computational and Graphical Statistics*, 12(3):475–511.

Schwarz, D. F., König, I. R., and Ziegler, A. (2010). On safari to random jungle: A fast implementation of random forests for high dimensional data. *Bioinformatics*, 26:1752–1758.

Schwender, H. (2012). Imputing missing genotypes with weighted k nearest neighbors. *Journal of Toxicology and Environmental Health, Part A.*

Schwender, H. and Ickstadt, K. (2008). Identification of SNP interactions using logic regression. *Biostatistics*, 9(1):187–198.

Schwender, H., Ruczinski, I., and Ickstadt, K. (2011). Testing SNPs and sets of SNPs for importance in association studies. *Biostatistics*, 12(1):18–32.

Silver, M., Montana, G., and Initiative, A. D. N. (2012). Fast identification of biological pathways associated with a quantitative trait using group lasso with overlaps. *Statistical Applications in Genetics and Molecular Biology*, 11(1):Article 7.

Steffens, M., Becker, T., Sander, T., Fimmers, R., Herold, C., Holler, D. A., Leu, C., Herms, S., Cichon, S., Bohn, B., Gerstner, T., Griebel, M., Nöthen, M. M., Wienker, T. F., and Baur, M. P. (2010). Feasible and successful: Genome-wide interaction analysis involving all 1.9e+11 pair-wise interaction tests. *Human Heredity*, 69(4):268–284.

Tutz, G. and Binder, H. (2005). Localized classification. *Statistics and Computing*, 15(3):155–166.

Tutz, G. and Binder, H. (2007). Boosting ridge regression. *Computational Statistics & Data Analysis*, 51(12):6044–6059.

VanderWeele, T. J. (2010). Epistatic interactions. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article 1.

Wichmann, H. E., Gieger, C., and Illig, T. (2005). KORA-gen – resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen*, 67, Suppl. 1:155–166.

Winham, S., Wang, C., and Motsinger-Reif, A. A. (2011). A comparison of multifactor dimensionality reduction and L_1-penalized regression to identify gene-gene interactions in genetic association studies. *Statistical Applications in Genetics and Molecular Biology*, 10(1):Article 4.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721.

Yang, Can, Wan, Xiang, Yang, Qiang, Xue, Hong, Yu, Weichuan, Yang, C., Wan, X., Yang, Q., Xue, H., and Yu, W. (2010). Identifying main effects and epistatic interactions from large-scale SNP data via adaptive group lasso. *BMC Bioinformatics*, 11(Suppl 1):S18.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B*, 68(1):49–67.

Zhang, Y. and Liu, J. S. (2007). Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics*, 39(9):1167–1173.

Ziegler, A., König, I. R., and Thompson, J. R. (2008). Biostatistical aspects of genome-wide association studies. *Biometrical Journal*, 50(1):8–28.