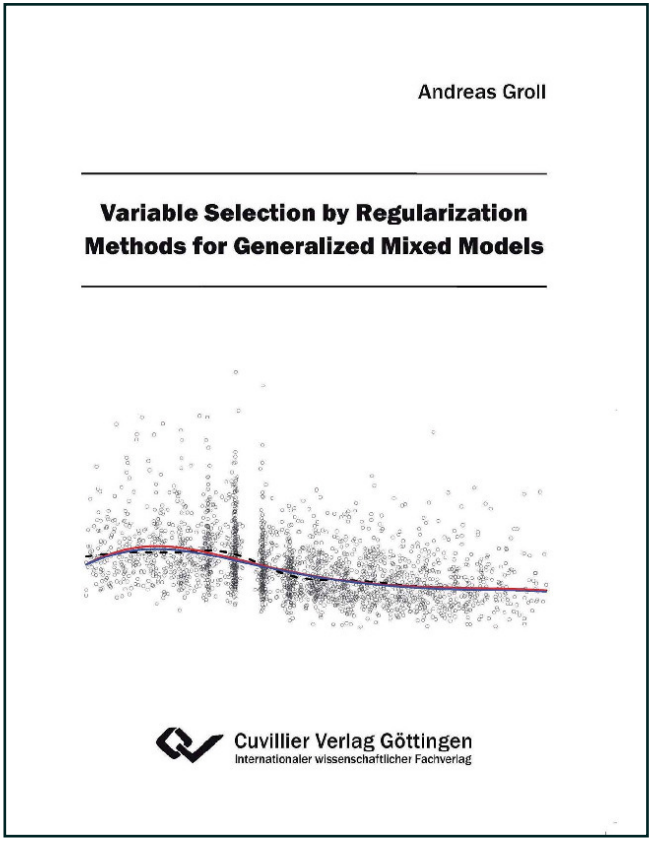

1. Introduction

Some Basic Concepts of Regularization

Generalized mixed models are widely used to model correlated and clustered responses. For example, the dependence structure of longitudinal data and of designs with repeated measurements can be captured. Due to heavy computational problems in the estimation of parameters modeling usually is restricted to a moderate number of predictor variables. When many predictors are available, estimates often become very unstable. Therefore, procedures to select the relevant variables are very important.

A classical approach to the selection of predictors is *subset selection*, which is based on test statistics with the usual stability problems of forward-backward algorithms, which are due to the inherent discreteness of the method (see for example Breiman, 1996b).

Boosting

A different and more timely approach to variable selection that has been developed in the machine learning community is based on *boosting methods*. According to Hastie et al. (2009), boosting is one of the most powerful learning ideas introduced in the last 20 years. Though it was originally designed for classification problems, it can be also applied to regression. An extensive and enlightening overview on recent boosting algorithms can be found in Bühlmann and Hothorn (2007). The general concept of converting a weak learning algorithm into one that achieves arbitrarily high accuracy has been developed by Schapire (1990). Thereby a "weak learner" characterizes a classification method that performs only slightly better than random guessing. This concept can be seen as the break through for several new methods, so-called *ensemble schemes*, which rely on the principle of generating repeated predictions by reweighting or resampling the original data set and finally averaging among the individual classifiers. Examples include *bagging* (Breiman, 1996a) or *random forests* (see for example Breiman, 2001). With the emergence of ensemble schemes also the most well known boosting algorithm has been developed, namely the AdaBoost algorithm for binary classification (Freund and Schapire, 1996, 1997), which uses a suitable base procedure[1] as classifier, such as, for example, a classification and regression tree (CART, Breiman et al., 1984).

It was not until Breiman (1998) found the decomposition of the prediction error of a classifier into bias and variance, that the success of AdaBoost could be satisfyingly explained, namely that it is able to reduce both bias and variance of a base procedure. Another important aspect of boosting concerns the optimal number of boosting steps. Contrary to the initial assumption, that AdaBoost is immune to overfitting, it is clear nowadays that boosting algorithms eventually overfit and, thus, the optimal number of iterations represents a tuning parameter which needs to be determined in some data-driven way, for example using some cross-validation scheme.

The next important step in the history of boosting was the new finding, that the AdaBoost algorithm can be represented as a functional gradient descent algorithm (Breiman,

[1] Note, that "base procedure" and "weak learner" are often used as equivalent terms in the boosting literature.

1998, 1999). This inspired Friedman et al. (2000) to transfer the idea of boosting to logistic regression, by modifying the AdaBoost algorithm, using a base procedure that returns class probabilities instead of labels. Friedman (2001) further improved the concept of boosting as a gradient descent optimization technique and extended boosting methods to include regression problems. He suggests to minimize the empirical version of the expected loss for some specified loss function, using a steepest gradient descent approach. An overview of robust loss functions for regression can be found in Hastie et al. (2009). Moreover, Friedman (2001) has demonstrated, that a small step-size factor in the boosting update, denoted by ν, can be often beneficial and almost never yields substantially worse predictive performance of boosting estimates. In Bühlmann and Yu (2003) the L_2-loss has been investigated. They showed how to fit smoothing splines by boosting base learners and introduced the idea of componentwise boosting, which may be exploited to select predictors. Furthermore, they succeeded in proving an exponential dependence between the bias and the variance of the boosted model, which explains to a certain extent that boosting algorithms are rather resistant against overfitting. These findings represent some of the most important results concerning theoretical properties of boosting algorithms.

Another form of boosting is likelihood-based boosting, which may be seen as an extension of boosting based on the L_2-loss (L_2Boosting). In case of the logit model and binomial likelihood Friedman et al. (2000) have proposed the LogitBoost algorithm. The more general case of semiparametrically structured regression in the form of additive models is considered in Tutz and Binder (2006), where all kinds of link functions and distributions that are used in generalized additive models are covered and also variable selection is achieved by using componentwise learners. Furthermore, penalized regression splines as well as penalized stumps are considered as weak learners. As the so-called GAMBoost algorithm with penalized regression splines as learners is of fundamental character for the boosting approaches presented in this thesis, we give some more details about the algorithm. In each boosting iteration the procedure uses a single step in Fisher scoring[2] for the update, based on a single smooth component. Thus, it is necessary to decide which of the available predictor variables should be used for the update. The straightforward criterion that is proposed here is to link the choice of the variable to the improvement of fit by one Fisher scoring step, which for likelihood-based models is given by the deviance. The componentwise update has the advantage that a selection of variables is implicitly performed by fitting simple models, containing only one predictor variable. For appropriate stopping of the algorithm a suitable information criterion is used, which specifies the trade-off between model complexity and goodness-of-fit. It represents an attractive alternative to cross-validation, which especially for larger data sets becomes very time consuming. The model complexity at boosting iteration l is given by the trace of the corresponding approximate hat matrix H_l, which is defined in a recursive manner, consisting of all hat matrices corresponding to earlier iterations.

[2]The Fisher scoring algorithm represents a variant of the Newton-Raphson method. Both methods coincide in exponential family model settings, if the canonical link function is used.

A componentwise likelihood-based boosting procedure for additive mixed models, called BoostMixed, has been proposed by Tutz and Reithinger (2007), which incorporates random effects. The procedure is very similar to the approach of Tutz and Binder (2006), with the main difference, that selection is based on information criteria, in contrast to measures of deviance, thereby exhibiting superior performance in simulations. Thus, the selection of components is performed in a way that minimizes the new lack-of-fit, including the augmented complexity. Furthermore, an additional step for the computation of the random effects variance components is included in the procedure.

The boosting methods presented in Chapters 2 - 4 of this thesis are based on the methodology of the latter two mentioned approaches, GAMBoost and BoostMixed, and extend the concept of componentwise likelihood-based boosting to generalized linear and additive mixed models.

Penalization

A different approach to variable selection in linear models that has received much attention is based on *penalized regression* techniques. The so-called *Lasso* (Tibshirani, 1996) can be seen as the break through of a new technique in regression that uses an L_1-penalty on the regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$. The basic idea is to maximize the log-likelihood $l(\boldsymbol{\beta})$ of a regression model while constraining the L_1-norm of the parameter vector $\boldsymbol{\beta}$, usually excluding the intercept. This has the effect that all coefficients are shrunken towards zero and some can be set exactly to zero. If some coefficients are set to zero, the corresponding covariates have no effect on the dependent variable and a sparser model is obtained.

The Lasso estimate $\hat{\boldsymbol{\beta}}$ can be defined as the solution of the following *constrained* likelihood optimization problem:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\ l(\boldsymbol{\beta}), \qquad \text{subject to } ||\boldsymbol{\beta}||_1 \leq s, \tag{1.1}$$

with $s \geq 0$ and with $||\cdot||_1$ denoting the L_1-norm. Equivalently the Lasso estimate can be derived by solving the *penalized* likelihood optimization problem

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\ \left(l(\boldsymbol{\beta}) - \lambda||\boldsymbol{\beta}||_1\right), \tag{1.2}$$

with penalty parameter $\lambda \geq 0$. The latter representation explains why this technique is called *penalized regression*. Note, that both s and λ are tuning parameters of the procedure and have to be determined, for example, by cross-validation. This can be very time-consuming, especially in high-dimensional data settings. Thus, in problems with intricate log-likelihood function, efficient algorithms are needed to derive solutions of equations (1.1) or (1.2) in order to keep computation time practical.

In Section 5.1 several approaches for efficient solutions of equations (1.1) or (1.2) are discussed. Furthermore, an overview of the manifold extensions and advancements of the Lasso is given, whereas the use of penalization techniques for the selection of variables in mixed models is still in the beginning. Based on the approach of Goeman (2010),

in Chapter 5 suitable L_1-penalty approaches for the generalized linear mixed model are developed, which work by combining gradient ascent optimization with the Fisher scoring algorithm.

At this point it should be mentioned, that, following Bühlmann and Hothorn (2007), boosting based on the L_2-loss and Lasso are not equivalent in general, but can be seen as "related". Hastie et al. (2009) have been the first to draw an astounding connection between L_2Boosting with componentwise least squares and the Lasso. Next, Efron et al. (2004) further concretized this connection. They showed, that their version of L_2Boosting (the so-called forward stagewise linear regression, FSLR) with infinitesimal small step-sizes ν produces a set of solutions, which is approximately equivalent to the set of Lasso solutions when varying the Lasso penalty parameter λ from equation (1.2). This was achieved by representing both FSLR and Lasso as two different modifications of their computationally efficient least angle regression (LARS) algorithm (for generalized linear models, see Park and Hastie, 2007). Hence, boosting and penalization represent related alternatives for attaining regularized regression models. Both approaches are employed in the regularization methods presented in this thesis.

Guideline through the Thesis

The main part of this thesis consists of four basic chapters, which show the favorable qualities of regularization methods in different types of generalized mixed models. All of the boosting techniques proposed in Chapters 2 - 4 are based on the likelihood function and work by iterative, componentwise fitting of the residuals using weak learners. In order to keep individual chapters selfcontained, some passages repeat themselves with only small modifications and adjustments due to the different frameworks. Additionally, in Chapter 5 a completely different approach is presented, which enforces variable selection and shrinkage simultaneously by penalization of the L_1-norm of the regression parameters. In short, the single chapters may be summarized as follows:

Chapter 2: Boosting Approaches to Generalized Linear Mixed Models

Generalized linear mixed models are suitable for modeling the dependence structure of longitudinal data and of designs with repeated measurements. However, their use is typically restricted to few covariates, because the presence of many predictors yields unstable estimates. We propose a componentwise likelihood-based boosting approach which can be used in high-dimensional settings when many potentially influential explanatory variables are present. It allows fitting generalized linear mixed models for many covariates with implicit selection of relevant variables. For the determination of the complexity of the resulting estimator we use information criteria. Moreover, we can incorporate "random slopes" on linear effects, resulting in flexible generalized linear mixed models which are appropriate in cases where a simple random intercept

is unable to capture the entire variation of effects across subjects. The method is investigated both in extensive simulation studies and in an application to a real data example.

Chapter 3: Boosting Approaches to Ordinal Random Effects Models

A componentwise likelihood-based boosting approach for the fitting of binary and ordinal mixed models is presented. It is based on the same principle as the boosting approach from Chapter 2, extending it to generalized linear mixed models with ordinal response variables. Again, the technique works well, even if a large number of potentially influential explanatory variables is available and conventional approaches tend to fail. The method is investigated in simulation studies both for cumulative and sequential models and is illustrated in three applications.

Chapter 4: Boosting Approaches to Generalized Additive Mixed Models

With the emergence of semi- and nonparametric regression the generalized linear mixed model has been expanded to account for additive predictors. In this chapter the concept of likelihood-based boosting is extended to generalized additive mixed models. The procedure is constructed as a componentwise boosting method and hence is able to perform variable selection, with the selection being restricted to additive predictors. In contrast to common procedures it can be used in high-dimensional settings where many covariates are available, with unknown and potentially nonlinear influence. The complexity of the resulting estimator is determined by information criteria. Simulation studies for binary and Poisson responses as well as real data set examples shall demonstrate the properties of the suggested approach.

Chapter 5: L_1-Penalized Generalized Linear Mixed Models

As already mentioned, though generalized linear mixed models are a widely used tool for modeling longitudinal data, their use is typically restricted to few covariates, because the presence of many predictors yields unstable estimates. The presented approach to the fitting of generalized linear mixed models includes an L_1-penalty term that enforces variable selection and shrinkage simultaneously. A gradient ascent algorithm is proposed that allows to maximize the penalized log-likelihood yielding models with reduced complexity. In contrast to common procedures it can be used in high-dimensional settings where a large number of potentially influential explanatory variables is available. For categorical predictors the method enforces simultaneous selection of all the dummies that are linked to the categorical predictor. The method is investigated in simulation studies and illustrated by use of several real data sets.

Software

For all computations in the thesis the statistical programm R (R Development Core Team, 2008 – 2011, depending on the time when the respective research was done) was used

in combination with related packages. R-functions for the different boosting approaches from Chapters 2 - 4 are implemented in the R add-on package `GMMBoost` (Groll, 2011b) and are publicly available via CRAN (see `http://www.r-project.org`). The `glmmLasso` function for the fitting of generalized linear mixed models using L_1-penalization (Chapter 5) is implemented in the correspondent R add-on package `glmmLasso` (Groll, 2011a) and is also publicly available via CRAN.