1 Introduction

Feature Extraction and Structured Predictors

When a statistical model is to be chosen, variable selection is usually an important task. In this dissertation, however, not only variable selection but *feature extraction* is investigated. Feature extraction, as the term is used in this thesis, goes beyond variable selection in the sense that not only variables are selected but features which depend on the special nature of the investigated data. Sometimes (also in this thesis), the word feature *selection* is used instead of feature extraction.

In particular, variables with a special structure are considered and used as predictors in regression and classification problems. High-dimensional *signal-like metric* covariates are one type of such 'structured predictors'. In this case, we are typically faced with (more or less smooth) functional predictors, which, however, can only be observed at (a high number of) distinct measurement points. Thus, functional data are given as hundreds/thouthands of (ordered) metric variables; but actually they are realizations of functions. In signal regression, where such curves (i.e., signals) are used as regressors, feature extraction can be defined as the 'identification of relevant parts of the signal', where *relevant* means *relevant with respect to the response* which is to be explained/predicted.

The term 'feature extraction' is also often found in mass spectrometry-based proteomics, where spectra of protein intensities are analyzed and, for example, used to predict clinical outcomes. Proteins and peptides can be characterized by their individual mass to charge (m/z) ratios, and in mass spectrometry, only those m/z ratios can be observed. So observed spectra arise from intensities of proteins and peptides which are defined by and ordered according to related m/z values. These spectra can be seen as quite spiky 'signals', and feature extraction means to select the relevant mass/charge ratios.

Another very interesting type of structured regressors are *categorical* covariates, which are usually dummy-coded and hence result in groups of dummy variables. In this case, seemingly simple variable selection means groupwise selection.

Very common in statistical analysis are *ordinally scaled* categorical predictors. A quite important question is how to incorporate the variables' ordinal structure into statistical modeling. Besides variable selection, the identification of relevant differences between categories – of both ordinal and nominal predictors – is an important aspect, too.

Aims, Scopes and Main Results

In the previous section, two types of structured predictors have been described, signal-like metric predictors and categorical covariates. Other examples could have been given, too – for instance, expression profiles of genes belonging to the same pathway. Though incorporating the latter type of structure into statistical analysis is also sketched at the end of this thesis, the focus is on signal-like metric and categorical predictors. It is aimed at developing new methods for feature extraction given such data. The proposed techniques are Boosting procedures and/or penalized likelihood approaches. Bayesian methods are not considered (with a few exceptions). The last chapter of the thesis, however, provides

some ideas about nonparametric feature extraction. These methods can also be applied to 'standard' data without a special structure. The main topics and results of this dissertation can be summarized as follows:

- A new Boosting procedure for feature extraction in signal regression and mass spectrometry-based proteomics is proposed. Simulation studies and real world data applications show that the presented technique is a highly competitive alternative to existing approaches.

- Fitting methods for regression models are proposed which are especially suited for ordinal predictors – with or without variable selection. The usefulness of the introduced methods is illustrated, for example, by analyzing new data – the ICF Core Sets for chronic widespread pain.

- Besides variable selection, the identification of relevant differences between categories of both ordinal and nominal covariates is considered, and appropriate L_1-type regularization techniques for supervised clustering of categories are presented. The methods are investigated from a practical and a theoretical point of view. It is shown that the proposed procedures perform quite well, especially in comparison with existing 'standard' approaches.

- Finally, a new nonparametric method for feature extraction is introduced: the nearest neighbor ensemble. The performance of the proposed technique is quite encouraging.

Guideline through the Thesis

The main part of this dissertation consists of three chapters, which deal with different aspects of structured predictors and feature extraction. The single chapters can be read independently of each other. The same applies (with a few exceptions) to main sections within chapters. Within each chapter and main section a separate introduction is found for better orientation.

In Chapter 2, we deal with ordered metric covariates, and present a Boosting technique that is able to select subsets of adjacent variables. Typical applications come from signal regression where functional predictors are observed at a large number of adjacent measurement points (Section 2.2) and feature extraction in mass spectrometry-based predictive proteomics (Section 2.3). In Section 2.2 we deal with (signal) regression problems with metric (approx. normal) response, in Section 2.3 binary classification problems are considered.

In Chapter 3, which is the largest and most important part of this thesis, we consider categorical predictors. At first, we present approaches for smooth modeling of ordinal predictors, in its generic form (Section 3.2) and in combination with variable selection (Section

3.3). In Section 3.4 we propose regularization techniques for sparse parameterizations of categorical – nominal and/or ordinal – independent variables. In this context, sparse parameterizations do not only result from variable selection but also from fusion of categories of covariates. Categorical effect modifiers (in varying-coefficient models) are treated in Section 3.5. Since most of the described methods are introduced within the classical linear model, we show in Section 3.6 how the proposed approaches can be generalized to clearly non-normal response distributions as, for example, (binary) classification problems.

In Chapter 4, some approaches for nonparametric feature selection using nearest neighbor methods are presented. We (shortly) investigate the issue of standardization when nearest neighbor methods are applied (Section 4.2), and introduce a new type of nearest neighbor ensemble (Section 4.3). Since nearest neighbors are mostly used for discriminant analysis, the focus of Chapter 4 is on classification problems.

Publications

Parts of this thesis are based on research which has also been published in peer reviewed journals or as technical reports, and has been done in cooperation with coauthors. Large parts of Chapter 2 are also found in

- Tutz, G. and **J. Gertheiss** (2010). Feature extraction in signal regression: A boosting technique for functional data regression. *Journal of Computational and Graphical Statistics 19*, 154–174. (Section 2.2)

- **Gertheiss, J.** and G. Tutz (2009). Supervised feature selection in mass spectrometry-based proteomic profiling by blockwise boosting. *Bioinformatics 25*, 1076–1077. (Section 2.3)

Chapter 3 contains work from

- **Gertheiss, J.** and G. Tutz (2009). Penalized regression with ordinal predictors. *International Statistical Review 77*, 345–365. (Section 3.2)

- **Gertheiss, J.**, S. Hogger, C. Oberhauser, and G. Tutz (2011). Selection of ordinally scaled independent variables with applications to ICF Core Sets. (to appear in the) *Journal of the Royal Statistical Society C (Applied Statistics).* (Section 3.3)

- **Gertheiss, J.** and G. Tutz (2010). Sparse modeling of categorial explanatory variables. (to appear in) *The Annals of Applied Statistics.* (Section 3.4)

- **Gertheiss, J.** and G. Tutz (2010). Regularization and model selection with categorial effect modifiers. (revised/submitted version of) Technical Report 73, Department of Statistics, Ludwig-Maximilians-Universität München. (Section 3.5)

And Chapter 4 is mainly based on

- **Gertheiss, J.** and G. Tutz (2009). Variable scaling and nearest neighbor methods. *Journal of Chemometrics 23*, 149–151. (Section 4.2)
- **Gertheiss, J.** and G. Tutz (2009). Feature selection and weighting by nearest neighbor ensembles. *Chemometrics and Intelligent Laboratory Systems 99*, 30–38. (Section 4.3)

Software

All computations were carried out using the statistical programm R (R Development Core Team, 2007 – 2010, depending on the time when the respective research was done), and related packages (which are indicated in the respective chapters/sections). R-functions for blockwise Boosting as presented in Chapter 2 are available at `http://www.statistik.lmu.de/~gertheiss/research.html`. Functions for selecting and/or smoothing ordinal predictors using a Group Lasso or generalized Ridge penalty (Sections 3.2 and 3.3) are implemented in the R add-on package `ordPens` (Gertheiss, 2010), which will be made publicly available via CRAN (see `http://www.r-project.org`); a test version of the package can be downloaded from `http://www.statistik.lmu.de/~gertheiss/research.html`. An R package for sparse modeling of categorical explanatory variables (Section 3.4) is in preparation.