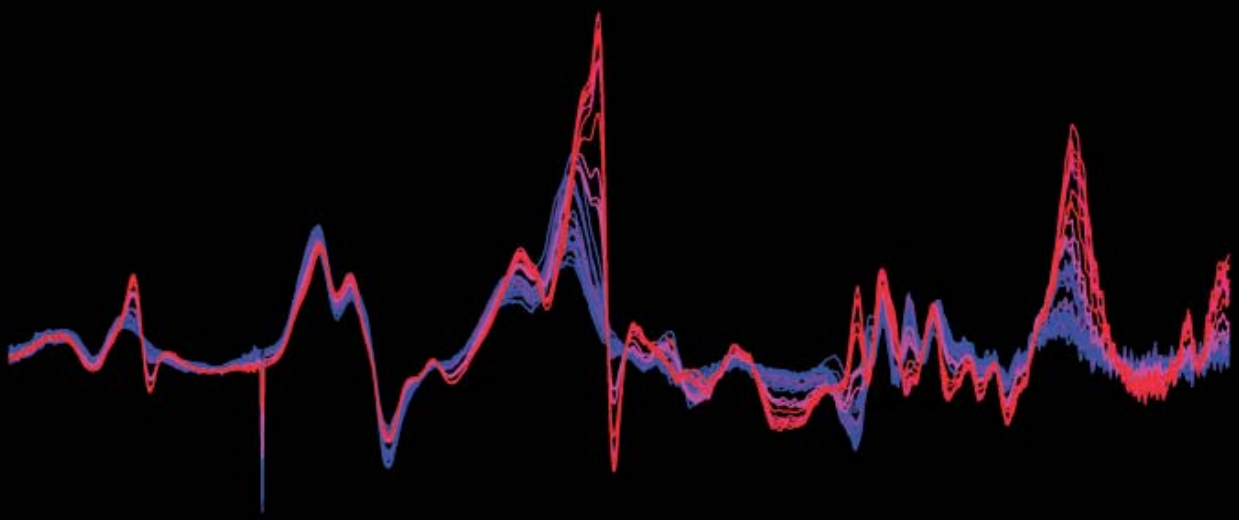
Jan Gertheiss

Feature Extraction in Regression and Classification with Structured Predictors



Cuvillier Verlag Göttingen
Internationaler wissenschaftlicher Fachverlag

Feature Extraction in Regression and Classification with Structured Predictors

Jan Gertheiss

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität
München

München, den 16. Dezember 2010

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet über http://dnb.d-nb.de abrufbar.
1. Aufl. - Göttingen: Cuvillier, 2011
Zugl.: München, Univ., Diss., 2011

978-3-86955-665-9

Erstgutachter: Prof. Dr. Gerhard Tutz
Zweitgutachter: Prof. Dr. Göran Kauermann
Tag der mündlichen Prüfung: 16. Februar 2011

© CUVILLIER VERLAG, Göttingen 2011
Nonnenstieg 8, 37075 Göttingen
Telefon: 0551-54724-0
Telefax: 0551-54724-21
www.cuvillier.de

Alle Rechte vorbehalten. Ohne ausdrückliche Genehmigung des Verlages ist es nicht gestattet, das Buch oder Teile daraus auf fotomechanischem Weg (Fotokopie, Mikrokopie) zu vervielfältigen.
1. Auflage 2011
Gedruckt auf säurefreiem Papier.

978-3-86955-665-9

Zusammenfassung

Eine typische Aufgabe bei der statistischen Modellierung ist die Selektion von Variablen. In der vorliegenden Arbeit wird jedoch nicht nur Variablenselektion sondern vielmehr *Feature Extraction* näher untersucht. Feature Extraction geht über bloße Variablenselektion hinaus, in dem Sinne, dass nicht einfach Variablen ausgewählt sondern bestimmte Merkmale erfasst werden sollen, die je nach Art der betrachteten Daten unterschiedlich sein können.

In dieser Dissertationsschrift werden Variablen mit einer speziellen Struktur betrachtet, wobei diese Größen als Prädiktoren in Regressions- oder Klassifikationsproblemen dienen sollen. Den ersten untersuchten Datentyp stellen hochdimensionale signalartige (metrische) Kovariablen dar. Ein typisches Beispiel für diese Art von Daten sind funktionale Prädiktoren in der Signalregression, die zwar nur an (einer Vielzahl von) einzelnen Messpunkten erfasst werden können, aber dennoch als Realisationen (mehr oder weniger) glatter Kurven angesehen werden sollten. Hier kann Feature Extraction als die 'Identifikation der relevanten Teile des Signals' definiert werden. Zu diesem Zweck wird in der vorliegenden Arbeit ein Boosting-Verfahren entwickelt, welches auch auf Protein-Massenspektren wie sie in der Proteomik vorkommen angewandt werden kann. Mit Hilfe von Simulationsstudien sowie an Hand realer Daten kann gezeigt werden, dass das vorgestellte Verfahren eine äußerst konkurrenzfähige Alternative zu bestehenden Verfahren darstellt.

Kategoriale Kovariablen sind eine weitere hochinteressante Art von speziell strukturierten Prädiktoren. Kategoriale Kovariablen werden in der Regel dummy-kodiert und resultieren folglich in Gruppen von Dummy-Variablen. Haben die betrachteten Größen allerdings ordinales Skalenniveau, wird diese Ordnung der Kategorien bei der Modellierung oftmals ignoriert, oder aber es werden (fälschlicherweise) Methoden angewandt, die eigentlich für Variablen mit metrischem Niveau gedacht sind. In dieser Arbeit werden nun penalisierte Likelihood-Ansätze vorgeschlagen, die ordinales Skalenniveau in den unabhängigen Größen über eine Differenzen-Penalty auf benachbarten Dummy-Koeffizienten berücksichtigen. Neben dem Aspekt der Variablenselektion wird auch die Identifikation relevanter Differenzen zwischen Kategorien sowohl ordinal als auch nominal skalierter Prädiktoren betrachtet und es werden geeignete L_1-Regularisierungstechniken vorgestellt. Die Verfahren werden dabei sowohl aus einem praktischen als auch einem theoretischen Blickwinkel heraus untersucht. Es wird gezeigt, dass die vorgestellten Methoden sinnvoll einsetzbar sind und auch im Vergleich mit alternativen Ansätzen sehr gut abschneiden. Darüber hinaus werden auch kategoriale (potentiell) Effekt-modifizierende Faktoren in Modellen mit variierenden Koeffizienten betrachtet.

Zu guter Letzt werden Ansätze zur nonparametrischen Feature Extraction unter Verwendung von Nearest-Neighbor-Verfahren vorgestellt. Das Abschneiden des in diesem Zusammenhang vorgeschlagenen Nearest-Neighbor-Ensembles ist dabei äußerst vielversprechend.

Summary

A typical task in statistical modeling is variable selection. In this thesis, however, not only variable selection but *feature extraction* is investigated. Feature extraction goes beyond variable selection in the sense that not only variables are selected but features which depend on the special nature of the data considered.

In this dissertation, variables with a special structure are considered and used as predictors in regression and classification problems. High-dimensional signal-like (metric) covariates are the first type of data investigated. A typical example for this kind of data are functional predictors in signal regression, which can only be observed at (a high number of) distinct measurement points but are realizations of (more or less) smooth functions. In this case, feature extraction can be defined as 'the identification of relevant parts of the signal'. For that purpose, a Boosting technique is developed, which can also be applied to curves of protein intensities obtained from mass spectrometry in proteomics. Simulation studies and real world data applications show that the proposed procedure is a highly competitive alternative to existing approaches.

Categorical covariates, which are usually dummy-coded and hence result in groups of dummy variables, are another very interesting type of structured regressors. If predictors are ordinal, however, the levels' ordering is typically ignored in regression modeling, or methods for metric covariates are (wrongly) applied. In this thesis, penalized likelihood methods are proposed which take the ordinal scale level into account using a difference penalty on adjacent dummy coefficients. Besides variable selection, the identification of relevant differences between categories of both ordinal and nominal predictors is considered, and appropriate L_1-type regularization techniques 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, also in comparison with alternative approaches. Categorical covariates serving as (potentially) effect modifying factors in varying-coefficient models are considered, too.

Finally, approaches for nonparametric feature extraction using nearest neighbor methods are presented. The performance of the proposed nearest neighbor ensemble technique is quite encouraging.

Vorwort und Danksagung

Die vorliegende Arbeit entstand zum größten Teil im Rahmen meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Institut für Statistik der Ludwig-Maximilians-Universität München. Zuallererst möchte ich daher meinem Doktorvater Herrn Prof. Dr. Gerhard Tutz dafür danken mir diese Stelle angeboten und mich in den darauf folgenden Jahren so hervorragend betreut zu haben. Ebenso möchte ich Herrn Prof. Dr. Göran Kauermann für die Bereitschaft danken als Zeitgutachter zu fungieren. Darüberhinaus gilt mein Dank meinen Freunden und Kollegen für die gute Zusammenarbeit, Unterstützung bei Problemen und Korrekturlesen. Da große Teile der Arbeit bereits in Fachzeitschriften veröffentlicht (oder zumindest akzeptiert) wurden, möchte ich hier auch den zahlreichen und größtenteils anonymen Referees und (Associate) Editors für ihre – manchmal nervigen oder schwer nachvollziehbaren aber doch meist sehr hilfreichen – Kommentare und Anregungen danken.

Da meine Tätigkeit an der LMU zu großen Teilen aus DFG-Projektmitteln (DFG Projekt TU62/4-1) finanziert wurde, möchte ich an dieser Stelle auch der Deutschen Forschungsgemeinschaft für die entsprechende Unterstützung danken.

Ohne vorheriges Studium wäre eine Promotion aber selbstverständlich nicht möglich gewesen. Daher gilt mein aufrichtiger Dank insbesondere auch meinen Eltern, die mir das Studium ermöglichten und mich in meinen Entscheidungen immer unterstützten. Auch möchte ich der Studienstiftung des deutschen Volkes und meinem damaligen persönlichen Vertrauensdozenten Herr Prof. Dr. August König für die finanzielle und ideelle Unterstützung während meines Studiums herzlich danken.

München, im Dezember 2010 *Jan Gertheiss*

Contents

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.

2 Handling Signal-like Predictors by Blockwise Boosting

2.1 Introduction

In this chapter, we deal with the first type of structured predictors investigated in this thesis: signal-like metric predictors. As already described, the typical example for this kind of data are functional regressors in signal regression, where a scalar quantity is to be explained or predicted by a curve – the signal. For technical reasons these curves can only be observed at (a large number of) adjacent measurement points. However, the knowledge that observations represent (more or less smooth) functions should be used when analyzing the data.

The main objectives of feature extraction in signal regression are improved accuracy of the prediction of future data and the identification of relevant parts of the signal. In Section 2.2 we will introduce a feature extraction procedure that uses boosting techniques to select the relevant parts of the signal, whereby the proposed blockwise Boosting procedure simultaneously selects intervals in the signal's domain and estimates the effect on the response. The blocks that are defined explicitly use the underlying metric of the signal. The method is based on the so-called functional linear model, where a metric (e.g., approx. normal) response is assumed; see Section 2.2 for details.

Another interesting type of high-dimensional signal-like metric predictors comes from proteomics where mass spectrometry is used to measure protein concentrations, resulting in (rather jagged) curves of protein intensities. These spectra are typically used to discriminate between groups, as, for example, cancer/control. In Section 2.3 we will show how blockwise Boosting can be successfully used for feature extraction and classification in mass spectrometry-based predictive proteomics. Moreover, it is shown how the method can be generalized to other (clearly non-normal) outcomes, and some computational aspects a discussed.

Large parts of this chapter are also found in Tutz and Gertheiss (2010) and Gertheiss and Tutz (2009c).

2.2 Feature Extraction in Signal Regression

2.2.1 Introduction

Signal regression has been extensively studied in the chemometrics community, with an excellent summary of tools by Frank and Friedman (1993). With the recent surge of interest in functional data, signal regression may be embedded into the framework of functional data, as outlined by Ramsay and Silverman (2005). Figure 2.1 shows signal regressors from near-infrared spectroscopy, as applied to a compositional analysis of 32 marzipan samples (Christensen et al., 2004). Each *signal* is represented by 600 digitizations along the wavelength axis, where the objective of the analysis is to determine moisture and sugar content (for details, see Section 2.2.4).

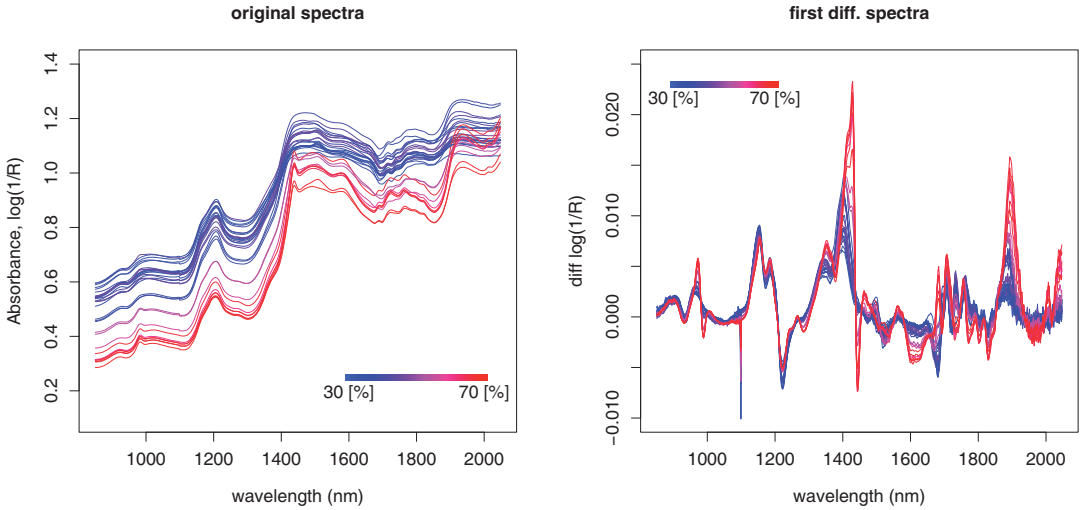


Figure 2.1: NIR spectra of 32 marzipan samples; colors corresponding to sugar content.

Objectives in functional regression are manifold. However, the focus here is on two aspects: accuracy of prediction on future data and feature extraction. When the main concern is prediction, feature extraction is secondary but may assist in obtaining better prediction performance. In other cases, feature extraction is of interest because of its interpretability. Here the object is to determine which predictors effect the response; in case of the spectroscopy data, that means to identify the relevant areas of wavelength.

In this section, a method of feature extraction is proposed which focuses on intervals in the signal's domain. As later illustrated, in a discretized setting, these intervals turn into groups of adjacent variables. By using Boosting techniques it is possible to select subsets of adjacent predictors whose coefficients are interpretable. Moreover, it is demonstrated that the selection of groups of predictors improves the accuracy of prediction when compared to alternative procedures.

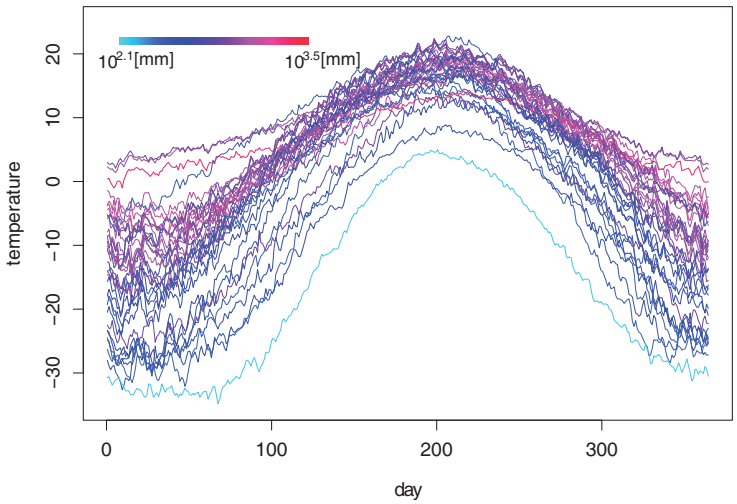


Figure 2.2: Temperature profiles of 35 Canadian weather stations; colors corresponding to (log) total annual precipitation.

As a second example the (benchmark) rainfall data from Ramsay and Silverman (2005) will be analyzed. Figure 2.2 shows the temperature profiles (in degrees Celsius) of Canadian weather stations across the year, averaged over 1960 to 1994. We will use the same response variable as Ramsay and Silverman, namely the base 10 logarithm of the total annual precipitation. There is a wide range of methods that apply to such functional data. Classical instruments are partial least squares (PLS) and principal-component regression (PCR). More recently developed tools aim at constraining the coefficient vector to be a smooth function (see Hastie and Mallows, 1993; Marx and Eilers, 1999, 2005); however, the much older Ridge regression (Hoerl and Kennard, 1970), the new Elastic Net (Zou and Hastie, 2005) and the Fused Lasso (Tibshirani et al., 2005) are also able to handle high-dimensional predictor spaces. A first illustration of the differences between methods is given in Figure 2.3, where the coefficient function resulting from Lasso (Tibshirani, 1996), Fused Lasso, Ridge regression, generalized Ridge regression with first-difference penalty, functional data approach (Ramsay and Silverman, 2005) and the proposed BlockBoost are shown for the Canadian weather data. It is seen that Lasso selects a minimal number of variables (i.e. measurement points), theoretically at most n variables (with n denoting the number of observations), as pointed out by Zou and Hastie (2005). Thus, in the considered example only a few days are selected whose mean temperature is assumed to be relevant for the total annual precipitation. In contrast, Ridge regression takes into account all variables. Smoothing the resulting coefficient function is possible by penalizing differences between adjacent coefficients – or by using smooth basis functions – as proposed by Marx and Eilers (1999) and Ramsay and Silverman (2005). For the functional data

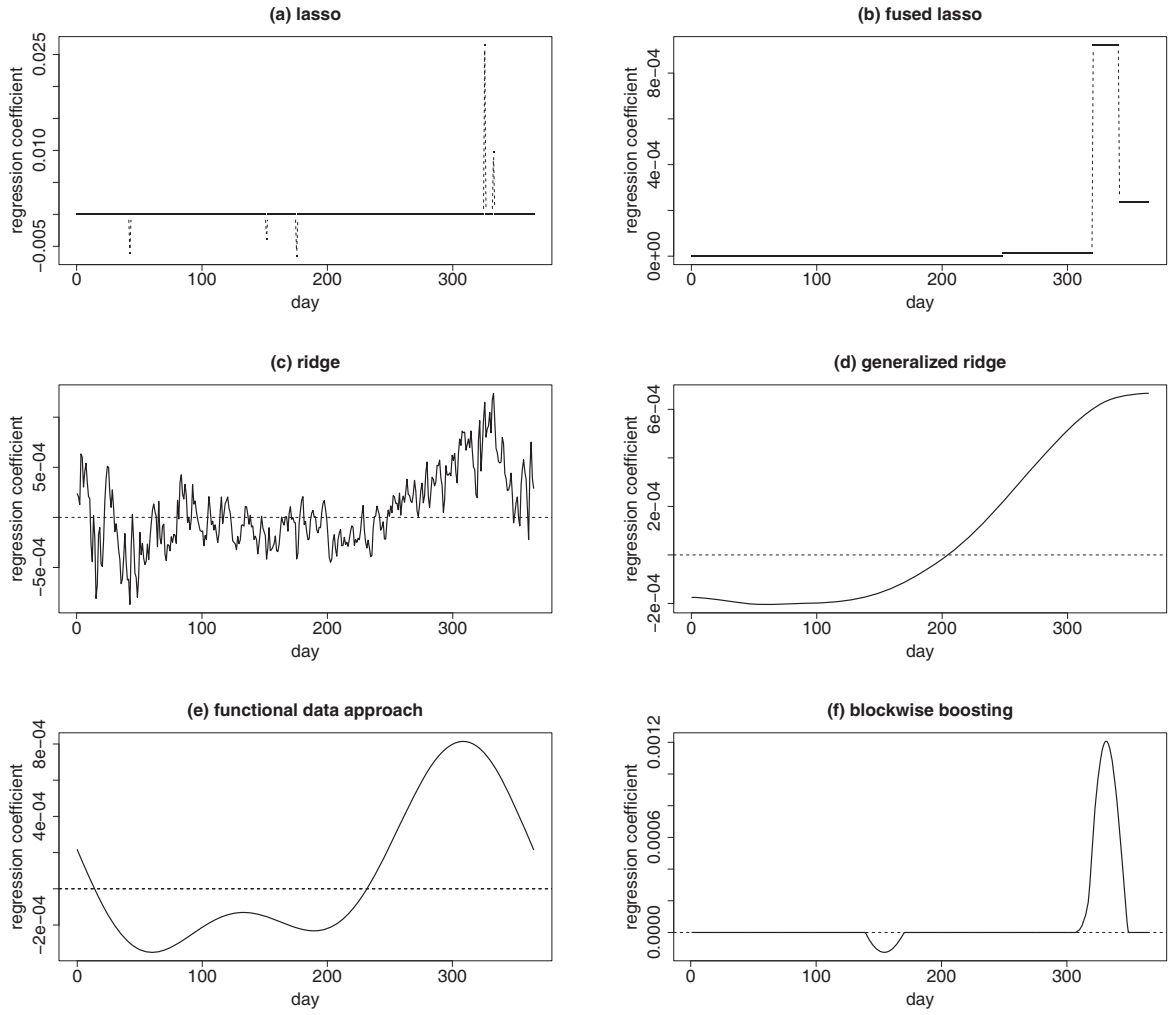


Figure 2.3: Regression coefficients estimated by various methods: Lasso, Fused Lasso, Ridge, generalized Ridge with first-difference penalty, functional data approach, blockwise Boosting; temperature profiles of 35 Canadian weather stations as predictors, log total annual precipitation as response.

approach (see Figure 2.3 (e)), for example, Fourier basis functions for smoothing both the functional regressors and the coefficient function have been used; for details see Ramsay and Silverman (2005). However, almost all of the data are considered to be important. Alternatively, the Fused Lasso and the BlockBoost select only certain periods, for example, a number of weeks in late autumn/early winter. In general, though, smooth estimates seem to make more sense than coefficient functions with abrupt jumps. Why should the effect of temperature substantially change from day to day? The smooth BlockBoost estimates result from penalizing (squared first) differences between adjacent coefficients. Details of the procedure are given in Section 2.2.2.