Boosting for Regression Problems with Complex Data

by

Xiaomeng Ju

B.Sc., Renmin University of China, 2013
M.Sc., University of Michigan, 2015

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate Studies

(Statistics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

July 2021

Abstract

The `genthesis.cls` LaTeX class file and accompanying documents, such as this sample thesis, are distributed in the hope that it will be useful but without any warranty (without even the implied warranty of fitness for a particular purpose). For a description of this file's purpose, and instructions on its use, see below.

These files are distributed under the GPL which should be included here in the future. Please let the author know of any changes or improvements that should be made.

Michael Forbes. mforbes@physics.ubc.ca

Preface

You must include a preface if any part of your research was partly or wholly published in articles, was part of a collaboration, or required the approval of UBC Research Ethics Boards.

The Preface must include the following:

- A statement indicating the relative contributions of all collaborators and co-authors of publications (if any), emphasizing details of your contribution, and stating the proportion of research and writing conducted by you.
- A list of any publications arising from work presented in the dissertation, and the chapter(s) in which the work is located.
- The name of the particular UBC Research Ethics Board, and the Certificate Number(s) of the Ethics Certificate(s) obtained, if ethics approval was required for the research.

Examples

Chapter **??** is based on work conducted in UBC's Maple Syrup Laboratory by Dr. A. Apple, Professor B. Boat, and Michael McNeil Forbes. I was responsible for tapping the trees in forests X and Z, conducted and supervised all boiling operations, and performed frequent quality control tests on the product.

A version of chapter **??** has been published **?**. I conducted all the testing and wrote most of the manuscript. The section on "Testing Implements" was originally drafted by Boat, B. Check the first pages of this chapter to see footnotes with similar information.

Note that this preface must come before the table of contents. Note also that this section "Examples" should not be listed in the table of contents, so we have used the starred form: `\section*{Example}`.

Table of Contents

Abstract ii

Preface iii

Table of Contents iv

List of Tables vi

List of Figures vii

List of Programs viii

Acknowledgements ix

Dedication x

1 Introduction 1
1.1 Outline of the Thesis 4

2 Background 6
2.1 The regression problem 6
2.2 Gradient boosting 7
2.3 Robust regression estimators 9
2.3.1 M-estimators 9
2.3.2 S-estimators 11
2.3.3 MM-estimators 11
2.4 Functional regression estimators 11
2.4.1 Functional linear estimator 11
2.4.2 Functional nonparametric estimator 11
2.4.3 Functional single-index estimator 11
2.5 Functional principal component analysis 11

3 Robust boosting for regression problems 12
3.1 Introduction 12
3.2 SBoost 14
3.3 RRBoost 14
3.3.1 Two stages 14
3.3.2 The initial fit 14
3.3.3 Early stopping 14
3.4 Robust variable importance 14
3.5 Simulation studies 14
3.6 Empirical studies 14

4 Tree-based boosting for functional data 15
4.1 Introduction 15
4.2 Functional multi-index tree 15
4.2.1 Type A tree 15
4.2.2 Type B tree 15
4.3 TFBoost 15
4.4 Simulation studies 15
4.5 Analysis of the German electricity data 15
4.5.1 Functional regression 15
4.5.2 Partial-functional regression 15

5 Robust tree-based boosting for functional data 16
5.1 Introduction 16
5.2 Robust TFBoost 16
5.3 Simulation studies 16

6 Applying boosting to sparse functional data 17
6.1 Introduction 17

7 Concluding Remarks and Future Work 18
7.1 Conclusion 18
7.2 Future work 18

Bibliography 19

Appendices

A First Appendix 22

B Second Appendix 23

List of Tables

List of Figures

List of Programs

Acknowledgements

This is the place to thank professional colleagues and people who have given you the most help during the course of your graduate work.

Dedication

The dedication is usually quite short, and is a personal rather than an academic recognition. The *Dedication* does not have to be titled, but it must appear in the table of contents. If you want to skip the chapter title but still enter it into the Table of Contents, use this command `\chapter[Dedication]{}`.

Note that this section is the last of the preliminary pages (with lowercase Roman numeral page numbers). It must be placed *before* the `\mainmatter` command. After that, Arabic numbered pages will begin.

Chapter 1

Introduction

With the advancement of modern technology, the data collected for various applications not only grow in the number of objects, but also in the complexity of features and error distributions. Due to new data-gathering technology and improved storage capacity, people are able to collect and store a large number of features of a given objects. Depending on the nature of data, these features can be viewed as realizations of multiple covariates or evaluations of a functional covariate on a finite grid.

The "large p small n" setting poses challenges to regression problems and leads the development of new statistical methodologies. When the number of objects n exceeds the number features p, it becomes difficult to obtain a stable estimate of the regression function. Hence, it is natural to conduct variable selection and impose assumptions on the form of the estimator. A lot of exciting research has been conducted in the multivariate setting to fit regression models. In particular, boosting has received great attention being a power technique widely applied to regression with a large number of features. When there is not sufficient data to fit a single accurate regression model, boosting's strategy of combining multiple "simple" base learners often leads to a better fit. Friedman (2001a) derived a gradient-descent based boosting algorithm from a statistical framework. This gradient boosting algorithm has been one of the most popular statistical learning techniques, and have shown considerable success across applications in various scientific domains.

Boosting methods generally assume that the data does not contain atypical observations. However, this is usually not the case. In real applications, collected data is often noisy and may contain "outliers". These "outliers" can be points that follow a different model that is remarkably different from the one that best describes the majority of the data, or simply "extreme" values that deviates from rest. For example, a data point may be regarded as an outlier if the variable is believed to be normally distributed, but it in fact comes from a very different normal distribution. Another common example is skewed or heavy-tailed distributions. In this case, data points at the skewed-end or heavy tails are typically viewed as outliers.

Outlying values can occur in the x-direction, y-direction, or both. A data point that has extreme predictor x-values is referred to as a high-leverage point. With multiple predictors, extreme x values may be particularly high or low for one or more predictors, or may be "unusual" combinations of predictor values. For example, with two positively correlated predictors, an unusual combination might be a high value of one predictor paired with a low value of the other predictor. A data point whose response y does not follow the general model of the rest of data is called a "vertical outlier". For regression purposes, the outliers are typically referring to vertical outliers, which could have high-leverage or not. Based on the type of the estimator.the leverage of outliers can impact regression estimators differently. For linear regression estimators, high-leverage outliers cause more damage than low-leverage ones. They pull the fit towards them at the cost of severely increased residuals for low-leverage points. By comparison, some nonparametric estimators constructs local fits using neighbouring points only and thus reduces the influence of high-leverage points on the rest of the data.

There are many causes for outliers — measurement errors, entry mistake, sampling problems, or natural variation due to variable distributions. In practice, the proportion and cause of outliers are typically unknown, and it is usually impossible to make assumptions on their distributions and model them explicitly. When a boosting algorithm is applied without accounting for outliers, the regression estimator can be highly influenced, leading to unstable predictions that provide misleading conclusions.

The impact of outliers has recognized and studied by numerous researchers in the past 50 years. A series of robust location estimators have been developed by minimizing a robust objective function down-weighting the contribution of potential outliers. Some robust regression estimators have been proposed following the same principle. However, these methods have so far been focused on fitting linear models, with very few options available for non-parametric models with many features. The ones proposed for boosting either yield estimators with low efficiency or requires an ad hoc residual scale estimator calculated at every iteration, which may be unstable or severely biased. These issues require the development of new boosting methods to perform regression in the presence of outliers.

The first main contribution of this thesis is to present a boosting algorithm that is expected to be robust and highly efficient. We propose a two-stage approach similar as it is done in regression MM-estimators: first stage computes an estimator with high robustness but possibly low efficiency, and then improves its efficiency in the second stage. We recommend the use of regression tree as base learners, which are fast to compute, scalable to

high-dimensional data, and robust to high-leverage points.

From finite dimensional data, we move to infinite dimensional data, also called functional data. This type of data consist of a sample of functions defined on some continuous domain, and in theory, measurements could be taken at any time points on functional curves. Examples of functional data are encountered across various scientific fields, including analysis of growth curves, climate variation, handwriting, spectrometry data, medical research, and many others.

When each functional curve is observed at the same set of design points, the observations may look like multivariate data. In a number of situations, classical statistical methods can be applied to such data treating it as multivariate. However, they do not take advantage of the smoothness of the underlying functions, ignoring the additional information contained in the function and its derivatives. In addition, functional methods do not require each individual to be observed at the same design points, and the time intervals do not have to be equally spaced. For a functional object, the observations at different time points are typically dependent, which violates the assumption of many multivariate techniques. These aspects motivate the development of regression methods in recent decades for functional data.

Functional regression models are categorized into "scalar-on-function", "function-on-scalar", and "function-on-function" depending on the role played by the functional data. We focus on the first case with scalar responses and functional predictors. On this topic, most existing methods can be classified into linear, nonparametric, or single-index models. Different from them, we consider a multi-index model that provides greater flexibility compared to linear and single index models and circumvents the "curse of dimensionality" arises in nonparametric models. This makes up the second main contribution of this thesis, developing a boosting method that constructs a multi-index estimator for "scalar-on-function" regression. We propose two types of multi-index functional trees as base learners and present algorithms to compute them. The method is extendable to the partial-functional setting in which the data contains functional and scalar covariates.

Same for the finite dimension case, functional evaluations may also contain outliers. These abnormal values include evaluations on entire outlying curves (global outliers) or on curves with outlying values in some parts across the domain (local outliers). The former may not contain extreme values throughout the entire domain, but have a different shape from the bulk of data. This is similar as the concept of high-leverage point in the finite dimensional setting. Besides functional outliers, the scalar response y may have vertical outliers, which would adversely influence the fit of the

regression function.

To address this problem, we explore robust boosting method for functional data by combining the ideas of our two previous proposals: one on robust boosting, and the other on functional boosting. `Challenges of high-leverage points..., if we need to propose something else...`

Most functional methods assume the functions to be evaluated on a a large number of time points, usually on a dense regular grid. In real situations, however, the observed design points can be sparse and vary greatly across objects. In this case, the "direct method" of pre-smoothing each individual curve then compute the estimator with the smoothed curves may not be viable. For example, individuals with few sampled time points (e.g. less than 5) will produce unreliable curve estimates and consequently affects the regression fit. This implies that one needs to take the sparseness into account when analyzing such data in a regression framework.

A better strategy to model sparse functional data is to borrow strength across individuals, combining information from different evaluations on the same curve and across curves. This can be achieved by sparse functional principle components analysis (FPCA) which estimates the covariance function with sparse data using measurements from periods where data have been collected to effectively estimates the correlation from periods with little or no measurements. Based on sparse FPCA, we provide an idea to extend functional boosting to sparsely sampled functional data. From there, we explore its generalization to sparsely sampled data with functional outliers as well as vertical outliers.

1.1 Outline of the Thesis

This thesis covers boosting for regression with two main types of complex data: (1) finite dimensional (including high-dimensional) data with outliers, (2) densely or sparsely sampled functional data with or without outliers. It is organized as follows:

Chapter 2 provides a background on the topics in this thesis. We define the regression model and introduce gradient boosting for its estimation. We also provide a description of robust regression estimators, functional regression estimators, and functional principle component analysis, which are used to develop new boosting methods in later chapters.

Chapter 3 addresses the problem of regression with outliers in the finite-dimensional setting. We introduce our proposed robust boosting estimator, and a variable importance score based on permutations. Then, we report the

results of comparing our method with previously proposed robust boosting algorithms on simulated data as well as benchmark data.

Chapter 4 introduces a tree-based boosting estimator for regression with functional data. The method uses functional multi-index trees as the base-learners, which includes two types of trees (type A and type B) that differ in how they find the optimal index. We establish the identifiability conditions of the trees and present algorithms to compute them. We report results of simulation studies comparing our method with competing alternatives, and close this chapter by showing a case study using our method to predict electricity demand in the German power market.

Chapter 5 develops a robust tree-based boosting estimator for functional data based on the proposal in Chapter 4. (`unfinished`)

Chapter 6 expands the use of boosting in Chapter 4 and Chapter 5 to sparsely sampled functional data. We investigate the impact induced by the sparseness of data on our functional boosting estimator and offer solutions to this problem. (`unfinished`)

Chapter 7 makes concluding remarks and points out areas of future work. Valuable extensions of this thesis are discussed, including obvious next steps and directions that has yet to be explored.

Chapter 2

Background

In this chapter, we introduce the terminology and methods used throughout this thesis. To set the stage, we begin by defining the regression model and the associated prediction problem in Section 2.1. In Section 2.2, we provide an overview of gradient boosting and detail the algorithm proposed by Friedman (2001a). We follow this with an introduction of robust regression estimators and functional regression estimators in Section 2.3 and Section 2.4 respectively. Lastly, we review the basics of functional principle component analysis (FPCA) in Section 2.5.

2.1 The regression problem

Consider that we are interested in building a model to predict the response $Y \in \mathcal{Y}$ using variables $\mathbf{X} = \{X_1, ..., X_p\} \in \mathcal{X}$. The general form of a regression model is:

$$Y = F(\mathbf{X}) + \epsilon, \tag{2.1}$$

where ϵ is random error that is independent from $\mathbf{X}$. Typically, we assume ϵ to have zero mean and finite variance. The variables $\mathbf{X}$ are interchangeably referred to as "features", "covariates", "predictors", or "exploratory variables".

Given samples $(y_i, \mathbf{x}_i), ..., (y_i, \mathbf{x}_i)$ that are i.i.d realizations of $(\mathbf{X}, Y)$, the goal of regression analysis is to estimate the function F in (2.1) in order to make predictions for future observations. Besides prediction, practitioners sometimes make inference based on the estimation, interpreting how the response is influenced by the features.

Throughout this thesis, we assume the response to be a real-valued variable: $y_i \in \mathbb{R}$. In Chapter 3, we assume the features to be finite dimensional real variables $\mathbf{X}_i \in \mathbb{R}^p$. In Chapters 4 to 7, $\mathbf{x}_i$ are measurements taken on a functional curve. The dimension of $\mathbf{x}_i$ may differ across individuals, subject to the number of measurements taken for each curve. In what follows,

we use bold face fonts to represent vectors, upper case letters for random variables, and lower case letters for sampled values of random variables.

To estimate the regression function F in (2.1), one of the most commonly used approach is to minimize the empirical risk. We assume there is a joint probability distribution $P(\mathbf{x}, y)$ over $\mathbf{X}$ and Y, and we are given a loss function $L(y, \hat{F}(\mathbf{x}))$ that is non-negative and measures the discrepancy between the true value y and prediction $\hat{F}(\mathbf{x})$. The risk associated with any function $f : \mathbf{X} \to Y$ is defined by the expectation of the loss function $E[L(Y, f(\mathbf{X}))] = \int L(y, f(\mathbf{x}))dP(\mathbf{x}, y)$, which we seek to minimize over f. Since the distribution $P(\mathbf{x}, y)$ is unknown, we define a regression estimator that minimizes the empirical risk (the average of the loss on the training data):

$$\hat{F} = \underset{f}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i)). \tag{2.2}$$

If $L(y, \hat{F}(\mathbf{x})) = (y - \hat{F}(\mathbf{x}))^2$, the resulted $\hat{F}$ is referred to as the "least squares estimator" that is often used as the default choice.

2.2 Gradient boosting

Gradient boosting provides a way to solve (2.2) via an iterative step-wise algorithm. It considers a function estimation of the form

$$\hat{F}(\mathbf{x}) = \sum_{t=1}^{T} \alpha_t h_t(\mathbf{x}),$$

where T is the number of iterations and the functions $h_t : \mathcal{X} \to \mathbb{R}$ belongs to a class of base learners $\mathcal{H}$. The h_t as fitted in a sequential fashion; at the t-th iteration, the newly added h_t is computed while keeping the previous ones h_1, ..., h_{t-1} fixed.

We view the empirical risk $\frac{1}{n}\sum_{i=1}^{n} L(y_i, F(\mathbf{x}_i))$ as a function of the vector $(F(x_1), ..., F(x_n))^T$. If we were to apply gradient descent to minimize the empirical risk treating $(F(x_1), ..., F(x_n))^T$ as parameters, at the t-th iteration, we add the negative gradient vector computed at the point (-$\mathbf{g}_t$) to the current estimates $(\hat{F}_{t-1}(x_1), ..., \hat{F}_{t-1}(x_n))^T$. The negative gradient vector $\mathbf{g}_t$ is computed at the point obtained form the previous iteration $(\hat{F}_{t-1}(x_1), ..., \hat{F}_{t-1}(x_n))^T$:

$$g_{t,i} = -\frac{\partial L(y_i, b)}{\partial b}\Big|_{b=\hat{F}_{t-1}(x_i)}, \ i = 1, ..., n,$$

Similar to gradient descent, gradient boosting starts with an initial guess $\hat{F}_0(x)$, generally a constant fit

$$F_0(x) = \underset{a \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, a),$$

and adds to the current function estimate an approximation to the negative gradient vector (-$\mathbf{u}_t$). More specifically, at the t-th iteration, base learner h_t is chosen to minimize

$$\sum_{i=1}^{n} (h_t(\mathbf{x}_i) + g_{t,i})^2 \,,$$

over members h_t of the family of base learners ($\mathcal{H}$). We then set

$$\hat{F}_t(\mathbf{x}) = \hat{F}_{t-1}(\mathbf{x}) + \alpha_t h_t(\mathbf{x}),$$

where α_t is the optimal step size given by

$$\alpha_t \;=\; \underset{\alpha}{\operatorname{argmin}} \sum_{i \in \mathcal{I}_{\text{train}}} L(y_i, \hat{F}_{t-1}(\mathbf{x}_i) + \alpha h_t(\mathbf{x}_i)) \,. \tag{2.3}$$

The resulting gradient boosting algorithm is shown in Algorithm 1. Among different choices of base learners, many prefer regression trees due to its fast computation, automatic variable selection, and flexibility as a nonparametric estimator. For these reasons, our boosting implementations in later chapters are all based on tree learners, either the original or variants of regression tress.

Many boosting algorithms also include a "shrinkage" option (Friedman, 2001b). The idea is to reduce the size of the function updates by multiplying the optimal step size α_t in (2.3) by a small fixed constant $\gamma \in (0, 1)$, which correspond to replacing line Line 5 in Algorithm 1 with

$$\hat{F}_t(\mathbf{x}) \;=\; \hat{F}_{t-1}(\mathbf{x}) + \gamma \, \alpha_t \, h_t(\mathbf{x}) \,.$$

Shrinkage is expected to reduce the impact of each base learner and approximate more finely the search path in gradient descent (Telgarsky, 2013). A small γ typically improves the predictive accuracy, but requires a larger number of iterations and thus increases the computational cost of the algorithm. This simple and effective strategy can be viewed as a form of regularization.

Algorithm 1: Gradient boosting

Input : A training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$
A loss function $L(y, F(\mathbf{x}))$
The number of iterations T
The class of weak learners $\mathcal{H}$

Initialize: $\hat{F}_0(\mathbf{x}_i) = \text{argmin}_{c\in\mathbb{R}} \sum_{i=1}^n L(y_i, c)$

for $t = 1 : T$ **do**

$g_t(\mathbf{x}_i) = -\frac{\partial L(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)}\Big|_{F(\mathbf{x}_i)=\hat{F}_{t-1}(\mathbf{x}_i)}$

$h_t = \text{argmin}_{h\in\mathcal{H}} \sum_{i=1}^n (g_t(\mathbf{x}_i) - h(\mathbf{x}_i))^2$

$\alpha_t = \text{argmin}_\alpha \sum_{i=1}^n L(y_i, \hat{F}_{t-1}(\mathbf{x}_i) + \alpha h_t(\mathbf{x}_i))$

$\hat{F}_t(\mathbf{x}_i) = \hat{F}_{t-1}(\mathbf{x}_i) + \alpha_t h_t(\mathbf{x}_i)$

end

Output : $\hat{F}_T(\mathbf{x}) = \sum_{t=1}^T \alpha_t h_t(\mathbf{x})$

2.3 Robust regression estimators

For regression problems, it is well known least squares estimators are highly sensitive to outliers. Robust regression aims to reduce the impact outliers and learn the model that represents the bulk of the data. We begin by introducing the

2.3.1 M-estimators

Before defining robust boosting, we set up the notion and review M-estimators in the standard regression setting. Assume that data are generated from a linear model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, i = 1, ..., n$$

where $\mathbf{x}_i \in \mathbb{R}^p$ are the exploratory variables, ϵ_i follows a symmetric distribution centred around zero with standard deviation σ, and $\boldsymbol{\beta} \in \mathbb{R}^p$ are parameters that need to be estimated. The fitted values and the residuals r_i are defined respectively as

$$\hat{y}_i(\boldsymbol{\beta}) = \mathbf{x}_i^T \boldsymbol{\beta} \text{ and } r_i(\boldsymbol{\beta}) = y_i - \hat{y}_i(\boldsymbol{\beta}).$$

The regression M-estimator **??** is defined as the solution to

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \rho \left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}} \right), \tag{2.4}$$

where ρ is a ρ-function (see Section 2.3.3) and $\hat{\sigma}$ is a scale estimator of ϵ_i (see Section **??**).

The function ρ in (2.4) defines different types of M-estimators. For instance, $\rho(u) = u^2$ defines a *least-squares* estimator and $\rho(u) = |u|$ defines a *least absolute deviations (LAD)* estimator. According to the definition of a ρ-function in **?**, a ρ-function is a function ρ that satisfies: $\rho(u)$ is nondecreasing of $|u|$, $\rho(0) = 0$, and $\rho(u)$ is increasing for $u > 0$ such that $\rho(u) < \rho(\infty)$. A ψ-function is the derivative of a ρ-function. We say that a ρ-function is redescending if $\rho(\infty) = 1$ and $\lim_{|u|\to\infty} \psi(u) = 0$. The redescending ρ-functions are bounded, enabling them to eliminate the influence of very large outliers also allowing for a continuous transition of treating the data from "fully good" to "fully bad". M-estimators defined by a redescending ρ-function are redescending estimators that completely eliminate the influence of extreme outliers.

Because of redescending estimators's desirable property of completely ignoring extremely outliers, we design robust boosting methods with a redescending ρ-function and ψ-function. Specifically, we consider Tukey's bisqure functions defined as

$$\rho_{\text{Tukey},c}(u) = \begin{cases} 1 - \left(1 - (\frac{u}{c})^2\right)^3 & \text{if } |u| \le c \\ 1 & \text{if } |u| > c \end{cases}, \tag{2.5}$$

$$\psi_{\text{Tukey},c}(u) = \begin{cases} \frac{6u(1-(u/c)^2)2}{c^2} & \text{if } |u| \le c \\ 0 & \text{if } |u| > c \end{cases} \tag{2.6}$$

where the constant c controls the balance between robustness and efficiency. Figure **??** shows the Tukey's bisquare ρ-functions and ψ-functions with $c = 4.685$ and $c = 1.547$.

Key is the trade-off between robustness and efficiency.

2.3.2 S-estimators

2.3.3 MM-estimators

2.4 Functional regression estimators

2.4.1 Functional linear estimator

2.4.2 Functional nonparametric estimator

2.4.3 Functional single-index estimator

2.5 Functional principal component analysis

Chapter 3

Robust boosting for regression problems

3.1 Introduction

The goal of robust regression methods is to provide reliable estimates for the unknown regression function and predictions for future observations, even when the training data set may contain atypical observations. These "outliers" need not be "extreme" or aberrant values, but might simply be points following a different model from the one that applies to the majority of the data. It is well known that classical methods can provide seriously misleading results when trained on such heteregenous or contaminated data sets, and thus obtaining robust alternatives is of immediate practical importance.

In the robust regression literature much of the attention has been devoted to linear models, and a rich variety of proposals exists for them. For a review see, for example, Maronna et al. (2018). Fewer options are available for robust non-parametric regression methods, and generally they have not been developed as thoroughly as the linear ones. Among them we can mention: robust locally weighted regression (Cleveland, 1979; Wang and Scott, 1994), local M-estimators (Boente and Fraiman, 1989), local regression quantiles (Welsh, 1996), and the proposals in Härdle and Tsybakov (1988), Härdle (1990), and Oh et al. (2007). Unfortunately, when applied to problems with several explanatory variables, these methods suffer from the well known curse of dimensionality (Hastie et al., 2009; Bellman, 1961), requiring training sample sizes that grow exponentially with the number of covariates. An exception is the class of robust estimators for additive models recently proposed in (Boente et al., 2017). However, selecting appropriate bandwidths for each of the components of an additive model can be computationally prohibitive when a moderate or large number of explanatory variables are involved.

In this chapter, we study robust non-parametric regression estimators based on gradient boosting (Friedman, 2001b). These ensemble methods

are scalable to high-dimensional data, and typically require selecting fewer tuning parameters than additive models. Our proposal applies to practical situations where the proportion of potential outliers in the training and validation sets is unknown, and no parametric structure for the regression model is available.

Most robust boosting algorithms in the literature were designed for classification problems (Kégl, 2003; Rosset, 2005; Freund, 2009; Miao et al., 2015; Li and Bradic, 2018). Previous proposals to robustify regression boosting methods replaced the squared loss with the absolute value function, or with Huber's loss. Lutz et al. (2008) introduced several ways to robustify boosting, primarily focusing on linear regression. One of their methods (`Robloss`) can be extended to the scope of our study (nonlinear regression) by replacing the simple linear regression base learners with other types of base learners.

A different class of robust nonparametric regression methods is based on random forests. Roy and Larocque (2012) proposed variations of random forests that use the median to aggregate trees and make the tree splits. Meinshausen (2006) introduced a quantile random forest which models the full conditional distribution of the response variable, not just the conditional mean. Li and Martin (2017) established a connection between random forests and locally weighted regression. They suggested a forest-type regression framework which includes Meinshausen (2006)'s quantile random forest as a special case. It is also applicable to other loss functions, such as Huber's or Tukey's.

Our main concern with these proposals is that they use robust loss functions that either may yield estimators with low efficiency (e.g. the L_1 loss), or require an auxiliary residual scale estimator (e.g. Huber's and Tukey's loss functions). For the latter, previous proposals suggested using in each step a robust scale estimator obtained with the residuals from the fit at the previous iteration (Friedman, 2001b; Lutz et al., 2008). It is easy to see that this changing scale estimator may not work well, since the scale can be overestimated in early iterations and this might result in observations with large residuals not being considered outlying. In fact, our numerical experiments confirm that this is the case even in relatively simple settings. To address this problem, we propose a robust boosting method that directly minimizes a robust scale of the residuals, and hence does not need to compute an auxiliary residual scale estimator. Although in principle our approach can be used with any scale estimator, gradient boosting uses the partial derivatives of the objective function. Hence, in this paper we focus on minimizing an M-estimator of scale, for which we can compute

the corresponding gradient. This approach can be seen as an extension of S-estimators (Rousseeuw and Yohai, 1984) to this class of non-parametric regression estimators. Moreover, this robust boosting estimator can be used (along with its associated robust residual scale estimator) as the initial fit for a boosting M-type estimator using Huber's or Tukey's loss. As in the case of linear models, this second stage is expected to result in a robust estimator with higher efficiency (less variability) than the S-estimator.

The remainder of this paper is organized as follows. Section **??** introduces our proposed robust boosting regression estimator, and a variable importance score based on permutations. Section **??** reports the results of our simulation studies comparing our method with previously proposed robust boosting algorithms, while results obtained on benchmark data are discussed in Section **??**. Finally, Section **??** summarizes our findings and conclusions.

3.2 SBoost

3.3 RRBoost

3.3.1 Two stages

3.3.2 The initial fit

3.3.3 Early stopping

3.4 Robust variable importance

3.5 Simulation studies

3.6 Empirical studies

Chapter 4

Tree-based boosting for functional data

4.1 Introduction

4.2 Functional multi-index tree

4.2.1 Type A tree

4.2.2 Type B tree

4.3 TFBoost

4.4 Simulation studies

4.5 Analysis of the German electricity data

4.5.1 Functional regression

4.5.2 Partial-functional regression

Chapter 5

Robust tree-based boosting for functional data

5.1 Introduction

5.2 Robust TFBoost

5.3 Simulation studies

Chapter 6

Applying boosting to sparse functional data

6.1 Introduction

Chapter 7

Concluding Remarks and Future Work

7.1 Conclusion

7.2 Future work

Bibliography

Bellman, R. (1961). *Adaptive Control Processes*. Princeton University Press.

Boente, G. and Fraiman, R. (1989). Robust nonparametric regression estimation. *Journal of Multivariate Analysis*, 29(2):180–198.

Boente, G., Martínez, A., and Salibián-Barrera, M. (2017). Robust estimators for additive models using backfitting. *Journal of Nonparametric Statistics*, 29(4):744–767.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

Freund, Y. (2009). A more robust boosting algorithm. *arXiv preprint arXiv:0905.2138*.

Friedman, J. H. (2001a). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Friedman, J. H. (2001b). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Härdle, W. (1990). *Applied Nonparametric Regression*. Econometric Society Monographs. Cambridge University Press.

Härdle, W. and Tsybakov, B. (1988). Robust nonparametric regression with simultaneous scale curve estimation. *The Annals of Statistics*, 16(1):120–135.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning – Data Mining, Inference, and Prediction*. Springer Series in Statistics.

Kégl, B. (2003). Robust regression by boosting the median. *Learning Theory and Kernel Machines*, 2777:258–272.

Li, A. H. and Bradic, J. (2018). Boosting in the presence of outliers: adaptive classification with nonconvex loss functions. *Journal of the American Statistical Association*, 113(522):660–674.

Li, A. H. and Martin, A. (2017). Forest-type regression with general losses and robust forest. *Proceedings of the 34th International Conference on Machine Learning*, 70:2091–2100.

Lutz, R. W., Kalisch, M., and Bühlmann, P. (2008). Robustified l2 boosting. *Computational Statistics & Data Analysis*, 52(7):3331–3341.

Maronna, R. A., Martin, R. D., Yohai, V. J., and Salibián-Barrera, M. (2018). *Robust Statistics: Theory and Methods (with R)*. John Wiley & Sons.

Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999.

Miao, Q., Cao, Y., Xia, G., Gong, M., Liu, J., and Song, J. (2015). Rboost: label noise-robust boosting algorithm based on a nonconvex loss function and the numerically stable base learners. *IEEE Transactions on Neural Networks and Learning Systems*, 27(11):2216–2228.

Oh, H.-S., Nychka, D., and Lee, T. (2007). The role of pseudo data for robust smoothing with applications to wavelet regression. *Biometrika*, 94(4):893–904.

Rosset, S. (2005). Robust boosting and its relation to bagging. *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 249–255.

Rousseeuw, P. and Yohai, V. (1984). Robust regression by means of S-estimators. *Robust and Nonlinear Time Series Analysis*, 26:256–272.

Roy, M.-H. and Larocque, D. (2012). Robustness of random forests for regression. *Journal of Nonparametric Statistics*, 24(4):993–1006.

Telgarsky, M. (2013). Margins, shrinkage, and boosting. In *International Conference on Machine Learning*, pages 307–315. PMLR.

Wang, F. T. and Scott, D. W. (1994). The L1 method for robust non-parametric regression. *Journal of the American Statistical Association*, 89(425):65–76.

Welsh, A. (1996). Robust estimation of smooth regression and spread functions and their derivatives. *Statistica Sinica*, 6(2):347–366.

Appendix A

First Appendix

Here you can have your appendices. Note that if you only have a single appendix, you should issue `\renewcommand{\appendicesname}{Appendix}` before calling `\appendix` to display the singular "Appendix" rather than the default plural "Appendices".

Appendix B

Second Appendix

Here is the second appendix.

Additional Information

This chapter shows you how to include additional information in your thesis, the removal of which will not affect the submission. Such material should be removed before the thesis is actually submitted.

First, the chapter is unnumbered and not included in the Table of Contents. Second, it is the last section of the thesis, so its removal will not alter any of the page numbering etc. for the previous sections. Do not include any floats, however, as these will appear in the initial lists.

The `ubcthesis` LaTeX class has been designed to aid you in producing a thesis that conforms to the requirements of The University of British Columbia Faculty of Graduate Studies (FoGS).

Proper use of this class and sample is highly recommended—and should produce a well formatted document that meets the FoGS requirement. Notwithstanding, complex theses may require additional formatting that may conflict with some of the requirements. We therefore *highly recommend* that you consult one of the FoGS staff for assistance and an assessment of potential problems *before* starting final draft.

While we have attemped to address most of the thesis formatting requirements in these files, they do not constitute an official set of thesis requirements. The official requirements are available at the following section of the FoGS web site:

`http://www.grad.ubc.ca/current-students/dissertation-thesis-preparation`

We recommend that you review these instructions carefully.