Boosting for partially linear additive models

Tang, Xingyu

2016

Tang, X. (2016). Boosting for partially linear additive models. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/69082>

<https://doi.org/10.32657/10356/69082>

NANYANG
TECHNOLOGICAL
UNIVERSITY

BOOSTING FOR PARTIALLY LINEAR ADDITIVE MODELS

TANG XINGYU

**DIVISION OF MATHEMATICAL SCIENCES
SCHOOL OF PHYSICAL AND MATHEMATICAL
SCIENCES**

2016

BOOSTING FOR PARTIALLY LINEAR ADDITIVE MODELS

TANG XINGYU

Division of Mathematical Sciences
School of Physical and Mathematical Sciences

A thesis submitted to Nanyang Technological University
in fulfillment of the requirement for the degree of
Doctor of Philosophy in Mathematical Sciences

2016

Acknowledgements

First of all, I would like to thank the three anonymous reviewers for their delightful and valuable comments and suggestions, which have helped a lot improving the thesis.

I am grateful to my dearest three supervisors Prof. Qin Yingli, Prof. Lian Heng, and Prof. Xiang Liming, without whom this thesis would be impossible. The instructions, helps and supports from my supervisors are the most valuable things that I have ever gained during my PhD studies in Division of Mathematical Sciences (MAS), School of Physical and Mathematical Sciences (SPMS), Nanyang Technological University (NTU). I am deeply impressed and encouraged by their intelligence and diligence.

I want to express my gratitude to Prof. Pan Guangming, Prof. Nicolas Privault, Prof. Elise Frédérique Oggier, and all the other faculty members in MAS for their unconditional helps and supports.

I would like to thank my fellow students and friends Yang Qing, Zhao Kaifeng, Fan Zengyan, Tian Ye, Sun Ruimin, Wu Hailing, Li Mengling, Liu Shiqiu, Wu Liyasi, Liu Yue, You Yuan, Pan Ying, Li Chunming, Long Yulin, Chen Yi, Huang Bin, and Meng Jin for their friendships which make my life in NTU (and Singapore) so valuable and unforgettable.

Last but not least, I want to express my love to my parents who have always been unconditionally loving and supporting.

Abstract

Additive models are widely applied in statistical learning. The partially linear additive model is a special form of additive models, which combines the strengths of linear and nonlinear models by allowing linear and nonlinear predictors to coexist. One of the most interesting questions associated with the partially linear additive model is to identify nonlinear, linear, and non-informative covariates with no such pre-specification given, and to simultaneously recover underlying component functions which indicate how each covariate affects the response.

In this thesis, algorithms are developed to solve the above question. Main technique used is gradient boosting, in which simple linear regressions and univariate penalized splines are together used as base learners. In this way our proposed algorithms are able to estimate component functions and simultaneously specify model structure. Twin boosting is incorporated as well to achieve better variable selection accuracy. The proposed methods can be applied to mean and quantile regressions as well as survival analysis. Simulation studies as well as real data applications illustrate the strength of our proposed approaches.

Contents

Acknowledgements	i
Abstract	iii
1 Introduction	1
1.1 The Models	1
1.1.1 Generalized Additive Model	1
1.1.2 Partially Linear Additive Model	4
1.1.3 Quantile Regression	6
1.1.4 Survival Analysis	8
1.2 The Boosting	11
1.2.1 Gradient Boosting	12
1.2.2 Quantile Regression with Boosting	14
1.2.3 Survival Analysis with Boosting	15
1.2.4 Boosting for PLAM	16
1.2.5 Twin Boosting	19
1.3 Thesis Outline	20

2	Mean and Quantile Regressions with Boosting	23
2.1	Methodology	23
2.1.1	Partially Linear Additive Model	23
2.1.2	Base Learners	25
2.1.3	Gradient Boosting	29
2.1.4	Twin Boosting	33
2.2	Simulation Illustrations	35
2.2.1	Estimation Accuracy	37
2.2.2	Prediction Accuracy	39
2.2.3	Variable Selection	44
2.3	Real Data Applications	52
2.3.1	House Price Data	52
2.3.2	Concrete Compressive Strength Data	62
3	Survival Analysis with Boosting	73
3.1	Methodology	73
3.1.1	Survival Time	73
3.1.2	Proportional Hazards Model	74
3.1.3	Censoring and Partial Likelihood	75
3.1.4	Boosting for Survival Analysis	77
3.2	Simulation Illustrations	77
3.2.1	Estimation Accuracy	79
3.2.2	Prediction Accuracy	81
3.2.3	Variable Selection	89
3.3	Real Data Applications	93
4	Conclusions and Discussions	105
4.1	Discussions on Boosting and PLAM	107

4.1.1	Asymptotic Theories	107
4.1.2	More Types of Predictor Effects	109
4.1.3	The Degree and Internal Knots of Spline Basis	110
4.1.4	Degrees of Freedom	112
4.1.5	Stability Selection	113
4.2	Discussions on Quantile Regression	120
4.2.1	Quantile Crossing	123
4.2.2	Discrete Responses	124
4.3	Discussions on Survival Analysis	125
4.3.1	Baseline Hazard Function	125
4.3.2	Tied Times	126
4.3.3	Time-dependent Covariates	126
Appendix		129
A.1	Proof of Proposition 2.3	129
A.2	Supplementary Numerical Results of Chapter 2	131
A.2.1	Estimation Accuracy	131
A.2.2	Prediction Accuracy	131
A.2.3	Variable Selection	131
A.3	Sample R Codes	155
Bibliography		173

List of Tables

2.1	Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates by different approaches, for $N = 300$, $p = 15$, and Gaussian errors. .	40
2.2	Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates by different approaches, for $N = 300$, $p = 15$, and Gaussian errors. (Continued)	41
2.3	Average (standard deviation) of testing errors in mean regression over 100 replicates by different approaches with Gaussian errors. “RMSE” and “ADE” indicate the root mean squared errors and the absolute deviation errors, respectively.	45
2.4	Average (standard deviation) of check loss errors in quantile regression over 100 replicates by different approaches with Gaussian errors.	46

2.5	Average (standard deviation) of variable selection results over 100 replicates by different approaches, for $N = 300$, $p = 15$, and Gaussian errors. “NV”, “NVT”, “NL”, and “NLT” indicate the number of nonzero covariates selected, the number of nonzero covariates selected which are truly informative, the number of linear covariates selected, and the number of linear covariates selected which are truly linear, respectively.	51
2.6	House price data. Variable selection results by PL-Tw.	55
2.7	House price data. Variable selection results by different approaches.	62
2.8	Concrete compressive strength data. Variable selection results by PL-Tw.	65
2.9	Concrete compressive strength data. Variable selection results by different approaches.	65
3.1	Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates by different approaches for $C = 0.2$	82
3.2	Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates by different approaches for $C = 0.5$	83
3.3	Average (standard deviation) of negative log partial likelihood on an independent testing dataset over 100 replicates, by different approaches for different scenarios.	90

3.4	Average (standard deviation) of variable selection results over 100 replicates by different approaches for $C = 0.2$. “NV”, “NVT”, “NL”, and “NLT” indicate the number of nonzero covariates selected, the number of nonzero covariates selected which are truly informative, the number of linear covariates selected, and the number of linear covariates selected which are truly linear, respectively.	94
3.5	Average (standard deviation) of variable selection results over 100 replicates by different approaches for $C = 0.5$. “NV”, “NVT”, “NL”, and “NLT” indicate the number of nonzero covariates selected, the number of nonzero covariates selected which are truly informative, the number of linear covariates selected, and the number of linear covariates selected which are truly linear, respectively.	95
3.6	DLBCL data. Variable selection results by different approaches. .	98
4.1	Comparison of two definitions of degrees of freedom. Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates, for $N = 300$, $p = 15$, and Gaussian errors.	114
4.2	Comparison of two definitions of degrees of freedom. Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates, for $N = 300$, $p = 15$, and Gaussian errors. (Continued)	115
4.3	Comparison of two definitions of degrees of freedom. Average (standard deviation) of testing errors in mean regression over 100 replicates, for $N = 300$, $p = 15$, and Gaussian errors. “RMSE” and “ADE” indicate the root mean squared errors and the absolute deviation errors, respectively.	116

4.4	Comparison of two definitions of degrees of freedom. Average (standard deviation) of check loss errors in quantile regression over 100 replicates, for $N = 300$, $p = 15$, and Gaussian errors. . .	116
4.5	Comparison of two definitions of degrees of freedom. Average (standard deviation) of variable selection results over 100 replicates, for $N = 300$, $p = 15$, and Gaussian errors. “NV”, “NVT”, “NL”, and “NLT” indicate the number of nonzero covariates selected, the number of nonzero covariates selected which are truly informative, the number of linear covariates selected, and the number of linear covariates selected which are truly linear, respectively.	117
4.6	Comparison of twin boosting and stability selection. Average (standard deviation) of variable selection results over 100 replicates, for $N = 300$, $p = 15$, and Gaussian errors, mean regression only. “NV”, “NVT”, “NL”, and “NLT” indicate the number of nonzero covariates selected, the number of nonzero covariates selected which are truly informative, the number of linear covariates selected, and the number of linear covariates selected which are truly linear, respectively.	121
4.7	Comparison of twin boosting and stability selection. Average (standard deviation) of variable selection results over 100 replicates, for $N = 300$, $p = 15$, and Student’s t errors, mean regression only. “NV”, “NVT”, “NL”, and “NLT” indicate the number of nonzero covariates selected, the number of nonzero covariates selected which are truly informative, the number of linear covariates selected, and the number of linear covariates selected which are truly linear, respectively.	122

A.1	Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates by different approaches, for $N = 300$, $p = 30$, and Gaussian errors. .	132
A.2	Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates by different approaches, for $N = 300$, $p = 30$, and Gaussian errors. (Continued)	133
A.3	Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates by different approaches, for $N = 500$, $p = 15$, and Gaussian errors. .	134
A.4	Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates by different approaches, for $N = 500$, $p = 15$, and Gaussian errors. (Continued)	135
A.5	Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates by different approaches, for $N = 500$, $p = 30$, and Gaussian errors. .	136
A.6	Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates by different approaches, for $N = 500$, $p = 30$, and Gaussian errors. (Continued)	137
A.7	Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates by different approaches, for $N = 300$, $p = 15$, and Student's t errors.	138
A.8	Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates by different approaches, for $N = 300$, $p = 15$, and Student's t errors. (Continued)	139
A.9	Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates by different approaches, for $N = 300$, $p = 30$, and Student's t errors.	140
A.10	Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates by different approaches, for $N = 300$, $p = 30$, and Student's t errors. (Continued)	141
A.11	Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates by different approaches, for $N = 500$, $p = 15$, and Student's t errors.	142

A.12 Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates by different approaches, for $N = 500$, $p = 15$, and Student's t errors. (Continued)	143
A.13 Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates by different approaches, for $N = 500$, $p = 30$, and Student's t errors.	144
A.14 Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates by different approaches, for $N = 500$, $p = 30$, and Student's t errors. (Continued)	145
A.15 Average (standard deviation) of testing errors in mean regression over 100 replicates by different approaches with Student's t errors. "RMSE" and "ADE" indicate the root mean squared errors and the absolute deviation errors, respectively.	146
A.16 Average (standard deviation) of check loss errors in quantile re- gression over 100 replicates by different approaches with Stu- dent's t errors.	147
A.17 Average (standard deviation) of variable selection results over 100 replicates by different approaches, for $N = 300$, $p = 30$, and Gaussian errors.	148
A.18 Average (standard deviation) of variable selection results over 100 replicates by different approaches, for $N = 500$, $p = 15$, and Gaussian errors.	149
A.19 Average (standard deviation) of variable selection results over 100 replicates by different approaches, for $N = 500$, $p = 30$, and Gaussian errors.	150
A.20 Average (standard deviation) of variable selection results over 100 replicates by different approaches, for $N = 300$, $p = 15$, and Student's t errors.	151

A.21 Average (standard deviation) of variable selection results over 100 replicates by different approaches, for $N = 300$, $p = 30$, and Student's t errors.	152
A.22 Average (standard deviation) of variable selection results over 100 replicates by different approaches, for $N = 500$, $p = 15$, and Student's t errors.	153
A.23 Average (standard deviation) of variable selection results over 100 replicates by different approaches, for $N = 500$, $p = 30$, and Student's t errors.	154

List of Figures

2.1	Boxplots of $\sqrt{\widehat{ISE}_{nonlinear}}$ as well as $\sqrt{\widehat{ISE}_{linear}}$ over 100 replicates by different approaches, for $N = 300$, $p = 15$, and Gaussian errors.	42
2.2	Boxplots of $\sqrt{\widehat{ISE}_{nonlinear}}$ as well as $\sqrt{\widehat{ISE}_{linear}}$ over 100 replicates by different approaches, for $N = 300$, $p = 15$, and Gaussian errors. (Continued)	43
2.3	Boxplots of testing errors in mean regression over 100 replicates by different approaches, for $N = 300$, $p = 15$, and Gaussian errors. “RMSE” and “ADE” indicate the root mean squared errors and the absolute deviation errors, respectively.	47
2.4	Boxplots of check loss errors in quantile regression over 100 replicates by different approaches, for $N = 300$, $p = 15$, and Gaussian errors.	48
2.5	House price data. Estimated component functions by PL-Tw. . .	56
2.6	House price data. Estimated component functions by PL-St. . .	57
2.7	House price data. Estimated component functions by NI-Tw. . .	58
2.8	House price data. Estimated component functions by PL-mb. . .	59

2.9	House price data. Estimated component functions by mgcv. . . .	60
2.10	House price data. Estimated component functions by mcmc. . .	61
2.11	Concrete compressive strength data. Estimated component functions by PL-Tw.	66
2.12	Concrete compressive strength data. Estimated component functions by PL-St.	67
2.13	Concrete compressive strength data. Estimated component functions by NL-Tw.	68
2.14	Concrete compressive strength data. Estimated component functions by PL-mb.	69
2.15	Concrete compressive strength data. Estimated component functions by mgcv.	70
2.16	Concrete compressive strength data. Estimated component functions by mcmc.	71
3.1	Boxplots of $\sqrt{\widehat{ISE}_{nonlinear}}$, $\sqrt{\widehat{ISE}_{linear}}$, as well as $\sqrt{\widehat{ISE}_{non-informative}}$ over 100 replicates by different approaches, for $p = 20$ and $C = 0.2$. 84	
3.2	Boxplots of $\sqrt{\widehat{ISE}_{nonlinear}}$, $\sqrt{\widehat{ISE}_{linear}}$, as well as $\sqrt{\widehat{ISE}_{non-informative}}$ over 100 replicates by different approaches, for $p = 100$ and $C = 0.2$. 85	
3.3	Boxplots of $\sqrt{\widehat{ISE}_{nonlinear}}$, $\sqrt{\widehat{ISE}_{linear}}$, as well as $\sqrt{\widehat{ISE}_{non-informative}}$ over 100 replicates by different approaches, for $p = 20$ and $C = 0.5$. 86	
3.4	Boxplots of $\sqrt{\widehat{ISE}_{nonlinear}}$, $\sqrt{\widehat{ISE}_{linear}}$, as well as $\sqrt{\widehat{ISE}_{non-informative}}$ over 100 replicates by different approaches, for $p = 100$ and $C = 0.5$. 87	
3.5	Estimated functions by PL-Tw, for $N = 200$, $p = 20$, and $C = 0.2$. In each panel, light blue dot-dash curves correspond to estimates in 100 independent replicates, and the red solid curve indicates the underlying true function.	88

3.6	Boxplots of negative log partial likelihood on an independent testing dataset over 100 replicates by different approaches for $C = 0.2$.	91
3.7	Boxplots of negative log partial likelihood on an independent testing dataset over 100 replicates by different approaches for $C = 0.5$.	92
3.8	DLBCL data. Estimated component functions by PL-Tw.	99
3.9	DLBCL data. Estimated component functions by PL-St.	100
3.10	DLBCL data. Estimated component functions by NL-Tw.	101
3.11	DLBCL data. Estimated component functions by PL-mb.	102
3.12	DLBCL data. Estimated component functions by mgcv.	103
4.1	Overfitting. Error to minimum error ratios over iterating. Training errors against testing errors.	118

Chapter 1

Introduction

1.1 The Models

1.1.1 Generalized Additive Model

Additive models are commonly used in the field of statistical learning to study the relationships between predictors and responses. The simplest version of additive model might be the linear additive model, in which it is assumed that

$$Y = \beta_0 + X^T \beta + \epsilon,$$

with X being the predictors, Y being the response, β being the linear coefficients, β_0 being the intercept, and ϵ being the error term and following a centered Gaussian distribution. This model is still widely in use in numerous fields like environmental researches (Khan et al. (2013)), biological researches (Wu et al. (2014)), and chemical researches (Geiß et al. (2015)).

Although it can be convenient in computations and interpretations, there have been concerns that a linear additive model may not be always appropriate. A major one of the concerns is that the response might not, even

approximately or asymptotically, follow a Gaussian distribution. Therefore, a so-called link function $g()$ is introduced to generalize the linear additive model:

$$g(\mu) = \beta_0 + X^T \beta,$$

with $\mu = E[Y|X]$, which thus allows the response to be linked to the predictors in a more general way. One popular example of generalized linear models is the logistic regression model (Kleinbaum and Klein (2010), Allison (2012), Hosmer et al. (2013)) with a binary response Y . In this case, $\mu = E[Y|X] = Pr[Y = 1|X]$, and the log odd $\log(\frac{\mu}{1-\mu}) \triangleq g(\mu)$ is assumed to be linear in predictors. More examples of generalized linear models include the Poisson regression model (Frome (1983)) and the Gamma regression model (Mittlböck and Heinzl (2002)).

Another major concern of linear additive models is that, as widely recognised, it is quite unlikely in the real world that all the predictors contribute to the response variability purely linearly. The nonlinear additive model is thus relatively much more flexible and practical. For instance, Hastie and Tibshirani (1987b) studies a coronary heart disease dataset, and finds that a moderate systolic blood pressure results in a low risk of heart disease, while both low and high blood pressures tend to increase the risk. Holst et al. (1996) and Ruppert and Carroll (2000) reveal, in the studies of a light detection and ranging dataset, that the ratio of received signals at frequencies on and off the resonance frequency of the chemical species of interest is roughly monotonically decreasing in the distance travelled before the light is reflected back to its source, but the slope quite varies for different values of distance.

Combining the two major concerns of the linear additive model, Hastie and Tibshirani (1986) and Hastie and Tibshirani (1990) propose the gener-

alized additive model which assumes

$$g(\mu) = \beta_0 + \sum_{j=1}^p f_j(X_j),$$

where $\mu = E[Y|X]$, $g()$ is the link function, p denotes the number of covariates, and component functions $f_j()$'s can be linear or nonlinear in functional forms. The generalized additive model has been popularly studied in various fields ever since it was proposed. Hastie and Tibshirani (1987a) adapts the generalized additive model to logistic regression in a coronary risk factor study. The model is also applied in Yee and Mitchell (1991) to analyze plant species distributions with respect to the climate. Guisan et al. (2002) uses the nonlinear Poisson regression model to examine the relation of health and air pollution. Wood (2006) applies the nonlinear Gamma regression model on a brain imaging dataset to study the activity levels in different voxels of the brain.

However, sometimes it is some other quantity of the distribution of Y conditional on X , which may not be able to be formularized as a function of $E[Y|X]$, that is of interest. More information on how the response relies on the predictors is learnt by examining such kind of quantities. This concern leads to variants of the generalized additive model. For instance, Rigby and Stasinopoulos (2005) and Mayr et al. (2012a) adopt the generalized additive model to study the scale, the skewness, and the kurtosis of the conditional distribution of the response. Other examples include the quantile regression model and the proportional hazards model in survival analysis. These two models are what this thesis is mainly focusing on. As argued in Scheipl et al. (2013), literatures on partially linear additive models concentrate more on Gaussian or exponential family responses, while extensions to other regression models, such as quantile regressions, survival models, or other

beyond mean regression models, could be relatively lacking.

1.1.2 Partially Linear Additive Model

The core model studied in this thesis is the partially linear additive model (PLAM). PLAM is a special case of the additive model and by “partially linear”, it is emphasized that the coexistence of linear and nonlinear predictors is allowed. PLAM therefore combines the advantages of linear models and nonlinear models. PLAM is actively studied in numerous literatures. For instance, Xia and Härdle (2006) embeds PLAM into the single index models. Liu et al. (2007) connects PLAM with least-squares kernel machines and linear mixed models. Aneiros-Pérez and Vieu (2008) associates PLAM with functional data analysis. Fahrmeir and Lang (2001) develops a Bayesian semiparametric regression technique for multicategorical time-space data. Kammann and Wand (2003) applies PLAM to geostatistics and thus proposes the geoaddivitive models. Interesting real-world applications of PLAM can be found in various works as well. Liang (2006) applies PLAM in a study of the earning with respect to personal characteristics. A linear predictor, the level of schooling a person has completed, and a nonlinear predictor, the number of years spent in the labor market, coexist in the model. Liang et al. (2008) uses PLAM to examine how environmental chemical exposures affect semen quality. In this study of the sperm concentration, linear predictors, the abstinence time and the urinary monobutyl phthalate concentration, as well as nonlinear predictors, the age and the body mass index, are involved. Mayr et al. (2012a) analyzes rental prices in Munich with PLAM. Besides a number of linear covariates, three nonlinear covariates, the size of the flat, the year of construction, and a spatial effect based on the neighbourhood, are included in the model.

Model specification is a research topic strongly connected with PLAM as the strength of PLAM can be realised only when linear and nonlinear covariates can be correctly identified. Model specification in PLAM could be treated as a special case of variable selection. Variable selection is a problem of interest that naturally arises from multivariate analysis. Almost in every field of research, one may have the need to identify truly informative ones from a set of candidate covariates involved in the study. In the early days after the problem of variable selection arose, literatures mainly focused on identifying informative and non-informative covariates. With the drawbacks of all-subsets algorithms (Simonoff and Tsai (1999)) and step-wise selection/deletion procedures (Steyerberg et al. (1999), Wagner and Shimshak (2007)) being widely noticed, penalization methods might have become the most popular category of methods in the topic of variable selection. A penalization method usually achieves a trade-off between the model fit and the model complexity by optimizing a penalized likelihood function. A subset of the candidate predictors are thus left out of the model, while the other covariates are selected as informative. The ridge regression (Hoerl and Kennard (1970)) and the least absolute shrinkage and selection operator (LASSO) (Tibshirani (1996)) are two popular examples of penalization methods which are widely in use. Zou and Hastie (2005) proposes the elastic net which combines the advantages of ridge regression and LASSO. Fan and Li (2001) provides a thorough review of penalization methods and proposes a now commonly used penalty term called the smoothly clipped absolute deviation (SCAD) penalty, which satisfies all the three desired properties of a penalization estimator: unbiasedness, sparsity, and continuity.

PLAM has brought new challenges to variable selection since there are two categories of informative predictors, linear and nonlinear, to be iden-

tified respectively while the detection of non-informative covariates is still required. However, in some of the literatures applying PLAM or the so-called semiparametric additive model (Ruppert et al. (2003), Ruppert et al. (2009)), pre-specification of linear covariates is required. Some examples can be found in Speckman (1988), Yatchew (1997), and He et al. (2005). Zhang et al. (2011) has a good summary of literatures studying PLAM with pre-specified linear covariates while points out that one might need a more principled method to separate linear covariates from the nonlinear ones. Therefore in this thesis, we aim to identify nonlinear, linear, as well as non-informative predictors simultaneously when no prior knowledge, especially no pre-specification of linear components, is provided. Model specification in PLAM with no prior knowledge has recently attracted more attention. Both frequentist (Zhang et al. (2011), Lian et al. (2012)) and Bayesian (Scheipl et al. (2012), Hu et al. (2015)) approaches have been proposed. Scheipl et al. (2013) contains a comprehensive review and provides detailed numerical comparisons of various approaches. Methods involved in this work are mainly penalized likelihood methods, Bayesian methods, as well as boosting methods. Comparisons include the prediction accuracy as well as variable selection performances, while the numerical results find no approach able to dominate the others.

1.1.3 Quantile Regression

Mean regression is the most common and basic topic of regression analysis which estimates the conditional expectation of the response given the predictors. The method of least squares is sometimes seen as the earliest form of mean regression. However quantile regression, aiming to estimate conditional quantile values of the response, offers a more complete view

than mean regression by enabling us to study the entire conditional distribution of the response. Taking the discussions in Fenske et al. (2011) as an instance, mean regression only estimates the contribution of each covariate in the average nutritional status, but we might have more interest in studying the 0.05 or 0.1 quantile of the response which gathers information of the children who are suffering from extreme malnutrition. There are more examples, like researches on wage structures (Buchinsky (1994), Buchinsky (1998)), economic growth studies (Koenker and Machado (1999)), and county homeless researches (Kriegler and Berk (2007)), illustrating the necessity of quantile regression.

In models with responses assumed to follow Gaussian distributions, quantile research can be done by associating mean regression and conditional variance estimation. However there have been much more researches on quantile regression in numerous fields and in various conditions. Koenker (2005) provides overviews of quantile regression in theoretical and computational aspects with a lot of interesting illustrations. Buchinsky (1995) compares methods of estimating asymptotic covariance matrix for quantile regression. Buchinsky and Hahn (1998) studies quantile regression models with censored responses. Koenker (2004) associates quantile regression with longitudinal data. Xiao (2009) applies quantile regression in cointegrated time series models for stock pricing.

One popular choice of the loss function used in quantile regressions to measure the model fit is the so-called check loss function

$$\rho_{\tau}(Y - \hat{Y}) = (Y - \hat{Y})(\tau - \mathbf{1}_{[Y < \hat{Y}]}), \quad (1.1.1)$$

where τ ($0 < \tau < 1$) is the desired quantile level and $\hat{Y}$ is the estimate of quantile value of Y at level τ . The study of this loss function could be

traced back to Fox and Rubin (1964), and it has been widely in use among literatures on quantile regressions. We apply check loss in our study as well, technical details can be found in Chapter 2 of this thesis.

PLAM has been applied to quantile regression as well, where it is assumed that

$$Q_\tau[Y|X] = \beta_0 + \sum_{j=1}^p f_j(X_j),$$

with $Q_\tau[Y|X]$ being the quantile of Y , at level τ , conditional on X . Yee (2004) embeds the method developed by Cole and Green (1992) into vector generalized additive models for multivariate responses. The approach is applied to a cross-sectional workforce study. Bollaerts et al. (2006) develops methods for quantile regression using P-splines, with special attention to the quantile crossing issue. Two medical examples are studied in this work. Takeuchi et al. (2006) shares a similar motivation and compares several approaches on several real-world datasets. Choudhary (2007) uses Bayesian approach in quantile regression, especially for assessing agreement between measurement methods. This approach is illustrated by an oestradiol example and a body fat example. Wang et al. (2009) studies quantile regression for PLAM with varying coefficients and tests the methods on an AIDS dataset.

1.1.4 Survival Analysis

Survival analysis is a topic of statistics whose interest centers on the length of the time period, usually called the survival time or the failure time, until a certain event happens. This event can be various, like the death of a patient due to a certain disease, the breakdown of a machine currently in operation, or the employment of a person that is formerly jobless. Sur-

vival analysis can be applied in numerous fields, especially in biology and medical science as the popularity of biostatistics and gene researches rising nowadays. Some interesting examples of survival analysis can be found in Bender et al. (2005), Cai et al. (2007), and Benner et al. (2010).

Numbers of models have been developed for survival analysis. The Kaplan-Meier method developed in Kaplan and Meier (1958) and the Nelson-Aalen estimator proposed in Nelson (1969) are early techniques in survival analysis with no covariate taken into consideration. Feigl and Zelen (1965) and Glasser (1967) consider several forms of regression models in association with the exponentially distributed survival times. Lin and Ying (1995) and Zhao and Elsayed (2005) study the accelerated life model which assumes that the effect of a covariate is to accelerate or decelerate the survival time by some constant. The frailty model studied in Murphy (1995) and Sastry (1997) is a relatively advanced model with attention to random effects and clustered survival times.

However, nowadays the most popular model in survival analysis might be the Cox's proportional hazards model which is proposed in Cox (1972) and reviewed in Oakes (1981), Cox and Oakes (1984), and Therneau and Grambsch (2000). In this model, the hazard ratio for two subjects with fixed covariates is assumed to be constant over time. Cox's model is studied in numerous literatures. Schoenfeld (1983) considers the sample size needed to test the equality of two survival distributions using Cox's proportional hazards model. Arjas (1988) proposes a graphical method for assessing goodness of fit in proportional hazards model with an illustration of post-transplant survival data. Lin and Wei (1989) derives asymptotic properties of the estimator for the purely parametric version of Cox's model. Fine and Gray (1999) generalizes the proportional hazards model for competing-risks

data where two types of failures are taken into consideration together. This model is illustrated on a tumor dataset with the death and the recurrence being two types of failures. Tsiatis and Davidian (2001) develops estimators for proportional hazards model with longitudinal covariates and involves several simulation studies. Zhang and Lu (2007) associates adaptive LASSO with Cox's model and applies the approach to a cirrhosis dataset. Goeman (2010) also applies the L_1 penalty to proportional hazards model and provides a study on a cancer-related gene expression dataset.

A special difficulty of survival analysis arises from the phenomenon called censoring, which is a form of missing data problem that is frequently encountered in the real world. There are types of censorings, among which the right censoring might be the most common one. Taking the study of the mortality after myocardial infarctions in Lu et al. (2006) as an example, 89 out of 764 involved patients survived the research period thus their survival times are right censored. Studies with censored survival times can also be found in numerous literatures like Peto (1973), Chang and Yang (1987), Lindsey and Ryan (1998), and Klein and Moeschberger (2005), some of which are not restricted to right censoring. Motivated by the right censoring, a partial likelihood function, which is now widely in use, is designed for the Cox's proportional hazards model (Cox (1975), Oakes (1981)). The partial likelihood is essentially the multiplication of a sequence of conditional probabilities, while technical details are to be found in Chapter 3.

There have been literatures on PLAM in survival analysis as well. Cai et al. (2002) proposes a mixed model-based method to estimate hazard functions using penalized splines and interestingly applies the method to a cricket dataset. This approach is extended in Cai and Betensky (2003) to possibly right- or interval-censored survival times. A breast cosmesis

dataset and a hemophilia dataset are used to illustrate. Tutz and Binder (2004) studies semiparametric survival models with time-varying coefficients and includes a hospital dataset of schizophrenic disorders. Lambert and Eilers (2005) develops a penalized Poisson regression approach for Bayesian proportional hazards model, and the approach is illustrated with a clinical dataset of ovarian cancer. Kauermann and Khomski (2006) proposes an additive two-way hazards model and provides an application on unemployment data. Brown et al. (2007) uses a partial likelihood approach with penalized splines to recover non-proportional hazards models with dynamic covariate effects and involves a prevention study of coronary heart diseases. In this thesis, an association of Cox's proportional hazards model with PLAM is studied, which assumes the risk score function of covariates

$$f(\mathbf{X}) = \sum_{j=1}^p f_j(X_j),$$

with each of the component function $f_j()$ being possibly nonlinear or linear in functional form. Technical details are presented in Chapter 3 of this thesis.

1.2 The Boosting

Boosting, as an English word means increasing or improving, meanwhile as a term in statistics and machine learning, indicates a family of algorithms which convert weak learners to strong ones. Boosting originates from a question posed in Kearns (1988) and Kearns and Valiant (1989), that whether a set of weak learners are able to create a strong one. In the early days after the idea of boosting was introduced, discussions were restricted to classification problems. Thus a weak learner was defined to be

a classifier which was only slightly correlated with the true classification, in other words, was only slightly better than a random guessing. In contrast, a strong learner was a classifier that was arbitrarily well-correlated with the true classification. Schapire (1990) gives an affirmative answer to that question, and is seen to have led to the invention of boosting. More early studies of boosting include Drucker et al. (1994) and Freund (1995). Freund and Schapire (1995) and Freund and Schapire (1997) develop a new version of boosting called adaptive boosting, which adjusts adaptively to the errors of the weak learners. These works prove that adaptive boosting does turn weak learners into strong ones with smaller misclassification errors, meanwhile take a first step to extend boosting to regression problems. Jiang (2004) and Bartlett and Traskin (2007) investigate the consistency of adaptive boosting. Numerical reviews of this approach can be also found in Freund and Schapire (1996). Moreover, the adaptive resampling and combining (ARC) techniques (Breiman (1997), Breiman (1998), Breiman (1999)) are seen more or less synonymous with boosting.

1.2.1 Gradient Boosting

Gradient boosting is a statistical extension of boosting for both regression and classification problems. The idea of gradient boosting originates in the discussions in Breiman (1997) that boosting can be interpreted as an optimization algorithm on a suitable loss function. Mason et al. (1999) reiterates this argument and proposes abstract boosting algorithms which optimize a loss function in some function space by iteratively choosing a weak learner. Convergences are proved for these abstract algorithms, however this work is again restricted to classifications. Explicit gradient boosting algorithm is proposed in Friedman (2001) and Friedman (2002), which is an

iterative method optimizing a pre-specified proper loss function, based on combining weak learners. Nowadays, a weak learner, which is also called a base learner sometimes, can be any statistical estimation procedure which fits into a regression framework. At each iteration, gradient boosting selects a single base learner and makes a small update to the estimate. Technical details of gradient boosting, including choices of loss functions and base learners as well as iterating procedures, can be found in following chapters. Theoretical justification of gradient boosting approaches roughly includes the following four core points:

1. The regression model is formularized as an optimization problem in some function space

$$\min_{f(\cdot)} E_{Y,X}[L(Y, f(X))],$$

with X being the predictors, Y being the response, and $L(\cdot)$ being a pre-specified appropriate loss function.

2. To numerically implement the optimization, the gradient descent method of numerical optimization is adopted.
3. By the idea of boosting, the optimization is implemented by repeatedly selecting a base learner and making an update to the estimate. Selection in each iteration is made according to how base learners fit to the gradient direction in the sense of least squares.
4. Due to finite sample sizes, involved expectations are replaced by their sample-based analogs.

All these points are technically reasonable and with theoretical justifications to some extent respectively. However in an overall sense, asymptotic theories for gradient boosting estimators are relatively lacking. More reviews of

gradient boosting can be found in numerous works such as Friedman et al. (2000), Bühlmann and Yu (2003), and Bühlmann and Hothorn (2007).

1.2.2 Quantile Regression with Boosting

Application of boosting methods in quantile regressions does not seem to be an extremely popular topic so far. However related literatures can still be found. Kriegler and Berk (2007) suggests an interesting idea that studies of quantiles are needed in mean regression if costs of underestimates and overestimates are not even. It is one of the early works applying gradient boosting to quantile regressions. Fenske et al. (2011) studies semiparametric additive models in quantile regression with gradient boosting, however the detection of purely linear covariates is not paid attention to. An interesting application of risk factor study for severe childhood malnutrition in India is involved in this work as well as Fenske et al. (2013a). And an extension of this work to longitudinal data can be found in Fenske et al. (2013b). Zheng (2011) introduces a smooth approximation of the check loss function (1.1.1) for quantile regression, noticing the existence of a singular point $\hat{Y} = Y$. Gradient boosting is applied in this work as well. Mayr et al. (2012b) uses quantile regression with boosting to construct prediction intervals and applies the method in a study of childhood body mass index. Covariates involved in this study are pre-classified into linear and nonlinear categories. Zheng (2012) develops quantile boosting algorithms and extends the approaches to binary classification problems by introducing a hidden continuous variable. The methods are tested on several real-world datasets. However when it comes to variable selection in this work, only detections of informative and non-informative covariates are taken into consideration.

In Chapter 2 of this thesis, we consider PLAM in both mean regression

and quantile regression. Gradient boosting algorithms are developed for component function estimation as well as model specification with no prior knowledge on the model structure.

1.2.3 Survival Analysis with Boosting

Boosting methods have been applied to survival analysis as well for various models. Hothorn et al. (2006) introduces a loss function for survival model by applying the inverse probability of censoring (IPC) weights (van der Laan and Robins (2003)), and applies the gradient boosting approach in an acute myeloid leukemia study and a node-positive breast cancer study. Binder and Schumacher (2008) considers high-dimensional linear survival models with special attention to the existence of mandatory covariates. A gradient boosting algorithm is developed and illustrated in a gene study. Lu and Li (2008) applies gradient boosting to nonlinear transformation models with censored survival data. A study of breast cancer is involved in this work. Schmid and Hothorn (2008) investigates the accelerated failure time models and adopts gradient boosting for the estimation of a model-specific scale parameter. Mayr and Schmid (2014) focuses on derivation and evaluation of biomarker combinations. Gradient boosting approach is developed to estimate the concordance index of survival data.

Boosting for Cox's proportional hazards model has been studied as well. Ridgeway (1999) proposes the idea to use the negative partial likelihood as loss function in researches of proportional hazards model. Li and Luan (2005) reiterates this idea and develops a gradient boosting approach for the nonlinear additive version of Cox's model. In Chapter 3 of this thesis, PLAM is associated with Cox's proportional hazards model to study relationships between the distribution of survival time and a set of predictors.

Again, nonlinear, linear, and non-informative covariates are to be detected simultaneously. The proposed gradient boosting approach is designed for partially censored data by applying the negative partial likelihood as loss function.

1.2.4 Boosting for PLAM

Variable selection is automatically achieved by boosting when covariates are associated with base learners and some of the base learners are never selected during the iterating procedure of boosting. For instance in linear regression with gradient boosting, naturally each of the base learners is just the univariate linear regression approach for each of the covariates. A predictor is detected as informative only if its associated base learner is ever selected during the iterating procedure of boosting. Some examples can be found in Bühlmann (2006), González-Recio et al. (2010), and Champion et al. (2014).

There have been literatures applying boosting methods to PLAM, however some of these works require pre-specification of linear covariates. Tutz and Reithinger (2007) develops a boosting approach for semiparametric mixed models with pre-specified linear predictors and involves two real data studies. In the AIDS cohort study, recreational drug use and number of sex partners are grouped as linear covariates while years since seroconversion, age at seroconversion, and CESD, a mental illness score, are grouped as nonlinear ones. In the infant weight study, the age of infant and the age of mother are used as nonlinear covariates while other features involved are all treated to be with linear effects only. Mayr et al. (2012b) develops boosting approaches for construction of prediction intervals and includes a clinical study to predict the future body mass index of a child based on data

collected in the first two years after born. More than ten covariates are involved in this study with pre-specification of the linear ones. There are also works on semiparametric model with boosting, which do not focus on the existence of purely linear covariates at all. Tutz and Binder (2006) investigates boosting for generalized additive models and applies the approach to a study of the number of readmissions of psychiatric patients in Germany. Twelve predictors, including the age and several medical test scores, are all involved as nonlinear features in this study. Algorithms developed are compared with some other approaches by Binder and Tutz (2008) in a number of simulation studies.

Kneib et al. (2009) is the first to note that boosting can be adapted to model specification of PLAM without prior knowledge on model structure, by supplying separate base learners for the linear and nonlinear parts of a covariate influence, which also includes a study of habitat suitability for breeding bird communities. Each of the component function is split into a linear part and a nonlinear part. The linear part is associated with a simple linear regression base learner, while the nonlinear part is associated with a univariate penalized spline base learner. A predictor is detected as non-informative if both associated base learners are never selected in boosting procedure, while it is selected as purely linear if only the associated linear base learner is selected but not the nonlinear one. This point is reiterated in a few literatures. Hothorn et al. (2010) is sort of a manual of R package *mboost* which implements model-based boosting algorithms. This package allows users to involve separate base learners for each of predictors. Scheipl et al. (2013) investigates variable selection for types of generalized additive models. It reviews and compares several approaches, including boosting, in a number of numerical studies. Mayr et al. (2014) reviews

methodological developments of statistical boosting with respect to variable selection, different types of predictor effects, as well as extensions to more types of regression settings.

Since now each covariate is supplied with two base learners, corresponding to linear and nonlinear influences respectively, it is important that the boosting procedure is unbiased in selection without favoring one set of base learners over the other. The effect of bias is investigated in Hofner et al. (2011) in a more general context and it is noted that the degrees of freedom for the nonlinear parts need to be comparable with those for the linear parts. In ordinary multiple linear regressions, the estimated response vector $\hat{Y}$, although with length equal to the sample size, is in fact constrained in a vector space with dimension equal to the number of covariates. The idea of degrees of freedom is proposed for more complex regression frameworks, such as penalized linear regressions and nonlinear regressions, by analogy of ordinary linear regressions to measure the “effective number of parameters”. There exist multiple definitions of degrees of freedom (Buja et al. (1989)), all based on the so-called hat matrix, to the best of our knowledge. Technical details and discussions of degrees of freedom, including a quick numerical comparison of different definitions, are to be found in following chapters.

However, although it is demonstrated to be competitive in estimation by various empirical investigations, standard gradient boosting approach does not perform well as for variable selection or model identification. Base learners associated with informative covariates might be selected early and frequently during the iterating procedure of gradient boosting, but it is unlikely that base learners associated with non-informative components are never selected. This empirical property of standard boosting always re-

sults in small but non-zero estimates for the component functions of non-informative covariates, which is a well-known shortcoming of boosting. For example, Fenske et al. (2011) involves a simulation study with 24 covariates among which 4 are truly informative. Variable selection is only assessed with respect to first selection iteration (FSI) and proportion of selection iterations (PSI). Although empirically, truly informative covariates have distinctly smaller FSIs and larger PSIs, non-informative ones are rarely left out of the estimated model. Li and Luan (2005) conducts a simulation study with 50 predictors, 4 informative and 46 non-informative respectively. It introduces a measurement of the importance of a covariate in the estimated model, based on estimated component functions. Variable selection is only evaluated in terms of that truly informative covariates have relatively greater importance than non-informative ones. But again, estimated component functions for non-informative covariates are hardly zero.

1.2.5 Twin Boosting

Therefore it is proposed that twin boosting (Bühlmann and Hothorn (2010)) method can be applied to achieve better model identification. The core idea of twin boosting is to associate each base learner with an appropriately chosen weight, which can encourage the selection of informative components while suppress the selection of non-informative ones. A few literatures can be found investigating performances of twin boosting. Wang (2011) studies disease classification with high-dimensional data. A weighted hinge loss with misclassification costs incorporated is optimized by boosting. Twin boosting is embedded for a sparser model. He et al. (2015) investigates omics data with label noise and applies twin boosting for feature selection. Zhao (2015) compares several approaches including twin boosting on vari-

able selection with a number of numerical studies. Technical details of twin boosting are to be found in following chapters.

Another method that has been associated with boosting in order to improve variable selection performance is the stability selection (Meinshausen and Bühlmann (2010), Hofner et al. (2014)), which can potentially be adopted in our proposed algorithms as well. However, this approach has its own drawbacks and related discussions are put in Chapter 4.

1.3 Thesis Outline

To sum up, although maybe not as popular as the penalization methods or the Bayesian methods, gradient boosting has shown its strength in numerous literatures. Meanwhile works on PLAM in beyond mean regression models are relatively lacking. Therefore in this thesis, we aim to combine the two topics PLAM and gradient boosting, by developing boosting approaches for model recovery of PLAM. Algorithms are proposed for mean regression, quantile regression, as well as survival analysis. Component functions are estimated, and simultaneously, nonlinear, linear, and non-informative covariates are detected without prior knowledge on model specifications. Twin boosting is incorporated to improve variable selection. Simulations as well as real-world numerical studies are involved to illustrate the proposed algorithms.

The rest of the thesis is organized as follows. In Chapter 2, after reviewing the partially linear additive model, the details of gradient boosting method adapted to PLAM in the context of mean and quantile regressions are explained. How the twin boosting method is incorporated is introduced as well. Chapter 2 also contains the simulation studies and the analyses

of two real-world datasets to further illustrate the proposed approaches. Chapter 3 associates PLAM with the Cox's proportional hazards model, and develops a twin boosting approach for partially censored data. Simulation illustrations and real data applications are as well supplied in Chapter 3. The thesis is concluded in Chapter 4 with some discussions.

Chapter 2

Mean and Quantile Regressions with Boosting

2.1 Methodology

2.1.1 Partially Linear Additive Model

In this chapter, the partially linear additive model (PLAM), which includes p predictors $\{X_j\}_{j=1}^p$ and a univariate response Y , is studied for mean regression as well as quantile regression. For mean regression, it is assumed that the conditional expectation of Y given $\mathbf{X}$ can be expressed by

$$E[Y|\mathbf{X}] = f(\mathbf{X}) = \sum_{j=1}^p f_j(X_j).$$

While in the context of quantile regression, it is assumed that the conditional quantile of Y , at any given quantile level τ ($0 < \tau < 1$), given $\mathbf{X}$ can be expressed by

$$Q_\tau[Y|\mathbf{X}] = f_\tau(\mathbf{X}) = \sum_{j=1}^p f_{\tau,j}(X_j).$$

Here it is worth mentioning that we may omit the subscript “ τ ” of f_τ and $f_{\tau,j}$, throughout in this thesis for the simplicity of notations. However it is not assumed that $f = f_\tau$ for any value of τ or that $f_{\tau_1} = f_{\tau_2}$ for any pair of τ_1 and τ_2 . In PLAM, each f_j can be nonlinear, linear, or zero in functional form. In literatures, the linear covariates of PLAM used to be pre-specified, for example in Speckman (1988), Yatchew (1997), and He et al. (2005). This raises the question that whether there is a more principled way to identify the linear covariates without prior knowledge. The importance of this question has been addressed in some recent works including Zhang et al. (2011) and Scheipl et al. (2012), and both frequentist and Bayesian approaches have been proposed. However, literatures on PLAM with boosting, especially those which do not require prior model specifications, are relatively lacking. Given boosting is an important and popular approach in both statistics and machine learning, we feel that it is interesting to develop a model identification approach based on boosting for PLAM. Given a sample with size N , $\{(\mathbf{X}_i, Y_i)\}_{i=1}^N$, the purpose here is to estimate the underlying component functions $\{f_j\}_{j=1}^p$, and simultaneously identify which of the f_j ’s are nonlinear, linear, or zero, respectively. The motivation for identifying linear and zero components in an additive model comes from the desire to have a parsimonious model which is generally more interpretable.

Considering the task of variable selection and model identification, the component functions $\{f_j\}_{j=1}^p$ in PLAM are described by introducing a decomposition as follows. For each $j = 1, \dots, p$, it is written that

$$f_j = \beta_{0,j} + g_j + h_j, \quad (2.1.1)$$

where $\beta_{0,j}$ is a constant, g_j is a linear function and h_j is a nonlinear function. Moreover, $\beta_{0,j}$, g_j , and h_j are restricted to be pair-wise orthogonal. A

similar decomposition is used in Zhang et al. (2011). By the decomposition, we have

$$f(\mathbf{X}) = \beta_0 + \sum_{j=1}^p g_j(X_j) + \sum_{j=1}^p h_j(X_j),$$

where $f(\mathbf{X}) = E[Y|\mathbf{X}]$ or $Q_\tau[Y|\mathbf{X}]$, and $\beta_0 = \sum_{j=1}^p \beta_{0,j}$. Denote the corresponding estimates by

$$\hat{f}(\mathbf{X}) = \hat{\beta}_0 + \sum_{j=1}^p \hat{g}_j(X_j) + \sum_{j=1}^p \hat{h}_j(X_j).$$

X_j is identified as a nonlinear predictor if and only if $\hat{h}_j \neq 0$; X_j is identified as a linear predictor if and only if $\hat{h}_j = 0$ but $\hat{g}_j \neq 0$; X_j is identified as a non-informative predictor if and only if $\hat{g}_j = 0$ and $\hat{h}_j = 0$. Each covariate is standardized, using linear transformation, to be supported in $[0, 1]$ before the proposed method is applied. The rest of the section gives details on the estimation of f for both mean regression and quantile regression.

2.1.2 Base Learners

Boosting is a method taking some base learners as its builders. In our context, a base learner, also called a weak learner, can be any statistical estimation procedure which fits into a regression framework. To achieve the task of model structure recovery of PLAM, the boosting approaches proposed here apply two types of base learners: univariate least squares linear regression, which is used to build estimates of linear functions $\{g_j\}_{j=1}^p$, and univariate penalized spline regression, which is used to build estimates of nonlinear functions $\{h_j\}_{j=1}^p$.

Simple univariate linear regression solves the minimization problem

$$\min_{\beta} \sum_{i=1}^N (Y_i - \beta X_{ij})^2,$$

given data $\{(\mathbf{X}_i, Y_i)\}_{i=1}^N$, for any $j = 1, \dots, p$.

Univariate regression with splines (de Boor (1978), Wahba (1990), Eubank (1999)) is more flexible than linear regression. We start with truncated power splines (TPS). For a given degree q and a sequence of knots $\{t_l\}_{l=1}^k$ inside $[0, 1]$, the functions $\tilde{B}_{-1}(x) \equiv 1$, $\tilde{B}_0(x) = x$, $\tilde{B}_1(x) = x^2$, ..., $\tilde{B}_{q-1}(x) = x^q$, $\tilde{B}_q(x) = (x - t_1)^q \mathbf{1}_{[x > t_1]}$, ..., $\tilde{B}_{q+k-1}(x) = (x - t_k)^q \mathbf{1}_{[x > t_k]}$ form the TPS basis. The using of TPS basis is rationalized by Proposition 2.1:

Proposition 2.1. With a given degree q and a sequence of knots $\{t_l\}_{l=1}^k$ inside $[0, 1]$,

$$\begin{aligned} \tilde{B}_{-1}(x) &\equiv 1, \\ \tilde{B}_0(x) &= x, \\ \tilde{B}_1(x) &= x^2, \\ &\dots, \\ \tilde{B}_{q-1}(x) &= x^q, \\ \tilde{B}_q(x) &= (x - t_1)^q \mathbf{1}_{[x > t_1]}, \\ &\dots, \\ \tilde{B}_{q+k-1}(x) &= (x - t_k)^q \mathbf{1}_{[x > t_k]} \end{aligned}$$

form a basis of $\Pi_{q, \{t_l\}_{l=1}^k}$, where $\Pi_{q, \{t_l\}_{l=1}^k}$ denotes the function space formed by all piecewise polynomial functions, with break sequence $\{t_l\}_{l=1}^k$, which are of degree no higher than q and have a global continuous $(q-1)$ -th derivative in $[0, 1]$.

Proof of this proposition can be found in de Boor (1978).

Recalling the decomposition (2.1.1), as we are using splines to estimate the nonlinear component of each covariate, which is orthogonal to the linear component, we accordingly modify the TPS basis by removing $\tilde{B}_{-1}(x)$ and $\tilde{B}_0(x)$, and orthogonalizing $\tilde{B}_1(x)$, ..., $\tilde{B}_{q+k-1}(x)$ to $\tilde{B}_{-1}(x)$ and $\tilde{B}_0(x)$, as described in Proposition 2.2:

Proposition 2.2. Let X be any random variable supported in $[0, 1]$. q , k , and $\tilde{B}_l(x)$ ($l = -1, 0, 1, \dots, (q + k - 1)$) are defined in Proposition 2.1. $B_m(x) = \tilde{B}_m(x) + a_mx + b_m$ with

$$\begin{aligned} a_m &= \frac{E[X]E[X^{m+1}] - E[X^{m+2}]}{E[X^2] - E[X]^2}, \\ b_m &= \frac{E[X]E[X^{m+2}] - E[X^2]E[X^{m+1}]}{E[X^2] - E[X]^2}, \end{aligned}$$

$m = 1, \dots, (q - 1)$. $B_n(x) = \tilde{B}_n(x) + a_nx + b_n$ with

$$\begin{aligned} a_n &= \frac{E[X]E[Z_n] - E[XZ_n]}{E[X^2] - E[X]^2}, \\ b_n &= \frac{E[X]E[XZ_n] - E[X^2]E[Z_n]}{E[X^2] - E[X]^2}, \\ Z_n &= (X - t_{n-q+1})^q \mathbf{1}_{[X > t_{n-q+1}]}, \end{aligned}$$

$n = q, \dots, (q + k - 1)$. Then

$$\begin{aligned} E[\tilde{B}_{-1}(X)B_l(X)] &= 0, \\ E[\tilde{B}_0(X)B_l(X)] &= 0, \end{aligned}$$

$l = 1, \dots, (q + k - 1)$.

Proof of this proposition is trivial thus omitted. In practice, TPS basis is modified for each of the covariates, and all expectations involved are replaced by sample averages. Therefore, a_l and b_l , and thus B_l ($l = 1, \dots, (q+k-1)$), are different for each of the covariates. But we may not include more subscripts in spline basis for the simplicity of notations.

It is widely recognized that the flexibility of nonparametric regression may cause the problem of overfitting. Thus here penalized splines are used as base learners. Let

$$\mathbf{\Omega} = \begin{pmatrix} \int_0^1 B_1''(x)B_1''(x)dx & \cdots & \int_0^1 B_1''(x)B_{q+k-1}''(x)dx \\ \vdots & \ddots & \vdots \\ \int_0^1 B_{q+k-1}''(x)B_1''(x)dx & \cdots & \int_0^1 B_{q+k-1}''(x)B_{q+k-1}''(x)dx \end{pmatrix}$$

be the second derivative penalty matrix. For any $j = 1, \dots, p$, given the sample $\{(X_{ij}, Y_i)\}_{i=1}^N$, the penalized spline estimator is obtained by solving the minimization problem

$$\min_{\boldsymbol{\theta}_j} [(\mathbf{Y} - \mathbf{B}_j \boldsymbol{\theta}_j)^T (\mathbf{Y} - \mathbf{B}_j \boldsymbol{\theta}_j) + \lambda_j \boldsymbol{\theta}_j^T \mathbf{\Omega} \boldsymbol{\theta}_j],$$

where $\mathbf{Y} = (Y_1, \dots, Y_N)^T$, λ_j is a tuning parameter, and

$$\mathbf{B}_j = \begin{pmatrix} B_1(X_{1j}) & \cdots & B_{q+k-1}(X_{1j}) \\ \vdots & \ddots & \vdots \\ B_1(X_{Nj}) & \cdots & B_{q+k-1}(X_{Nj}) \end{pmatrix}_{N \times (q+k-1)}.$$

It was previously suggested that λ_j should be chosen such that the degrees of freedom, $\text{trace}(\mathbf{B}_j(\mathbf{B}_j^T \mathbf{B}_j + \lambda_j \mathbf{\Omega})^{-1} \mathbf{B}_j^T)$, equal some small integer, for example 3 (Bühlmann and Hothorn (2007)). However, in our case, the univariate penalized splines and univariate linear regression are both used as base learners. When the degrees of freedom of the penalized spline model are set to be 3, empirically it is observed that the boosting algorithm

will almost always choose the nonlinear base learners over the linear base learners. This is not surprising since the penalized splines provide a better fit to data due to the flexibility compared to the simple linear regression model (noting that the degrees of freedom represent the effective number of parameters in the model). To address this problem, it is proposed to choose λ_j such that

$$\text{trace}(\mathbf{B}_j(\mathbf{B}_j^T \mathbf{B}_j + \lambda_j \mathbf{\Omega})^{-1} \mathbf{B}_j^T) = 1,$$

because, as widely know, simple linear regressions have 1 degree of freedom. In this way, linear base learners are comparable to nonlinear ones. It is worth mentioning that λ_j can be pre-calculated and fixed before running the proposed boosting algorithm.

A slightly different definition of degrees of freedom is used in Hofner et al. (2011), which is argued to be exactly unbiased. Empirically we believe the two are very similar in performance and both are in popular use. A quick numerical comparison between two definitions is put in Chapter 4.

2.1.3 Gradient Boosting

Boosting is originally designed for classification (Freund and Schapire (1995), Freund and Schapire (1997)) and has been successfully extended to regression. Friedman (2001) proposes a statistical framework for boosting, called gradient boosting, that interprets boosting as a stage-wise function estimating method similar to forward regression. With a pre-specified loss function $L(Y, f)$ and given sample $\{(\mathbf{X}_i, Y_i)\}_{i=1}^N$, to find f (in some function space) minimizing

$$\frac{1}{N} \sum_{i=1}^N L(Y_i, f(\mathbf{X}_i)),$$

the sample-based analog of $E_{Y,\mathbf{X}}[L(Y, f(\mathbf{X}))]$, gradient boosting proceeds as follows.

Gradient boosting algorithm

1. Initialize $m = 0$ and $\hat{f}^{[0]} = \mathbf{0}$. Specify a set of base learners.
 2. Compute the negative gradient of loss function at the current estimate $\hat{f}^{[m]}$:
$$U_i^{[m]} = -\frac{\partial}{\partial f} L(Y_i, f)|_{f=\hat{f}^{[m]}(\mathbf{X}_i)}.$$
 3. Apply each of the base learners to data $\{(\mathbf{X}_i, U_i^{[m]})\}_{i=1}^N$, using $U^{[m]} = (U_1^{[m]}, \dots, U_N^{[m]})^T$ as the “pseudo-responses”. Find the base learner that fits the best, and denote the resulting estimate with $g^{[m]}$.
 4. Update $\hat{f}^{[m+1]} = \hat{f}^{[m]} + \nu g^{[m]}$, where ν ($0 < \nu < 1$) is known as the step size. Increase m by 1.
 5. Iterate steps 2~4 until some stopping criterion is met.
-

Gradient boosting adopts the gradient descent method to implement the optimization of $\frac{1}{N} \sum_{i=1}^N L(Y_i, f(\mathbf{X}_i))$. Gradient descent, also known as steepest descent, is a numerical optimization algorithm motivated by the observation that a multivariable objective function decreases the fastest when going from some point in the direction of the negative gradient at this point. Under regularization conditions, convergence to minimum by gradient descent is guaranteed. The gradient descent method can be traced

back to Cauchy (1847) and has been studied by numbers of literatures in the filed of numerical optimization. Examples include Avriel (2003) and Snyman (2005).

In this thesis, gradient boosting approach is adapted to PLAM, $f(\mathbf{X}) = \beta_0 + \sum_{j=1}^p g_j(X_j) + \sum_{j=1}^p h_j(X_j)$. To initialize, it is set that

$$\begin{aligned}\hat{\beta}_0^{[0]} &= 0, \\ \hat{\mathbf{g}}_j^{[0]} &= (\hat{g}_j^{[0]}(X_{1j}), \dots, \hat{g}_j^{[0]}(X_{Nj}))^T = \mathbf{0}, \\ \hat{\mathbf{h}}_j^{[0]} &= (\hat{h}_j^{[0]}(X_{1j}), \dots, \hat{h}_j^{[0]}(X_{Nj}))^T = \mathbf{0},\end{aligned}$$

for $j = 1, \dots, p$. For mean regression, we specify the loss function to be the widely used squared error loss $L(Y, f) = (Y - f)^2$. While in the context of quantile regression, the loss function used is the check loss $\rho_\tau(Y, f) = (Y - f)(\tau - \mathbf{1}_{[Y < f]})$. The using of these two loss functions is rationalized by Proposition 2.3:

Proposition 2.3. Let Y be any random variable. Then

$$\arg \min_{\eta} E_Y[(Y - \eta)^2] = E[Y].$$

If Y is continuous with the cumulative distribution function (CDF) denoted as $F_Y()$ and an arbitrary quantile level τ ($0 < \tau < 1$),

$$\arg \min_{\eta} E_Y[\rho_\tau(Y, \eta)] = F_Y^{-1}(\tau), \text{ the population quantile of } Y \text{ at level } \tau.$$

Proof of this proposition is put in the appendix of this thesis. In consequence, we have the corresponding negative gradients

$$U_i^{[m]} = 2(Y_i - \hat{f}^{[m]}(\mathbf{X}_i))$$

for mean regression and

$$U_i^{[m]} = \tau - \mathbf{1}_{[Y_i < \hat{f}^{[m]}(\mathbf{X}_i)]}$$

for quantile regression.

Let $\mathbf{U}^{[m]} = (U_1^{[m]}, \dots, U_N^{[m]})^T$. Base learners applied in our proposed approach are introduced in the previous subsection. For each of the linear base learners, we calculate

$$\begin{aligned} \Delta \hat{\beta}_j^{[m]} &= \arg \min_{\Delta \beta_j} [(\mathbf{U}^{[m]} - \Delta \beta_j \mathbf{X}_j)^T (\mathbf{U}^{[m]} - \Delta \beta_j \mathbf{X}_j)] \\ &= (\mathbf{X}_j^T \mathbf{X}_j)^{-1} \mathbf{X}_j^T \mathbf{U}^{[m]}, \\ \Delta \hat{\mathbf{g}}_j^{[m]} &= \Delta \hat{\beta}_j^{[m]} \mathbf{X}_j, \end{aligned}$$

where $\mathbf{X}_j = (X_{1j}, \dots, X_{Nj})^T$, and for each of the penalized spline base learners, we calculate

$$\begin{aligned} \Delta \hat{\boldsymbol{\theta}}_j^{[m]} &= \arg \min_{\Delta \boldsymbol{\theta}_j} [(\mathbf{U}^{[m]} - \mathbf{B}_j \Delta \boldsymbol{\theta}_j)^T (\mathbf{U}^{[m]} - \mathbf{B}_j \Delta \boldsymbol{\theta}_j) + \lambda_j \Delta \boldsymbol{\theta}_j^T \boldsymbol{\Omega} \Delta \boldsymbol{\theta}_j] \\ &= (\mathbf{B}_j^T \mathbf{B}_j + \lambda_j \boldsymbol{\Omega})^{-1} \mathbf{B}_j^T \mathbf{U}^{[m]}, \\ \Delta \hat{\mathbf{h}}_j^{[m]} &= \mathbf{B}_j \Delta \hat{\boldsymbol{\theta}}_j^{[m]}, \end{aligned}$$

$j = 1, \dots, p$. To select one of the base learners in each iteration that fits the best, the improvement of fit for each base learner in the current iteration is measured by

$$\mathbf{U}^{[m]T} \mathbf{U}^{[m]} - (\mathbf{U}^{[m]} - \Delta \hat{\mathbf{g}}_j^{[m]})^T (\mathbf{U}^{[m]} - \Delta \hat{\mathbf{g}}_j^{[m]}) \quad (2.1.2)$$

or

$$\mathbf{U}^{[m]T} \mathbf{U}^{[m]} - (\mathbf{U}^{[m]} - \Delta \hat{\mathbf{h}}_j^{[m]})^T (\mathbf{U}^{[m]} - \Delta \hat{\mathbf{h}}_j^{[m]}). \quad (2.1.3)$$

If the linear base learner associated with predictor j^* has the greatest improvement of fit among all the $2p$ (linear and nonlinear) base learners, it is updated that

$$\hat{\mathbf{g}}_{j^*}^{[m+1]} = \hat{\mathbf{g}}_{j^*}^{[m]} + \nu \Delta \hat{\mathbf{g}}_{j^*}^{[m]},$$

if the nonlinear base learner associated with predictor j^* has the greatest improvement of fit, we update $\hat{\mathbf{h}}_{j^*}^{[m]}$ similarly.

For the step size, the commonly suggested value $\nu = 0.1$ is used. Moreover, $\hat{\beta}_0$ is updated with the sample mean or sample quantile of the current residual $\mathbf{Y} - \hat{\mathbf{f}}^{[m]}$ (e.g. after every 10 iterations). For the stopping criterion, in our simulation studies, the method, which is very popular in use, of 5-fold cross validation (Stone (1974), Stone (1977), Efron and Gong (1983), Kohavi (1995)) is applied, treating the number of iterations as the tuning parameter. In the real data illustrations, since the sample sizes are relatively large, parts of the sample sets are reserved as validation samples and the iterating procedure is stopped when the validation error is minimized.

2.1.4 Twin Boosting

The standard boosting algorithm is more powerful in function estimation than in variable selection. Empirically, for an non-informative covariate, gradient boosting usually comes up with an estimated component function which is quite small, but not exactly zero. Base learners associated with informative components may be selected early and frequently during the iterating procedure of boosting, but it is quite unlikely that base learners associated with non-informative components are never selected. To achieve

better variable selection and model identification, the method of twin boosting is applied, which is firstly introduced in Bühlmann and Hothorn (2010). The core idea of twin boosting is to associate each of the base learners with an appropriately chosen weight which can encourage the selection of base learners associated with informative components while suppress the selection of those associated with non-informative components. More specifically, we replace (2.1.2) with the weighted improvement of fit

$$\text{Weight}(g_j) \times [\mathbf{U}^{[m]T} \mathbf{U}^{[m]} - (\mathbf{U}^{[m]} - \Delta \hat{\mathbf{g}}_j^{[m]})^T (\mathbf{U}^{[m]} - \Delta \hat{\mathbf{g}}_j^{[m]})]$$

and similarly for (2.1.3). In consequence a base learner with a larger weight will be selected more often. To decide the weights, a two-round strategy is applied as suggested in Bühlmann and Hothorn (2010):

Twin boosting algorithm

1. Implement the standard gradient boosting, which is equivalent to twin boosting with weight equal to 1 for every base learner, and obtain estimates denoted with $\hat{g}_{o,j}$ and $\hat{h}_{o,j}$ ($j = 1, \dots, p$).
2. Implement a second round of boosting with $\text{Weight}(g_j)$ naturally defined by:

$$\text{Weight}(g_j) = \frac{1}{N} \sum_{i=1}^N \hat{g}_{o,j}^2(X_{ij}),$$

$j = 1, \dots, p$, and similarly define the weight for h_j . Obtain the final estimates denoted with $\hat{g}_j$ and $\hat{h}_j$ ($j = 1, \dots, p$).

2.2 Simulation Illustrations

The performance of proposed boosting approach is evaluated by simulation studies. The simulation model includes a univariate response Y and p predictors, among which 4 are nonlinear, 3 are linear, and the rest are all non-informative. More specifically, $f_1(x) = [\sin(2\pi x) + 2\cos(2\pi x) + 3\sin^2(2\pi x) + 4\cos^2(2\pi x) + 5\sin^3(2\pi x)]/5$, $f_2(x) = -\sin(2\pi x)/[2 - \sin(2\pi x)]$, $f_3(x) = 5x(1-x)$, $f_4(x) = 1/(x+0.5)^2$, $f_5(x) = 2x$, $f_6(x) = -2x$, $f_7(x) = x$, and

$$Y = \sum_{j=1}^7 f_j(X_j) + (0.5 + X_4)\varepsilon.$$

The predictors are generated from the multivariate Gaussian distribution with mean $\mathbf{0}$ and covariance matrix Σ , where $\Sigma_{j_1 j_2} = 0.5^{|j_1 - j_2|}$, and then transformed to be marginally uniform on $[0, 1]$ by applying the cumulative distribution function of the univariate standard normal distribution. Moreover, two different distributions of the noise term ε are used. One is the Gaussian distribution with mean 0 and standard deviation 0.5. The other is the Student's t distribution with 2 degrees of freedom multiplied by a scale parameter 0.2.

The sample size is set as $N = 300$ or 500 . Total number of covariates is set as $p = 15$ or 30 . Degree of truncated power splines is fixed to be $q = 3$ and the internal knots used in TPS are set to be equally spaced $\{0.1, 0.2, \dots, 0.9\}$. Other choices of q and internal knots are discussed in Chapter 4. Step size is fixed to be $\nu = 0.1$. The proposed twin boosting algorithm with linear and nonlinear base learners is implemented for mean regression and quantile regressions at 5 quantile levels $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. Stopping time is decided by 5-fold cross validation. 100 independent replicates are used in each of the settings. For comparison, we also apply

five other approaches (on the same datasets):

1. Standard gradient boosting approach (equal weight 1 for all base learners) with linear and nonlinear base learners. This is to illustrate the power of twin boosting.
2. Twin boosting approach with only penalized spline base learners. This is to illustrate the power of PLAM.
3. The algorithm in existing R package *mboost*. This package implements model-based boosting approaches (Hothorn et al. (2010)) with mean regression and quantile regression both included. It is also allowed, in this package, to supply separate base learners for the linear parts and nonlinear parts of covariate influences as in Scheipl et al. (2013). However *mboost* does not involve twin boosting.
4. The algorithm in existing R package *mgcv*. This package is designed as a mixed generalized additive model computation vehicle mostly using penalization methods (Wood (2006)). To our knowledge, in this package, quantile regression can only be implemented by associating mean regression and conditional variance estimation (with Gaussian assumptions).
5. The algorithm proposed in Hu et al. (2015). This is a Bayesian approach designed for PLAM. Mean regression and quantile regression are both included.

Throughout in this thesis (including following chapters), in presentations of numerical results, “PL-Tw” indicates the proposed twin boosting approach with linear and nonlinear base learners; “PL-St” indicates the

standard gradient boosting approach (equal weight 1 for all base learners) with linear and nonlinear base learners; “NI-Tw” indicates the twin boosting approach with only penalized spline base learners; “PL-mb” indicates the algorithm in package *mboost*, supplying separate base learners for the linear parts and nonlinear parts of covariate influences; “mgcv” indicates the algorithm in package *mgcv*; “mcmc” indicates the Bayesian algorithm proposed in Hu et al. (2015). In the following subsections, the performances of different approaches are investigated in terms of estimation accuracy, prediction accuracy, and variable selection accuracy. For prediction accuracy, only the results for Gaussian errors are presented in the main text while the results for Student’s t errors are left in the appendix. For estimation accuracy and variable selection accuracy, we further restrict presentation of numerical results to the case where $N = 300$ and $p = 15$, while all the other settings are presented in the appendix. The results reported in the main text as well as in the appendix both demonstrate the favorable performance of the proposed approach, especially the ability of model identification for PLAM using twin boosting.

2.2.1 Estimation Accuracy

To measure the estimation accuracies of the methods, the integrated squared error (*ISE*) is calculated, which is defined as

$$ISE_j = \int_0^1 (\hat{f}_j(x) - f_j(x))^2 dx,$$

with the integral approximated on an equidistant sequence $\{t_1, \dots, t_{1000}\}$ inside the interval $[0, 1]$. Tables 2.1 and 2.2 present $\sqrt{\widehat{ISE}}$ of all the 7 nonzero component functions $(f_1, \dots, f_7)$ and the first zero component function (f_8) , when $N = 300$, $p = 15$, and the noises follow the Gaussian distribution.

Results of other cases can be found in the appendix. The average values over 100 replicates are provided together with the standard deviations in the brackets, for both mean regression and quantile regressions at the mentioned five quantile levels. Numerical results of our proposed twin boosting approach as well as the mentioned five approaches for comparison are displayed. It is worth mentioning that because the package *mgcv* can only implement quantile regression by associating mean regression with conditional variance estimation, it is enough to present $\sqrt{\widehat{ISE}}$ by “mgcv” for mean regression only. Figures 2.1 and 2.2 present more visible comparisons among different approaches using boxplots based on the same numerical results. To reduce the number of figures, truly nonlinear covariates are combined as $\widehat{ISE}_{nonlinear} = \sum_{j=1}^4 \widehat{ISE}_j$, as well as the truly linear covariates $\widehat{ISE}_{linear} = \sum_{j=5}^7 \widehat{ISE}_j$. We have made the following observations and comments on the numerical results:

1. Our proposed approach (“PL-Tw”) generally outperforms all the other approaches in mean regression, and is competitive as well in quantile regression.
2. Although twin boosting is originally motivated by searching for a better variable selection, it outperforms standard gradient boosting (“PL-St”) with respect to estimation accuracy.
3. With special attention to the linear covariates, PLAM outperforms the nonlinear model (“NL-Tw”) in estimations of linear component functions, but at the cost of inferiority in estimations of nonlinear component functions sometimes.
4. Our proposed algorithm generally outperforms the existing boosting algorithm in package *mboost*.

5. Our proposed approach performs better than the Bayesian approach “mcmc” in mean regression and quantile regressions at moderate quantile levels (0.3, 0.5, and 0.7). While “mcmc” may be slightly better in quantile regressions at extreme quantile levels (0.1 and 0.9).
6. Twin boosting approaches with either both nonlinear and linear base learners (“PL-Tw”) or only nonlinear base learners (“NL-Tw”) provide smaller $\sqrt{\widehat{ISE}}$ for non-informative covariates ($j = 8, \dots, p$), compared to the other approaches.

2.2.2 Prediction Accuracy

The prediction accuracies of the methods are evaluated by calculating 3 types of testing errors on another independent testing sample with size $N_t = 1500$. For mean regression, the root mean squared errors ($RMSE$)

$$RMSE = \sqrt{\frac{1}{N_t} \sum_{i=1}^{N_t} (\hat{y}_i - y_i)^2},$$

and the absolute deviation errors (ADE)

$$ADE = \frac{1}{N_t} \sum_{i=1}^{N_t} |\hat{y}_i - y_i|,$$

are calculated, while for quantile regression we compute the average check loss errors ($ACLE$)

$$ACLE = \frac{1}{N_t} \sum_{i=1}^{N_t} \rho_\tau(\hat{y}_i - y_i),$$

where y_i ’s are the observed responses in the testing sample and $\hat{y}_i$ ’s are the corresponding predicted values. Testing errors for different scenarios with

Table 2.1: Average (standard deviation) of $\sqrt{ISE}$ over 100 replicates by different approaches, for $N = 300$, $p = 15$, and Gaussian errors.

		f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
Mean	PL-Tw	0.094	0.082	0.055	0.077	0.034	0.040	0.037	0.005
		(0.025)	(0.023)	(0.020)	(0.023)	(0.024)	(0.026)	(0.024)	(0.012)
	PL-St	0.140	0.093	0.059	0.091	0.064	0.069	0.058	0.041
		(0.025)	(0.028)	(0.018)	(0.026)	(0.033)	(0.031)	(0.027)	(0.025)
	NL-Tw	0.099	0.083	0.066	0.079	0.061	0.068	0.054	0.004
		(0.024)	(0.023)	(0.022)	(0.024)	(0.022)	(0.024)	(0.024)	(0.012)
	PL-mb	0.162	0.096	0.057	0.102	0.062	0.059	0.054	0.037
		(0.018)	(0.025)	(0.020)	(0.019)	(0.034)	(0.032)	(0.030)	(0.027)
$\tau = 0.1$	mgcv	0.091	0.088	0.079	0.087	0.047	0.043	0.048	0.028
		(0.024)	(0.029)	(0.029)	(0.024)	(0.039)	(0.031)	(0.030)	(0.040)
	mcmc	0.102	0.101	0.096	0.115	0.058	0.061	0.054	0.007
		(0.024)	(0.030)	(0.020)	(0.017)	(0.031)	(0.028)	(0.028)	(0.027)
	PL-Tw	0.184	0.174	0.078	0.154	0.067	0.084	0.105	0.002
		(0.037)	(0.062)	(0.036)	(0.055)	(0.053)	(0.057)	(0.079)	(0.010)
	PL-St	0.215	0.160	0.094	0.213	0.129	0.118	0.101	0.060
		(0.034)	(0.046)	(0.036)	(0.064)	(0.060)	(0.062)	(0.049)	(0.035)
$\tau = 0.3$	NL-Tw	0.156	0.134	0.095	0.128	0.093	0.114	0.120	0.001
		(0.031)	(0.037)	(0.040)	(0.043)	(0.040)	(0.053)	(0.068)	(0.005)
	PL-mb	0.252	0.204	0.102	0.281	0.130	0.113	0.087	0.052
		(0.043)	(0.045)	(0.047)	(0.064)	(0.070)	(0.066)	(0.049)	(0.043)
	mcmc	0.141	0.165	0.152	0.227	0.065	0.068	0.067	0.017
		(0.053)	(0.085)	(0.090)	(0.103)	(0.034)	(0.037)	(0.043)	(0.044)
	PL-Tw	0.143	0.116	0.056	0.094	0.038	0.044	0.052	0.001
		(0.021)	(0.033)	(0.021)	(0.027)	(0.026)	(0.033)	(0.033)	(0.006)
	PL-St	0.175	0.117	0.066	0.115	0.076	0.076	0.066	0.046
		(0.019)	(0.029)	(0.023)	(0.035)	(0.034)	(0.038)	(0.033)	(0.031)
	NL-Tw	0.119	0.097	0.073	0.101	0.068	0.075	0.068	0.001
		(0.025)	(0.025)	(0.026)	(0.031)	(0.028)	(0.028)	(0.030)	(0.007)
	PL-mb	0.193	0.134	0.077	0.148	0.081	0.070	0.060	0.042
		(0.022)	(0.028)	(0.028)	(0.036)	(0.048)	(0.040)	(0.036)	(0.032)
	mcmc	0.103	0.106	0.103	0.130	0.058	0.061	0.047	0.013
		(0.023)	(0.033)	(0.024)	(0.027)	(0.036)	(0.033)	(0.027)	(0.042)

Table 2.2: Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates by different approaches, for $N = 300$, $p = 15$, and Gaussian errors. (Continued)

		f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
$\tau = 0.5$	PL-Tw	0.136	0.112	0.054	0.086	0.038	0.042	0.048	0.000
		(0.024)	(0.030)	(0.021)	(0.026)	(0.028)	(0.033)	(0.034)	(0.002)
	PL-St	0.172	0.114	0.065	0.100	0.069	0.069	0.064	0.039
		(0.019)	(0.034)	(0.024)	(0.029)	(0.037)	(0.036)	(0.031)	(0.029)
	NL-Tw	0.112	0.097	0.070	0.094	0.069	0.071	0.061	0.001
		(0.027)	(0.028)	(0.023)	(0.028)	(0.025)	(0.026)	(0.028)	(0.006)
	PL-mb	0.189	0.131	0.073	0.121	0.069	0.062	0.056	0.040
		(0.020)	(0.027)	(0.027)	(0.026)	(0.041)	(0.038)	(0.033)	(0.029)
	mcmc	0.100	0.101	0.103	0.118	0.045	0.053	0.047	0.006
		(0.023)	(0.031)	(0.024)	(0.023)	(0.028)	(0.026)	(0.031)	(0.020)
$\tau = 0.7$	PL-Tw	0.142	0.124	0.058	0.088	0.041	0.052	0.050	0.000
		(0.028)	(0.040)	(0.025)	(0.028)	(0.034)	(0.036)	(0.042)	(0.002)
	PL-St	0.175	0.120	0.071	0.099	0.074	0.079	0.062	0.044
		(0.018)	(0.034)	(0.025)	(0.029)	(0.038)	(0.038)	(0.030)	(0.028)
	NL-Tw	0.115	0.102	0.074	0.095	0.069	0.081	0.066	0.001
		(0.029)	(0.031)	(0.025)	(0.029)	(0.026)	(0.026)	(0.036)	(0.006)
	PL-mb	0.197	0.142	0.075	0.112	0.074	0.063	0.059	0.041
		(0.024)	(0.029)	(0.030)	(0.022)	(0.044)	(0.034)	(0.038)	(0.027)
	mcmc	0.107	0.105	0.098	0.116	0.057	0.048	0.045	0.011
		(0.029)	(0.052)	(0.024)	(0.018)	(0.033)	(0.027)	(0.027)	(0.024)
$\tau = 0.9$	PL-Tw	0.185	0.204	0.091	0.116	0.068	0.079	0.086	0.001
		(0.034)	(0.071)	(0.039)	(0.047)	(0.047)	(0.059)	(0.070)	(0.009)
	PL-St	0.219	0.177	0.097	0.116	0.123	0.124	0.094	0.051
		(0.034)	(0.045)	(0.035)	(0.038)	(0.062)	(0.060)	(0.052)	(0.034)
	NL-Tw	0.153	0.149	0.113	0.116	0.093	0.111	0.103	0.001
		(0.031)	(0.049)	(0.048)	(0.039)	(0.037)	(0.050)	(0.062)	(0.006)
	PL-mb	0.258	0.223	0.129	0.142	0.136	0.111	0.093	0.047
		(0.043)	(0.050)	(0.062)	(0.030)	(0.061)	(0.065)	(0.064)	(0.034)
	mcmc	0.140	0.204	0.168	0.153	0.070	0.057	0.062	0.014
		(0.066)	(0.106)	(0.103)	(0.053)	(0.042)	(0.050)	(0.038)	(0.037)

Figure 2.1: Boxplots of $\sqrt{\widehat{ISE}_{nonlinear}}$ as well as $\sqrt{\widehat{ISE}_{linear}}$ over 100 replicates by different approaches, for $N = 300$, $p = 15$, and Gaussian errors.

Figure 2.2: Boxplots of $\sqrt{\widehat{ISE}_{nonlinear}}$ as well as $\sqrt{\widehat{ISE}_{linear}}$ over 100 replicates by different approaches, for $N = 300$, $p = 15$, and Gaussian errors. (Continued)

normally distributed noises by different approaches are displayed in Tables 2.3 and 2.4. Results for the scenarios with Student's t distributed noises can be found in the appendix. The average values over 100 replicates are provided together with the standard deviations in the brackets. Figures 2.3 and 2.4 present more visible comparisons among different approaches using boxplots based on the same numerical results. To reduce the number of figures, only the case where $N = 300$, $p = 15$, and the noise terms follow a Gaussian distribution is plotted. We have made the following observations and comments on the numerical results:

1. In mean regression, our proposed twin boosting approach with both nonlinear and linear base learners ("PL-Tw") generally outperforms all the other approaches in terms of both $RMSE$ and ADE . Meanwhile among the five methods for comparison, twin boosting with only nonlinear base learners ("NL-Tw") and the Bayesian approach ("mcmc") perform better than the other three approaches.
2. In quantile regression, among all the six approaches, there are three, including our proposed approach, being competitive: "PL-Tw", "NL-Tw" and "mcmc". However, "mcmc" is slightly less stable than the other two with larger standard deviations. $ACLE$ results by the other three approaches are less satisfying.

2.2.3 Variable Selection

Besides estimation and prediction, model identification is an important goal of our proposed twin boosting approach. To evaluate, we calculate the number of informative covariates selected (NV), the number of informative

Table 2.3: *Average (standard deviation) of testing errors in mean regression over 100 replicates by different approaches with Gaussian errors. “RMSE” and “ADE” indicate the root mean squared errors and the absolute deviation errors, respectively.*

		$N = 300$	$N = 300$	$N = 500$	$N = 500$
		$p = 15$	$p = 30$	$p = 15$	$p = 30$
RMSE	PL-Tw	0.550(0.015)	0.551(0.016)	0.538(0.012)	0.539(0.011)
	PL-St	0.579(0.016)	0.604(0.020)	0.558(0.014)	0.574(0.012)
	NI-Tw	0.557(0.015)	0.558(0.017)	0.543(0.012)	0.544(0.012)
	PL-mb	0.585(0.017)	0.607(0.018)	0.565(0.013)	0.586(0.014)
	mgcv	0.569(0.017)	0.604(0.032)	0.543(0.015)	0.560(0.017)
	mcmc	0.560(0.017)	0.561(0.017)	0.545(0.011)	0.549(0.014)
ADE	PL-Tw	0.425(0.012)	0.426(0.012)	0.414(0.009)	0.415(0.009)
	PL-St	0.451(0.014)	0.473(0.017)	0.433(0.011)	0.446(0.010)
	NI-Tw	0.431(0.012)	0.432(0.013)	0.419(0.009)	0.419(0.009)
	PL-mb	0.458(0.014)	0.476(0.015)	0.439(0.010)	0.458(0.011)
	mgcv	0.442(0.014)	0.471(0.026)	0.421(0.012)	0.434(0.013)
	mcmc	0.435(0.014)	0.436(0.013)	0.423(0.009)	0.424(0.011)

Table 2.4: *Average (standard deviation) of check loss errors in quantile regression over 100 replicates by different approaches with Gaussian errors.*

		$N = 300$	$N = 300$	$N = 500$	$N = 500$
		$p = 15$	$p = 30$	$p = 15$	$p = 30$
$\tau = 0.1$	PL-Tw	0.107(0.007)	0.111(0.009)	0.099(0.004)	0.100(0.004)
	PL-St	0.116(0.006)	0.133(0.011)	0.105(0.004)	0.113(0.005)
	Nl-Tw	0.106(0.006)	0.107(0.006)	0.099(0.004)	0.100(0.004)
	PL-mb	0.122(0.008)	0.139(0.009)	0.108(0.004)	0.118(0.006)
	mgcv	0.148(0.012)	0.148(0.011)	0.150(0.007)	0.147(0.010)
	mcmc	0.111(0.013)	0.114(0.012)	0.095(0.003)	0.099(0.005)
$\tau = 0.3$	PL-Tw	0.193(0.006)	0.193(0.007)	0.186(0.005)	0.187(0.005)
	PL-St	0.204(0.007)	0.217(0.009)	0.195(0.005)	0.202(0.005)
	Nl-Tw	0.193(0.006)	0.192(0.006)	0.186(0.005)	0.186(0.004)
	PL-mb	0.210(0.008)	0.224(0.010)	0.199(0.005)	0.208(0.006)
	mgcv	0.206(0.008)	0.217(0.013)	0.198(0.006)	0.202(0.007)
	mcmc	0.195(0.019)	0.199(0.017)	0.184(0.005)	0.190(0.009)
$\tau = 0.5$	PL-Tw	0.219(0.007)	0.220(0.007)	0.211(0.005)	0.212(0.005)
	PL-St	0.232(0.007)	0.243(0.010)	0.222(0.006)	0.229(0.005)
	Nl-Tw	0.219(0.007)	0.219(0.007)	0.211(0.005)	0.212(0.004)
	PL-mb	0.237(0.008)	0.252(0.009)	0.227(0.005)	0.236(0.007)
	mgcv	0.221(0.007)	0.236(0.013)	0.210(0.006)	0.217(0.007)
	mcmc	0.218(0.007)	0.225(0.013)	0.213(0.006)	0.217(0.015)
$\tau = 0.7$	PL-Tw	0.194(0.007)	0.193(0.007)	0.185(0.005)	0.185(0.004)
	PL-St	0.205(0.006)	0.216(0.010)	0.195(0.005)	0.201(0.005)
	Nl-Tw	0.193(0.006)	0.193(0.007)	0.186(0.005)	0.185(0.004)
	PL-mb	0.210(0.007)	0.224(0.009)	0.199(0.005)	0.208(0.006)
	mgcv	0.204(0.009)	0.215(0.011)	0.197(0.007)	0.202(0.009)
	mcmc	0.194(0.009)	0.195(0.009)	0.186(0.004)	0.187(0.006)
$\tau = 0.9$	PL-Tw	0.107(0.008)	0.110(0.009)	0.098(0.004)	0.098(0.003)
	PL-St	0.114(0.006)	0.129(0.009)	0.104(0.005)	0.110(0.005)
	Nl-Tw	0.106(0.006)	0.108(0.007)	0.098(0.004)	0.098(0.004)
	PL-mb	0.119(0.007)	0.135(0.009)	0.107(0.004)	0.116(0.005)
	mgcv	0.144(0.012)	0.144(0.009)	0.148(0.010)	0.148(0.011)
	mcmc	0.109(0.012)	0.111(0.016)	0.098(0.003)	0.100(0.006)

Figure 2.3: Boxplots of testing errors in mean regression over 100 replicates by different approaches, for $N = 300$, $p = 15$, and Gaussian errors. “RMSE” and “ADE” indicate the root mean squared errors and the absolute deviation errors, respectively.

Figure 2.4: Boxplots of check loss errors in quantile regression over 100 replicates by different approaches, for $N = 300$, $p = 15$, and Gaussian errors.

covariates selected which are truly informative (NVT), the number of linear covariates selected (NL), and the number of linear covariates selected which are truly linear (NLT). Results are presented for the case where $N = 300$, $p = 15$, and errors follow the Gaussian distribution in Table 2.5. Results for the other scenarios can be found in the appendix. The average values over 100 replicates are provided together with the standard deviations in the brackets. Other quantities of interest can be extracted from these metrics. For example, the number of zero component functions estimated as nonzero (false positives) is $NV - NVT$, and the number of nonzero component functions estimated as zero (false negative) is $7 - NVT$. We have made the following observations and comments on the numerical results:

1. Overall, our proposed twin boosting approach with both nonlinear and linear base learners (“PL-Tw”) outperforms all the other approaches in terms of informative variable selection and linear covariate detection.
2. Standard boosting approach (“PL-St”) and algorithm in package *mboost* barely achieve the task of variable selection by always involving too many nonlinear covariates, which illustrates the power of twin boosting.
3. Twin boosting with only nonlinear base learners (“NL-Tw”) performs well in informative variable selection. However it is of course unable to detect linear covariates, which illustrates the usefulness of PLAM.
4. Again because “mgcv” implements quantile regression by associating mean regression with conditional variance estimation, it is enough to present variable selection results by “mgcv” for mean regression

only. To our knowledge, package *mgcv* does not include indicators of variable selection results in its numerical outputs. We have to manually examine the estimated component functions by “mgcv” to obtain model identification results of it. It turns out that “mgcv” selects all covariates as nonlinear in all the replicates of all scenarios. Fairly speaking, it is observed that estimated component function by “mgcv” of a truly linear covariate can sometimes be visibly closed enough to a purely linear function and thus suggests selecting the corresponding covariate as linear. The similar is observed for estimated component functions of truly non-informative covariates by “mgcv” as well. However, this requires a threshold to be set and so far it is not clear to us as for how to do this in a principled way. Meanwhile similar argument can be made for boosting approaches as well. In this thesis, a base learner is regarded as included in the estimated model as long as it is selected in at least one iteration. However if a base learner is selected only in a few iterations, corresponding estimated component function could be quite closed to zero and the covariate effect could be regarded as effectively excluded in practice. This is in fact part of motivation of the technique named stability selection, there are more discussions on this issue in Chapter 4.

5. The Bayesian approach (“mcmc”) is designed especially for PLAM and shows its strength in variable selection to some extent. However, “PL-Tw” outperforms “mcmc” in most of the cases with respect to model identification.
6. “PL-Tw”, “NI-Tw” and “mcmc” implement variable selection at the cost of occasionally missing some truly informative covariates.

Table 2.5: *Average (standard deviation) of variable selection results over 100 replicates by different approaches, for $N = 300$, $p = 15$, and Gaussian errors. “NV”, “NVT”, “NL”, and “NLT” indicate the number of nonzero covariates selected, the number of nonzero covariates selected which are truly informative, the number of linear covariates selected, and the number of linear covariates selected which are truly linear, respectively.*

		NV	NVT	NL	NLT
Mean	PL-Tw	8.54(1.28)	7.00(0.00)	2.91(0.71)	2.58(0.62)
	PL-St	14.7(0.77)	7.00(0.00)	0.74(1.00)	0.26(0.50)
	Nl-Tw	8.22(1.26)	7.00(0.00)	-	-
	PL-mb	14.6(0.82)	7.00(0.00)	0.86(0.95)	0.30(0.50)
	mgcv	15.0(0.00)	7.00(0.00)	0.00(0.00)	0.00(0.00)
	mcmc	7.17(0.51)	6.92(0.18)	1.40(0.93)	1.30(0.95)
$\tau = 0.1$	PL-Tw	7.35(0.64)	6.97(0.17)	3.14(0.51)	2.92(0.27)
	PL-St	15.0(0.17)	7.00(0.00)	0.18(0.61)	0.05(0.26)
	Nl-Tw	7.22(0.52)	6.96(0.20)	-	-
	PL-mb	15.0(0.41)	7.00(0.00)	1.01(1.11)	0.23(0.47)
	mcmc	8.20(1.03)	6.93(0.25)	4.00(1.49)	2.66(0.55)
$\tau = 0.3$	PL-Tw	7.27(0.49)	7.00(0.00)	3.09(0.35)	2.97(0.17)
	PL-St	14.8(0.56)	7.00(0.00)	0.35(0.80)	0.12(0.38)
	Nl-Tw	7.29(0.56)	7.00(0.00)	-	-
	PL-mb	14.9(0.39)	7.00(0.00)	0.89(1.10)	0.28(0.51)
	mcmc	7.61(1.32)	7.00(0.00)	2.52(1.07)	2.30(0.85)
$\tau = 0.5$	PL-Tw	7.17(0.43)	7.00(0.00)	3.04(0.32)	2.97(0.17)
	PL-St	14.8(0.48)	7.00(0.00)	0.34(0.70)	0.17(0.40)
	Nl-Tw	7.33(0.60)	7.00(0.00)	-	-
	PL-mb	14.9(0.39)	7.00(0.00)	1.05(1.05)	0.33(0.57)
	mcmc	7.68(1.10)	7.00(0.00)	2.50(0.94)	2.40(0.86)
$\tau = 0.7$	PL-Tw	7.25(0.48)	7.00(0.00)	3.01(0.41)	2.91(0.29)
	PL-St	14.9(0.51)	7.00(0.00)	0.31(0.61)	0.13(0.39)
	Nl-Tw	7.35(0.52)	7.00(0.00)	-	-
	PL-mb	14.9(0.33)	7.00(0.00)	0.99(1.07)	0.30(0.56)
	mcmc	8.09(1.45)	7.00(0.00)	2.70(1.15)	2.48(0.86)
$\tau = 0.9$	PL-Tw	7.43(0.66)	6.98(0.14)	3.00(0.53)	2.84(0.37)
	PL-St	15.0(0.22)	7.00(0.00)	0.12(0.41)	0.04(0.20)
	Nl-Tw	7.34(0.57)	6.99(0.10)	-	-
	PL-mb	14.9(0.43)	7.00(0.00)	1.09(1.32)	0.28(0.53)
	mcmc	7.50(1.25)	6.80(0.48)	3.47(1.41)	2.64(0.61)

2.3 Real Data Applications

2.3.1 House Price Data

The house price example concerns the relationship between house price and several features, including physical attributes of the residential houses, characteristics of the surrounding rural district, and environmental nuisances. The responses involved are the transaction prices for the residential houses sold during 1996 and 1997 in Brittany. Meanwhile there are $p = 10$ predictors: AGE (age of the house), REPAIR (state of repair), ROOMS (number of rooms), LOT (lot size), COUNTRY (whether the surrounding district is located in the department of Ille et Vilaine), VACANT (proportion of vacant houses), POP (population of the district), AVINC (average taxable family income), TMEAD (proportion of permanent grassland converted into cultivated grassland), and NITRO (amount of nitrogen emissions from livestock farming per hectare of arable land in the rural district where the residential house is located). The dataset with total size 2053 is used in Bontemps et al. (2008) to investigate how house prices are impacted by pollutions resulting from intensive livestock farming.

In our numerical study, the dataset is divided into a training sample with size $N = 1000$ and a validation sample with size $N_v = 1053$. 30 random partitions are generated to investigate the stability of variable selection. For the two binary predictors, REPAIR and COUNTRY, the nonlinear parts of these two component functions are forced to be 0. Our proposed twin boosting approach, as well as the five approaches for comparison mentioned in the previous simulation section, is applied to this dataset for mean regression as well as quantile regressions at five levels 0.1, 0.3, 0.5, 0.7, and 0.9. Variable selection results by the proposed approach (“PL-Tw”) are

summarized in Table 2.6, where the first number in each bracket is the number of replicates when corresponding covariate is identified as nonlinear and the second number is the number of times when the covariate is selected as (nonzero) linear. Table 2.7 summarizes variable selection results by different approaches. In each bracket, the two numbers are the average numbers of nonlinear and linear covariates, over 30 replicates, selected by the corresponding approach. Meanwhile for “Nl-Tw”, which is unable to detect linear covariates, the corresponding number indicates the average number of informative covariates selected. Algorithm in the package *mgcv* selects REPAIR and COUNTRY as linear (which is manually forced to) and all the other covariates as nonlinear in every single replicate, thus results by “mgcv” are not displayed. One of the 30 replicates is randomly picked and the regression functions are plotted in Figures 2.5 ~ 2.10, for different approaches respectively. In each panel, the regression function $\hat{f}(x)$ is plotted with the corresponding predictor varying and the others fixed at 0.5. Five black dashed curves correspond to the results of quantile regressions at five mentioned levels, while the blue solid curve indicates the results of mean regression. We have made the following observations and comments on the numerical results:

1. Overall, our proposed approach (“PL-Tw”) provides more parsimonious estimated models by detecting more linear and non-informative covariates, compared to all the other approaches. In this aspect, the Bayesian approach (“mcmc”) performs the second best in this example.
2. Generally, different approaches provide visually similar estimated functions, especially the four boosting approaches “PL-Tw”, “PL-St”,

“NI-Tw”, and “PL-mb”. Estimated functions by “mgcv” in quantile regressions have important differences from those by other approaches, maybe due to the different scheme of “mgcv” in quantile regression (associating mean regression with conditional variance estimation).

3. Estimated component functions for AGE, ROOMS and AVINC are roughly monotone but with empirically non-constant slopes. Therefore these covariates are selected as nonlinear with high frequencies. More specifically, the age of the house is found to have a negative impact on the price, while the number of rooms and the average taxable family income in the neighbourhood have positive ones, which all match intuitive expectations.
4. Other than the two binary predictors, VACANT and TMEAD are also detected as linear with high frequencies. The proportion of vacant houses in the neighbourhood and the proportion of permanent grassland converted into cultivated grassland both have negative effects on the house price, which are intuitive as well.
5. Factors like the population of the district and the amount of nitrogen emissions from livestock farming are not observed to have important influences on the house price.
6. Overall, physical attributes of the houses tend to affect the house price more than surrounding and environmental factors.
7. Heterogeneity is observed to some extent among mean regression and quantile regressions at different levels. For example, effect of ROOMS can be linear at quantile level 0.1 while nonlinear at higher quantile levels and in mean regression. Impact of VACANT is empirically

Table 2.6: *House price data. Variable selection results by PL-Tw.*

	Mean	$\tau = 0.1$	$\tau = 0.3$	$\tau = 0.5$	$\tau = 0.7$	$\tau = 0.9$
AGE	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(29,1)
REPAIR	(0,30)	(0,30)	(0,30)	(0,30)	(0,30)	(0,30)
ROOMS	(30,0)	(16,14)	(30,0)	(30,0)	(30,0)	(30,0)
LOT	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)
COUNTRY	(0,30)	(0,30)	(0,30)	(0,30)	(0,30)	(0,24)
VACANT	(28,2)	(5,15)	(12,16)	(7,21)	(4,25)	(6,11)
POP	(14,11)	(7,9)	(2,24)	(2,10)	(8,4)	(0,4)
AVINC	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(29,1)
TMEAD	(18,12)	(4,18)	(3,25)	(6,23)	(8,18)	(15,11)
NITRO	(22,8)	(4,0)	(22,2)	(17,6)	(8,14)	(12,2)

a little bit stronger in mean regression and quantile regressions at moderate quantile levels (0.3, 0.5, and 0.7) than at extreme quantile levels (0.1 and 0.9).

8. Bontemps et al. (2008) studies this dataset from the viewpoint of econometrics, with special attention to how the environmental nuisances impact consumers' willingness to pay. It is revealed that AGE, VACANT, TMEAD and NITRO have negative influences on the house price while the other six attributes have positive ones, which are confirmed again by our numerical results.

Figure 2.5: *House price data. Estimated component functions by PL-Tw.*

Figure 2.6: *House price data. Estimated component functions by PL-St.*

Figure 2.7: House price data. Estimated component functions by NL-Tw.

Figure 2.8: House price data. Estimated component functions by PL-mb.

Figure 2.9: *House price data. Estimated component functions by mgcv.*

Figure 2.10: *House price data. Estimated component functions by mcmc.*

Table 2.7: *House price data. Variable selection results by different approaches.*

	PL-Tw	PL-St	Nl-Tw	PL-mb	mcmc
mean	(6.73,3.10)	(8.00,2.00)	9.80	(7.93,2.07)	(5.13,3.67)
$\tau = 0.1$	(4.20,3.87)	(7.73,2.13)	8.20	(7.90,2.03)	(6.73,2.03)
$\tau = 0.3$	(5.30,4.23)	(7.87,2.13)	9.63	(7.80,2.20)	(6.97,2.43)
$\tau = 0.5$	(5.07,4.00)	(8.00,2.00)	9.40	(7.77,2.23)	(6.27,2.93)
$\tau = 0.7$	(4.93,4.03)	(8.00,2.00)	9.20	(7.77,2.20)	(6.40,2.03)
$\tau = 0.9$	(5.03,2.80)	(8.00,2.00)	8.87	(7.60,2.30)	(5.97,2.03)

2.3.2 Concrete Compressive Strength Data

The second example concerns the strong connection between concrete compressive strength and its age and ingredients as concrete is believed to be one of the most important materials in civil engineering nowadays. The concrete compressive strength is treated as the response while there are $p = 8$ covariates included: Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate, and Age. The dataset with total size 1030 is used in Yeh (1998) to show that the water-to-cement ratio is not the only factor that influences the concrete compressive strength.

In our numerical study, the dataset is divided into a training sample with size $N = 500$ and a validation sample with size $N_v = 530$, and again, 30 random partitions are generated. Our proposed twin boosting approach, as well as the five approaches for comparison, is applied to this dataset for mean regression as well as quantile regressions at five levels 0.1, 0.3, 0.5, 0.7, and 0.9. The numbers of replicates when each of the covariates is identified as nonlinear or linear respectively by our proposed approach are

summarized in Table 2.8. Table 2.9 summarizes variable selection results by different approaches while those by “mgcv” are omitted again for similar reason mentioned in the house price example. In this example the standard gradient boosting approach (“PL-St”) also selects all the covariates as non-linear in every single replicate, thus its numerical results are omitted as well. Like the previous example, one replicate is randomly picked and the regression functions are shown in Figures 2.11 ~ 2.16, for different approaches respectively. We have made the following observations and comments on the numerical results:

1. Overall, our proposed approach (“PL-Tw”) provides sparser estimated models by detecting more non-informative covariates. In this aspect, “NL-Tw” performs similarly, which however is unable to detect linear covariates. In this example, “PL-mb” detects more linear predictors while seldom leaves any covariate out of its estimated models.
2. Generally, different approaches provide visually similar estimated functions, especially the four boosting approaches “PL-Tw”, “PL-St”, “NL-Tw”, and “PL-mb”. Estimated functions by “mgcv” in quantile regressions have important differences from those by other approaches, maybe due to the different scheme of “mgcv” in quantile regression.
3. Among the seven ingredients, cement and water are observed to have the most important effects on the concrete compressive strength. The content of cement has a positive impact on the strength while the content of water has a negative one, which matches intuitive expectations and might be the reason why people have long treated water-to-cement ratio as the most important factor influencing the concrete

compressive strength.

4. The age of concrete is also observed to have a strong effect. The compressive strength increases sharply in the first two months and after that the impact of age does not change much over time. This point greatly coincides with the results in Yeh (1998), which studies this dataset from a viewpoint of civil engineering.
5. To our limited knowledge in civil engineering, blast furnace slag is one of the cementitious materials which should play a similar role as the cement. From the numerical results, it does have a roughly positive impact on the concrete compressive strength.
6. Some ingredients like fly ash, coarse aggregate, and fine aggregate are found to be without strong influence on the concrete compressive strength. However impacts of fly ash and coarse aggregate are a little bit stronger at lower quantile levels (0.1 and 0.3), while effect of fine aggregate is relatively greater at higher quantile levels (0.7 and 0.9).

Table 2.8: *Concrete compressive strength data. Variable selection results by PL-Tw.*

	Mean	$\tau = 0.1$	$\tau = 0.3$	$\tau = 0.5$	$\tau = 0.7$	$\tau = 0.9$
Cement	(30,0)	(17,13)	(23,7)	(22,8)	(24,6)	(8,22)
Blast Furnace Slag	(30,0)	(30,0)	(29,1)	(30,0)	(30,0)	(30,0)
Fly Ash	(30,0)	(19,0)	(14,0)	(4,1)	(1,0)	(3,0)
Water	(30,0)	(20,10)	(29,1)	(30,0)	(30,0)	(30,0)
Superplasticizer	(30,0)	(23,0)	(29,0)	(30,0)	(30,0)	(26,0)
Coarse Aggregate	(25,0)	(2,11)	(3,2)	(3,0)	(5,1)	(1,1)
Fine Aggregate	(30,0)	(4,0)	(3,2)	(11,2)	(17,5)	(0,27)
Age	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)	(30,0)

Table 2.9: *Concrete compressive strength data. Variable selection results by different approaches.*

	PL-Tw	Nl-Tw	PL-mb	mcmc
mean	(7.83,0.00)	7.87	(8.00,0.00)	(8.00,0.00)
$\tau = 0.1$	(4.83,1.13)	6.50	(6.77,1.10)	(8.00,0.00)
$\tau = 0.3$	(5.33,0.43)	5.93	(7.07,0.90)	(7.97,0.03)
$\tau = 0.5$	(5.33,0.37)	5.77	(6.83,1.17)	(8.00,0.00)
$\tau = 0.7$	(5.57,0.40)	5.70	(6.93,1.07)	(8.00,0.00)
$\tau = 0.9$	(4.27,1.67)	5.97	(6.13,1.83)	(7.97,0.03)

Figure 2.11: *Concrete compressive strength data. Estimated component functions by PL-Tw.*

Figure 2.12: *Concrete compressive strength data. Estimated component functions by PL-St.*

Figure 2.13: *Concrete compressive strength data. Estimated component functions by NI-Tw.*

Figure 2.14: *Concrete compressive strength data. Estimated component functions by PL-mb.*

Figure 2.15: *Concrete compressive strength data. Estimated component functions by mgcv.*

Figure 2.16: *Concrete compressive strength data. Estimated component functions by mcmc.*

Chapter 3

Survival Analysis with Boosting

3.1 Methodology

3.1.1 Survival Time

Survival analysis studies the distribution of the survival time T_s , which is the length of the time period before a certain event (e.g. death, breakdown, etc.) happens. The event is also popularly called the “failure” in literatures. In most of the cases, T_s is assumed to follow a continuous distribution, with the cumulative distribution function $F_{T_s}()$ and the probability density function $f_{T_s}()$. The survival function $S_{T_s}()$, also known as the survivor function or reliability function, of T_s is the probability that the subject survives any specified time t :

$$S_{T_s}(t) = \Pr[T_s \geq t] = 1 - F_{T_s}(t).$$

Another function which can represent the distribution of T_s is the hazard function, which describes the rate of the failure at any specified time t conditional on the survival until time t or later, and is defined as

$$\lambda_{T_s}(t) = \lim_{dt \rightarrow 0+} \frac{\Pr[t \leq T_s < t + dt | T_s \geq t]}{dt}.$$

By the definition of conditional probability, it can be easily derived that

$$\begin{aligned}
\lambda_{T_s}(t) &= \lim_{dt \rightarrow 0+} \frac{\Pr[t \leq T_s < t + dt | T_s \geq t]}{dt} \\
&= \lim_{dt \rightarrow 0+} \frac{\Pr[t \leq T_s < t + dt]}{\Pr[T_s \geq t]} \cdot \frac{1}{dt} \\
&= \lim_{dt \rightarrow 0+} \frac{F_{T_s}(t + dt) - F_{T_s}(t)}{dt} \cdot \frac{1}{S_{T_s}(t)} \\
&= \frac{f_{T_s}(t)}{S_{T_s}(t)}.
\end{aligned}$$

3.1.2 Proportional Hazards Model

Nowadays in survival analysis, literatures are always interested in the effects of covariates on the distribution of survival time T_s . Different models (Lin and Ying (1995), Murphy (1995), Sastry (1997)) have been proposed while the Cox's proportional hazards model might have become the most popular one in use. The model is firstly proposed in Cox (1972) and has been studied in numerous works including Therneau and Grambsch (2000), Goeman (2010), and Lian et al. (2013). Cox's model assumes that the predictors $\mathbf{X}$ influence the hazard function $\lambda_{T_s}(\cdot)$ of T_s in a multiplicative way:

$$\lambda_{T_s|\mathbf{X}}(t) = \lambda_0(t)\psi(\mathbf{X}),$$

where $\lambda_0(\cdot)$, free of $\mathbf{X}$, is called the baseline hazard function and needs not to be known in the model. It is called the proportional hazards model because the ratio of hazard functions for two subjects with fixed covariate vectors $\mathbf{X}_i$ and $\mathbf{X}_j$ is constant over time:

$$\frac{\lambda_{T_s|\mathbf{X}_i}(t)}{\lambda_{T_s|\mathbf{X}_j}(t)} = \frac{\lambda_0(t)\psi(\mathbf{X}_i)}{\lambda_0(t)\psi(\mathbf{X}_j)} = \frac{\psi(\mathbf{X}_i)}{\psi(\mathbf{X}_j)}.$$

To formularize $\psi(\mathbf{X})$, as it has to be non-negative overall, it is natural and conventional to consider $f(\mathbf{X}) = \ln \psi(\mathbf{X})$, where $f(\cdot)$ is sometimes

called the risk score function. For example in the most classical and basic linear additive proportional hazards model, it is assumed that $f(\mathbf{X}) = \sum_{j=1}^p \beta_j X_j$ where β_j 's are the underlying coefficients to be estimated. Here in this thesis, the partially linear additive model (PLAM) is adapted to the proportional hazards model, and it is thus assumed that

$$\ln \psi(\mathbf{X}) = f(\mathbf{X}) = \sum_{j=1}^p f_j(X_j),$$

where each component function f_j can be nonlinear, linear, or zero in functional form. We aim to develop a boosting approach to estimate the underlying functions f_j 's and simultaneously identify which f_j 's are nonlinear, linear, or zero, respectively. Recalling the decomposition (2.1.1), again we formularize

$$f(\mathbf{X}) = \sum_{j=1}^p g_j(X_j) + \sum_{j=1}^p h_j(X_j),$$

with g_j 's being the linear components and h_j 's being the nonlinear components. In the proportional hazards model, an intercept is not necessary since it can be merged into the baseline hazard function:

$$\lambda_0(t) e^{[\beta_0 + \sum_{j=1}^p g_j(X_j) + \sum_{j=1}^p h_j(X_j)]} = [e^{\beta_0} \lambda_0(t)] e^{[\sum_{j=1}^p g_j(X_j) + \sum_{j=1}^p h_j(X_j)]}.$$

In our proposed approach, g_j 's and h_j 's are estimated. X_j is identified as a nonlinear predictor if and only if $\hat{h}_j \neq 0$; X_j is identified as a linear predictor if and only if $\hat{h}_j = 0$ but $\hat{g}_j \neq 0$; X_j is identified as a non-informative predictor if and only if $\hat{g}_j = 0$ and $\hat{h}_j = 0$.

3.1.3 Censoring and Partial Likelihood

As mentioned in Chapter 1, we consider right censoring of survival times. There are many causes of censoring in practice, for instance that a subject

may survive the study period, a patient may die from reasons other than the disease of interest, contact may be lost, or follow-up may be refused. In the model, censoring time T_c is usually assumed to be independently distributed with T_s , and only the minimum of T_s and T_c can be observed. We denote the observed time with T and the censoring indicator with $\delta = \mathbf{1}_{[T_s < T_c]}$. To solve the proportional hazards model with partially censored observations $\{(\mathbf{X}_i, T_i, \delta_i)\}_{i=1}^N$, a partial likelihood function of the hazard function λ given the observed data is introduced in Cox (1972):

$$Lik(\lambda) = \prod_{\substack{1 \leq i \leq N \\ \delta_i = 1}} \frac{\lambda(T_i; \mathbf{X}_i)}{\sum_{j=1}^N \mathbf{1}_{[T_j \geq T_i]} \lambda(T_i; \mathbf{X}_j)}, \quad (3.1.1)$$

where $\lambda(t; \mathbf{X}_i) = \lambda_{T_s|\mathbf{X}=\mathbf{X}_i}(t)$. The rationale of this likelihood function can be interpreted as follows. Given $\forall i^* \in \{i : \delta_i = 1\}$, recalling that the hazard function measures the rate of the failure at any specified time t conditional on the survival until time t or later, $\lambda(T_{i^*}; \mathbf{X}_{i^*})$ is the rate of the i^* -th observation to fail at time T_{i^*} . Meanwhile $\sum_{j=1}^N \mathbf{1}_{[T_j \geq T_{i^*}]} \lambda(T_{i^*}; \mathbf{X}_j)$ is the rate of that there is one failure at time T_{i^*} among all the observations having not failed until time T_{i^*} , assuming no tied times are allowed. Therefore, $\frac{\lambda(T_{i^*}; \mathbf{X}_{i^*})}{\sum_{j=1}^N \mathbf{1}_{[T_j \geq T_{i^*}]} \lambda(T_{i^*}; \mathbf{X}_j)}$ is the conditional probability of that the i^* -th observation fails at time T_{i^*} given that there is one failure at time T_{i^*} among all the observations having not failed until time T_{i^*} , and the partial likelihood is essentially the product of a sequence of conditional probabilities. More details can be found in Cox (1975), Efron (1977) and Cox and Oakes (1984) to support the using of the partial likelihood. Recalling that $\lambda_{T_s|\mathbf{X}}(t) = \lambda_0(t)e^{f(\mathbf{X})}$, the baseline hazard function λ_0 is cancelled out in the fractions and the log partial likelihood function of the risk score function f

given the observed data is therefore

$$lik(f) = \sum_{i=1}^N \delta_i [f(\mathbf{X}_i) - \log(\sum_{j=1}^N 1_{[T_j \geq T_i]} e^{f(\mathbf{X}_j)})].$$

3.1.4 Boosting for Survival Analysis

The gradient boosting approach proposed in Chapter 2 is adapted to survival analysis for proportional hazards model with partially censored data. As for the choice of base learners, univariate linear regressions and univariate penalized splines are again used to recover PLAM structures. Furthermore, like in Li and Luan (2005), the negative log partial likelihood is used as the loss function in boosting approach, therefore the corresponding negative gradient at any current estimate $\hat{f}$ is

$$U_i = \delta_i - \sum_{j=1}^N \delta_j 1_{[T_j \geq T_i]} \frac{e^{\hat{f}(\mathbf{X}_i)}}{\sum_{k=1}^N 1_{[T_k \geq T_j]} e^{\hat{f}(\mathbf{X}_k)}},$$

with $i = 1, 2, \dots, N$ and $\mathbf{U} = (U_1, \dots, U_N)^T$. The twin boosting approach is as well incorporated to achieve better variable selection and model identification.

3.2 Simulation Illustrations

We firstly evaluate the proposed method by simulation studies. The used model consists of p predictors among which 2 are nonlinear, 3 are linear, and the rest are all non-informative. More specifically, $f_1(x) = 2 \sin(2\pi x)$, $f_2(x) = 9x^2 - 6x$, $f_3(x) = 3x$, $f_4(x) = -3x$, $f_5(x) = 2x$, and

$$f(\mathbf{X}) = \sum_{j=1}^5 f_j(X_j).$$

The predictors are generated from the multivariate Gaussian distribution with mean $\mathbf{0}$ and covariance matrix Σ , where $\Sigma_{j_1, j_2} = 0.1^{|j_1 - j_2|}$, and then transformed to be marginally uniform on $[0, 1]$ by applying the cumulative distribution function of the univariate standard normal distribution. The survival time T_s , given predictor $\mathbf{X}$, is generated from the exponential distribution with mean $1/\exp[f(\mathbf{X})]$. In this exponential setting, it can be proved that the risk score function is just $f(\mathbf{X})$. Meanwhile the censoring time T_c is independently generated from the exponential distribution with mean $1/C$, where C is sometimes called the censoring variable. The observed time T equals the minimum of T_s and T_c . And the censoring indicator δ equals $\mathbf{1}_{[T_s < T_c]}$.

We set the sample size as $N = 200$ or 400 . Total number of covariates is set as $p = 20$ or 100 . Choices of censoring variable C are 0.2 and 0.5 . Corresponding uncensored data rates are empirically approximately normally distributed with mean ≈ 0.7 and 0.6 respectively, and standard deviation ≈ 0.03 . Like the previous chapter, the degree of truncated power splines is fixed to be $q = 3$ and the internal knots used in TPS are set to be equally spaced $\{0.1, 0.2, \dots, 0.9\}$. Again we use the commonly suggested step size $\nu = 0.1$. The proposed twin boosting algorithm with linear and nonlinear base learners is implemented with the stopping time decided by 5-fold cross validation. 100 independent replicates are used in each of the settings. Again for comparison, four other approaches are applied (to the same datasets):

1. “PL-St”. Standard gradient boosting approach (equal weight 1 for all base learners) with linear and nonlinear base learners. This is to illustrate the power of twin boosting.

2. “NL-Tw”. Twin boosting approach with only penalized spline base learners. This is to illustrate the usefulness of PLAM.
3. “PL-mb”. The algorithm in existing R package *mboost*. This package implements model-based boosting approaches and is able to be applied to survival analysis with Cox’s model. It is also allowed, in this package, to supply separate base learners for the linear parts and non-linear parts of covariate influences. However *mboost* does not involve twin boosting.
4. “mgcv”. The algorithm in existing R package *mgcv*. This package is designed as a mixed generalized additive model computation vehicle mostly using penalization methods. It can be applied to survival analysis with Cox’s model.

In this chapter, the Bayesian approach introduced by Hu et al. (2015) is not included since it is so far not clear to us as for how to extend this approach to survival analysis. In the following subsections, the performances of different approaches are investigated in terms of estimation accuracy, prediction accuracy, and variable selection accuracy. Numerical results show that our proposed two-round twin boosting approach has a favorable performance among different methods, especially a highlighted ability of variable selection using twin boosting.

3.2.1 Estimation Accuracy

We firstly calculate the integrated squared error (*ISE*) to measure the estimation accuracy of different approaches. *ISE* is defined as

$$ISE_j = \int_0^1 (\hat{f}_j(x) - f_j(x))^2 dx,$$

and the integral is approximated on an equidistant sequence $\{t_1, \dots, t_{1000}\}$ inside the interval $[0, 1]$. Tables 3.1 and 3.2 present $\sqrt{\widehat{ISE}}$ of all the 5 nonzero component functions $(f_1, \dots, f_5)$, the first zero component function (f_6) , as well as the total, where it is defined that $\widehat{ISE}_{total} = \sum_{j=1}^p \widehat{ISE}_j$. Different combinations of N , p , and C are included. The average values over 100 replicates are provided together with the standard deviations in the brackets. Numerical results by our proposed twin boosting approach, as well as the mentioned four approaches for comparison, are displayed. It is worth mentioning that the package *mgcv* fails in all the scenarios where $p = 100$ due to too large number of covariates relative to the sample size. Therefore numerical results by “mgcv” are only available for cases with $p = 20$. Figures 3.1 ~ 3.4 present more visible comparisons among different approaches using boxplots based on the same numerical results. To reduce the number of figures, covariates are combined by the truly nonlinear, linear, and non-informative groups as $\widehat{ISE}_{nonlinear} = \sum_{j=1}^2 \widehat{ISE}_j$, $\widehat{ISE}_{linear} = \sum_{j=3}^5 \widehat{ISE}_j$, and $\widehat{ISE}_{non-informative} = \sum_{j=6}^p \widehat{ISE}_j$. To further visually illustrate the power of our proposed approach, estimated component functions are plotted in Figure 3.5 for the five truly informative covariates $X_1, \dots, X_5$ and the first non-informative covariate X_6 . In each panel, the light blue dot-dash curves correspond to estimates in 100 replicates, while the red solid curve indicates the underlying true function. For simplicity, only the case where $N = 200$, $p = 20$, and $C = 0.2$ is presented as an example. We have made the following observations and comments on the numerical results:

1. As for estimation of nonlinear component functions, the proposed approach “PL-Tw”, twin boosting with only nonlinear base learners “NI-Tw”, and the package *mgcv* are competitive. The other two approaches perform less satisfyingly.

2. In estimation of linear component functions, “PL-Tw” and “mgcv” generally outperform the other three approaches.
3. Two twin boosting approaches “PL-Tw” and “NL-Tw” provide smaller estimates of non-informative component functions than the other three approaches.
4. Overall, our proposed approach “PL-Tw” performs the best in terms of the smallest total *ISE*’s.
5. According to Figure 3.5, component functions can be recovered quite well by the proposed approach.

3.2.2 Prediction Accuracy

Bøvelstad et al. (2007) explicitly suggests that the partial likelihood can be used as a model evaluation criterion for Cox’s model. In this section, we calculate, for different approaches, the negative log partial likelihood (as a measurement of error)

$$-\frac{1}{N_t} \sum_{i=1}^{N_t} \delta_i (\hat{y}_i - \log(\sum_{j=1}^{N_t} 1_{[t_j \geq t_i]} e^{\hat{y}_i})),$$

given an independently generated testing sample with size $N_t = 1500$, where T_i ’s and δ_i ’s are the observed times and censoring indicators of the testing sample, and $\hat{y}_i$ ’s are the predicted values of the risk score function. Results for different scenarios are displayed in Table 3.3. The average values over 100 replicates are provided together with the standard deviations in the brackets. Again numerical results by “mgcv” are only available for cases with $p = 20$ for the mentioned reason. Figures 3.6 and 3.7 present more visible comparisons among different approaches using boxplots based on

Table 3.1: Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates by different approaches for $C = 0.2$.

		f_1	f_2	f_3	f_4	f_5	f_6	total
$N = 200$ $p = 20$	PL-Tw	0.209	0.166	0.105	0.108	0.118	0.002	0.371
		(0.079)	(0.063)	(0.067)	(0.077)	(0.090)	(0.012)	(0.096)
	PL-St	0.350	0.251	0.195	0.214	0.184	0.040	0.624
		(0.099)	(0.076)	(0.090)	(0.084)	(0.088)	(0.039)	(0.104)
	NL-Tw	0.208	0.158	0.161	0.168	0.166	0.000	0.414
		(0.077)	(0.056)	(0.071)	(0.069)	(0.084)	(0.000)	(0.087)
	PL-mb	0.278	0.230	0.156	0.155	0.126	0.038	0.517
		(0.096)	(0.083)	(0.081)	(0.086)	(0.070)	(0.045)	(0.106)
	mgcv	0.215	0.171	0.103	0.105	0.086	0.035	0.414
		(0.073)	(0.070)	(0.065)	(0.078)	(0.067)	(0.062)	(0.112)
	PL-Tw	0.141	0.130	0.073	0.076	0.072	0.002	0.251
		(0.044)	(0.040)	(0.050)	(0.055)	(0.054)	(0.009)	(0.060)
$N = 400$ $p = 20$	PL-St	0.240	0.192	0.147	0.157	0.123	0.038	0.448
		(0.071)	(0.060)	(0.060)	(0.067)	(0.058)	(0.033)	(0.081)
	NL-Tw	0.135	0.126	0.110	0.114	0.106	0.000	0.280
		(0.043)	(0.039)	(0.042)	(0.050)	(0.047)	(0.001)	(0.053)
	PL-mb	0.197	0.167	0.120	0.130	0.095	0.034	0.382
		(0.063)	(0.048)	(0.054)	(0.062)	(0.050)	(0.037)	(0.070)
	mgcv	0.140	0.117	0.064	0.072	0.056	0.023	0.276
		(0.047)	(0.046)	(0.038)	(0.047)	(0.038)	(0.043)	(0.062)
$N = 200$ $p = 100$	PL-Tw	0.224	0.175	0.116	0.117	0.133	0.004	0.403
		(0.082)	(0.070)	(0.073)	(0.077)	(0.095)	(0.019)	(0.107)
	PL-St	0.455	0.333	0.259	0.282	0.239	0.019	0.808
		(0.109)	(0.088)	(0.094)	(0.095)	(0.096)	(0.035)	(0.120)
	NL-Tw	0.198	0.158	0.158	0.159	0.167	0.003	0.410
		(0.068)	(0.070)	(0.068)	(0.059)	(0.093)	(0.022)	(0.091)
	PL-mb	0.406	0.321	0.209	0.225	0.203	0.017	0.720
		(0.117)	(0.092)	(0.108)	(0.080)	(0.076)	(0.031)	(0.124)
$N = 400$ $p = 100$	PL-Tw	0.161	0.125	0.078	0.081	0.070	0.001	0.266
		(0.051)	(0.048)	(0.050)	(0.053)	(0.054)	(0.007)	(0.066)
	PL-St	0.337	0.244	0.198	0.207	0.153	0.016	0.592
		(0.079)	(0.060)	(0.063)	(0.065)	(0.065)	(0.025)	(0.083)
	NL-Tw	0.152	0.120	0.112	0.113	0.105	0.000	0.286
		(0.051)	(0.046)	(0.047)	(0.045)	(0.053)	(0.000)	(0.062)
	PL-mb	0.294	0.228	0.159	0.171	0.135	0.010	0.526
		(0.066)	(0.063)	(0.061)	(0.058)	(0.057)	(0.019)	(0.071)

Table 3.2: Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates by different approaches for $C = 0.5$.

		f_1	f_2	f_3	f_4	f_5	f_6	total
$N = 200$ $p = 20$	PL-Tw	0.234	0.181	0.121	0.120	0.138	0.004	0.415
		(0.095)	(0.064)	(0.076)	(0.083)	(0.106)	(0.019)	(0.117)
	PL-St	0.376	0.261	0.209	0.221	0.197	0.047	0.665
		(0.110)	(0.073)	(0.096)	(0.095)	(0.094)	(0.047)	(0.116)
	NL-Tw	0.228	0.169	0.182	0.183	0.191	0.000	0.459
		(0.087)	(0.057)	(0.077)	(0.081)	(0.095)	(0.002)	(0.102)
	PL-mb	0.303	0.241	0.171	0.168	0.138	0.037	0.554
		(0.104)	(0.090)	(0.086)	(0.097)	(0.071)	(0.046)	(0.117)
	mgcv	0.227	0.179	0.107	0.115	0.091	0.025	0.434
		(0.077)	(0.072)	(0.073)	(0.082)	(0.069)	(0.051)	(0.107)
$N = 400$ $p = 20$	PL-Tw	0.155	0.142	0.079	0.084	0.080	0.002	0.278
		(0.049)	(0.045)	(0.051)	(0.059)	(0.059)	(0.008)	(0.061)
	PL-St	0.265	0.204	0.160	0.165	0.129	0.036	0.483
		(0.075)	(0.060)	(0.063)	(0.075)	(0.067)	(0.036)	(0.082)
	NL-Tw	0.153	0.136	0.120	0.119	0.114	0.001	0.305
		(0.048)	(0.046)	(0.045)	(0.054)	(0.053)	(0.006)	(0.057)
	PL-mb	0.207	0.165	0.122	0.132	0.096	0.034	0.393
		(0.067)	(0.048)	(0.055)	(0.062)	(0.056)	(0.039)	(0.069)
	mgcv	0.152	0.123	0.063	0.072	0.059	0.024	0.292
		(0.052)	(0.048)	(0.039)	(0.047)	(0.046)	(0.047)	(0.067)
$N = 200$ $p = 100$	PL-Tw	0.244	0.184	0.131	0.130	0.155	0.003	0.446
		(0.094)	(0.078)	(0.085)	(0.077)	(0.104)	(0.024)	(0.115)
	PL-St	0.498	0.347	0.286	0.310	0.263	0.018	0.879
		(0.107)	(0.098)	(0.110)	(0.101)	(0.107)	(0.033)	(0.119)
	NL-Tw	0.219	0.171	0.185	0.179	0.197	0.003	0.462
		(0.082)	(0.076)	(0.071)	(0.064)	(0.102)	(0.021)	(0.106)
	PL-mb	0.420	0.327	0.218	0.233	0.208	0.018	0.748
		(0.126)	(0.104)	(0.112)	(0.086)	(0.079)	(0.034)	(0.133)
$N = 400$ $p = 100$	PL-Tw	0.175	0.132	0.084	0.086	0.081	0.001	0.291
		(0.054)	(0.050)	(0.056)	(0.059)	(0.054)	(0.009)	(0.063)
	PL-St	0.365	0.254	0.208	0.223	0.169	0.012	0.634
		(0.080)	(0.061)	(0.069)	(0.070)	(0.069)	(0.024)	(0.083)
	NL-Tw	0.165	0.127	0.128	0.123	0.115	0.000	0.314
		(0.048)	(0.045)	(0.055)	(0.050)	(0.055)	(0.000)	(0.058)
	PL-mb	0.307	0.231	0.164	0.176	0.143	0.009	0.546
		(0.069)	(0.067)	(0.067)	(0.063)	(0.059)	(0.019)	(0.079)

Figure 3.1: Boxplots of $\sqrt{\widehat{ISE}_{nonlinear}}$, $\sqrt{\widehat{ISE}_{linear}}$, as well as $\sqrt{\widehat{ISE}_{non-informative}}$ over 100 replicates by different approaches, for $p = 20$ and $C = 0.2$.

Figure 3.2: Boxplots of $\sqrt{\widehat{ISE}_{nonlinear}}$, $\sqrt{\widehat{ISE}_{linear}}$, as well as $\sqrt{\widehat{ISE}_{non-informative}}$ over 100 replicates by different approaches, for $p = 100$ and $C = 0.2$.

Figure 3.3: Boxplots of $\sqrt{\widehat{ISE}_{nonlinear}}$, $\sqrt{\widehat{ISE}_{linear}}$, as well as $\sqrt{\widehat{ISE}_{non-informative}}$ over 100 replicates by different approaches, for $p = 20$ and $C = 0.5$.

Figure 3.4: Boxplots of $\sqrt{\widehat{ISE}_{nonlinear}}$, $\sqrt{\widehat{ISE}_{linear}}$, as well as $\sqrt{\widehat{ISE}_{non-informative}}$ over 100 replicates by different approaches, for $p = 100$ and $C = 0.5$.

Figure 3.5: *Estimated functions by PL-Tw, for $N = 200$, $p = 20$, and $C = 0.2$. In each panel, light blue dot-dash curves correspond to estimates in 100 independent replicates, and the red solid curve indicates the underlying true function.*

the same numerical results. We have made the following observations and comments on the numerical results:

1. Our proposed approach “PL-Tw” provides the smallest testing error in every scenario.
2. “PL-St” and “NL-Tw” are competitive with our proposed approach in terms of testing errors.
3. “PL-mb” and “mgcv” perform less satisfyingly as for testing errors.

3.2.3 Variable Selection

Finally in this section, we evaluate the approaches with variable selection performances. The number of nonzero covariates selected (NV), the number of nonzero covariates selected which are truly informative (NVT), the number of linear covariates selected (NL), and the number of linear covariates selected which are truly linear (NLT) are calculated and presented for different scenarios in Tables 3.4 and 3.5. The average values over 100 replicates are provided together with the standard deviations in the brackets. Other quantities of interest can be extracted from these metrics. We have made the following observations and comments on the numerical results:

1. Our proposed approach “PL-Tw” generally outperforms all the other approaches with the ability to identify three categories of covariates, nonlinear, linear, and non-informative, with high accuracy.
2. “PL-St” and “PL-mb” always include too many covariates in the estimated models, which illustrate the power of twin boosting.

Table 3.3: *Average (standard deviation) of negative log partial likelihood on an independent testing dataset over 100 replicates, by different approaches for different scenarios.*

			PL-Tw	PL-St	NI-Tw	PL-mb	mgcv
$C = 0.2$	$p = 20$	$N = 200$	4.080	4.115	4.091	4.710	4.714
			(0.081)	(0.082)	(0.080)	(0.073)	(0.065)
		$N = 400$	4.073	4.094	4.079	4.689	4.692
			(0.067)	(0.064)	(0.067)	(0.063)	(0.062)
	$p = 100$	$N = 200$	4.087	4.158	4.094	4.764	-
			(0.079)	(0.082)	(0.081)	(0.065)	
		$N = 400$	4.074	4.116	4.079	4.720	-
			(0.085)	(0.083)	(0.086)	(0.065)	
$C = 0.5$	$p = 20$	$N = 200$	3.402	3.435	3.412	4.161	4.165
			(0.081)	(0.082)	(0.079)	(0.087)	(0.070)
		$N = 400$	3.394	3.414	3.399	4.142	4.140
			(0.078)	(0.075)	(0.079)	(0.067)	(0.071)
	$p = 100$	$N = 200$	3.404	3.475	3.412	4.217	-
			(0.084)	(0.087)	(0.088)	(0.079)	
		$N = 400$	3.390	3.430	3.395	4.171	-
			(0.093)	(0.093)	(0.094)	(0.079)

Figure 3.6: Boxplots of negative log partial likelihood on an independent testing dataset over 100 replicates by different approaches for $C = 0.2$.

Figure 3.7: Boxplots of negative log partial likelihood on an independent testing dataset over 100 replicates by different approaches for $C = 0.5$.

3. “Nl-Tw” is competitive in non-informative covariate detection. However it is of course unable to identify truly linear covariates, which illustrate the usefulness of PLAM.
4. Since “mgcv” does not provide indicators of variable selection results in its numerical outputs, we have to manually examine the estimated component functions by “mgcv” to obtain model identification results of it. Similarity is observed as the previous chapter, “mgcv” selects all covariates as nonlinear in all the replicates of all scenarios.
5. Twin boosting approaches “PL-Tw” and “Nl-Tw” implement variable selection at the cost of occasionally missing a few truly informative covariates.

3.3 Real Data Applications

The different methods are applied to a real dataset of diffuse large-B-cell lymphoma (DLBCL) provided by Rosenwald et al. (2002) to illustrate their abilities to identify microarray genes related to the survival time. The dataset includes a total of 240 patients (138 uncensored) and 7399 genes. Firstly the univariate proportional hazards model is fitted for each of the 7399 genes and the genes are ranked by the resulting fits. Then we apply the different approaches to the $p = 100$ genes with best individual fits only, similarly as done in Jenssen et al. (2002) and Lian et al. (2013). The dataset is divided into a training sample with size $N = 160$ and a validation sample with size $N_v = 80$ to decide stopping times. 30 random partitions are generated to investigate the stability of variable selection. Among the pre-selected 100 genes, 7 are selected as informative in at least half of the

Table 3.4: Average (standard deviation) of variable selection results over 100 replicates by different approaches for $C = 0.2$. “NV”, “NVT”, “NL”, and “NLT” indicate the number of nonzero covariates selected, the number of nonzero covariates selected which are truly informative, the number of linear covariates selected, and the number of linear covariates selected which are truly linear, respectively.

		NV	NVT	NL	NLT
$N = 200$	PL-Tw	6.22(1.23)	5.00(0.00)	3.18(0.85)	2.75(0.46)
$p = 20$	PL-St	15.5(2.19)	5.00(0.00)	3.21(1.47)	1.22(0.89)
	NI-Tw	5.40(0.59)	5.00(0.00)	-	-
	PL-mb	15.5(2.31)	5.00(0.00)	5.23(1.88)	1.63(0.86)
	mgcv	20.0(0.00)	5.00(0.00)	0.00(0.00)	0.00(0.00)
$N = 400$	PL-Tw	5.79(1.17)	5.00(0.00)	3.16(0.65)	2.90(0.30)
$p = 20$	PL-St	16.0(2.22)	5.00(0.00)	3.54(1.68)	1.36(0.90)
	NI-Tw	5.24(0.57)	5.00(0.00)	-	-
	PL-mb	16.2(2.10)	5.00(0.00)	4.83(1.94)	1.34(0.87)
	mgcv	20.0(0.00)	5.00(0.00)	0.00(0.00)	0.00(0.00)
$N = 200$	PL-Tw	7.12(2.14)	5.00(0.00)	3.89(1.03)	3.00(0.00)
$p = 100$	PL-St	35.0(7.29)	5.00(0.00)	13.0(3.53)	2.46(0.66)
	NI-Tw	5.90(1.05)	5.00(0.00)	-	-
	PL-mb	34.7(7.00)	5.00(0.00)	20.1(4.16)	2.68(0.51)
$N = 400$	PL-Tw	6.20(1.31)	5.00(0.00)	3.43(0.71)	2.98(0.14)
$p = 100$	PL-St	40.5(7.54)	5.00(0.00)	12.9(3.17)	2.26(0.76)
	NI-Tw	5.33(0.65)	5.00(0.00)	-	-
	PL-mb	41.0(8.34)	5.00(0.00)	22.1(4.73)	2.40(0.72)

Table 3.5: *Average (standard deviation) of variable selection results over 100 replicates by different approaches for $C = 0.5$. “NV”, “NVT”, “NL”, and “NLT” indicate the number of nonzero covariates selected, the number of nonzero covariates selected which are truly informative, the number of linear covariates selected, and the number of linear covariates selected which are truly linear, respectively.*

		NV	NVT	NL	NLT
$N = 200$	PL-Tw	6.22(1.07)	4.99(0.10)	3.13(0.80)	2.77(0.47)
$p = 20$	PL-St	15.6(2.12)	5.00(0.00)	3.42(1.55)	1.34(0.86)
	Nl-Tw	5.52(0.75)	4.99(0.10)	-	-
	PL-mb	15.1(2.50)	5.00(0.00)	5.30(1.85)	1.68(0.86)
	mgcv	20.0(0.00)	5.00(0.00)	0.00(0.00)	0.00(0.00)
$N = 400$	PL-Tw	5.73(1.03)	5.00(0.00)	3.14(0.51)	2.93(0.26)
$p = 20$	PL-St	15.6(2.28)	5.00(0.00)	3.40(1.68)	1.28(0.91)
	Nl-Tw	5.35(0.61)	5.00(0.00)	-	-
	PL-mb	16.2(1.91)	5.00(0.00)	4.72(1.93)	1.33(0.90)
	mgcv	20.0(0.00)	5.00(0.00)	0.00(0.00)	0.00(0.00)
$N = 200$	PL-Tw	7.66(1.98)	5.00(0.00)	4.01(1.13)	2.98(0.14)
$p = 100$	PL-St	34.3(7.58)	5.00(0.00)	12.5(3.64)	2.45(0.63)
	Nl-Tw	6.09(1.22)	4.99(0.10)	-	-
	PL-mb	34.4(7.39)	5.00(0.00)	19.7(4.59)	2.66(0.55)
$N = 400$	PL-Tw	6.48(1.67)	5.00(0.00)	3.42(0.81)	2.95(0.22)
$p = 100$	PL-St	38.6(8.13)	5.00(0.00)	13.3(3.10)	2.33(0.68)
	Nl-Tw	5.33(0.62)	5.00(0.00)	-	-
	PL-mb	40.0(8.07)	5.00(0.00)	21.8(4.46)	2.47(0.72)

independent replicates by our proposed approach, and we present numerical results for these 7 genes. It is worth mentioning that as the package *mgcv* fails since $p = 100$ is too large relative to $N = 160$, “mgcv” is applied with only the 7 genes as covariates. Variable selection results over 30 independent replicates for the 7 genes by different approaches are summarized in Table 3.6. For “PL-Tw”, “PL-St”, “PL-mb” and “mgcv”, the first number in each bracket is the number of replicates when the corresponding gene is identified as nonlinear and the second number is the number of times when the gene is identified as (nonzero) linear. Meanwhile for “NI-Tw”, each number indicates the number of times the gene is selected as informative. Also in Table 3.6, the average numbers (over 30 replicates) of nonlinear and linear covariates selected by different approaches (only one number for “NI-Tw” of course) are summarized. One replicate is randomly picked and the estimated component functions, for the selected 7 genes, are plotted by different approaches in Figures 3.8 ~ 3.12. We have made the following observations and comments on the numerical results:

1. We choose to present numerical results for the 7 genes as they are detected by our approach “PL-Tw” as informative (nonlinear or linear) in at least half of the replicates. It turns out that generally other approaches also select these 7 genes with high frequencies.
2. The 7 genes all rank top 40 in terms of individual fits, including the top 3 genes U15552, BC012161 and Hs.79123.
3. Among the 7 genes, some interesting ones are involved. For example, Hs.79123 is identified to be useful for developing cancer diagnostic and prognostic tools by Wang et al. (2004). Hs.301946 is a lymphoid blast crisis oncogene shown to have differential expressions in impaired

cells and healthy cells by Holm et al. (2004). U15552 is an acidic protein related gene claimed to be responsible for the transition of the cells from a normal to premalignant stage according to Lee et al. (2001). BC012161 is a gene in proliferation signature found to be cancer survival associated by Liu et al. (2005).

4. Among these 7 genes, 5 are found to have linear effects mostly, while the other 2 have nonlinear influences. Binder and Schumacher (2008) uses linear additive Cox's model to study the same dataset and several approaches are compared. The 5 linear genes among the 7 genes in our study are all selected as informative by approaches in Binder and Schumacher (2008), and we agree on that Hs.79123 has a decreasing component function while Hs.301946, *AA829241, U15552, and BC012161 have increasing ones.
5. Li and Luan (2005) also examines the same dataset and finds 24 important genes including BC012161 and *AA824616, which are among the 7 genes in our study.
6. The same dataset is studied in Lu and Li (2008) as well. 30 genes are reported to be important, but the list is not fully presented. However, among the only four important genes this work chooses to present, we find BC012161 which is also among the 7 genes of our study. Moreover, Lu and Li (2008) also finds *AA293559 an important gene which is detected as informative by our approach for 12 out of 30 replicates.
7. On average, the twin boosting approaches "PL-Tw" and "NI-Tw" provides sparser models, than the other approaches in our study as well as some existing studies like Li and Luan (2005), Binder and

Table 3.6: *DLBCL data. Variable selection results by different approaches.*

Gene Name	PL-Tw	PL-St	NI-Tw	PL-mb	mgcv
Hs.79123	(0,29)	(3,27)	28	(0,30)	(30,0)
Hs.301946	(0,28)	(6,23)	27	(4,26)	(30,0)
*AA829241	(0,28)	(5,24)	24	(2,28)	(30,0)
U15552	(0,20)	(0,27)	4	(0,27)	(30,0)
BC012161	(1,17)	(5,23)	20	(0,30)	(30,0)
*AA824616	(18,0)	(26,0)	15	(19,0)	(30,0)
M14745	(14,1)	(20,1)	12	(18,1)	(30,0)
Overall	(2.43,8.23)	(9.20,16.50)	9.33	(3.80,22.90)	-

Schumacher (2008), and Lu and Li (2008).

8. As for the estimated component functions, visual similarities are observed among different approaches.

Figure 3.8: *DLBCL data. Estimated component functions by PL-Tw.*

Figure 3.9: *DLBCL data. Estimated component functions by PL-St.*

Figure 3.10: DLBCL data. Estimated component functions by Nl-Tw.

Figure 3.11: *DLBCL data. Estimated component functions by PL-mb.*

Figure 3.12: *DLBCL* data. Estimated component functions by *mgcv*.

Chapter 4

Conclusions and Discussions

Partially linear additive model is a special form of generalized additive model which combines the advantages of the linear model which is concise and the nonlinear model which is flexible. Penalization methods and Bayesian methods are more popular among literatures on PLAM. Gradient boosting is an iterating approach which has shown its strength in numerous works in statistics and machine learning. However it has been applied mostly to linear models, semiparametric models with no attention to existence of purely linear covariates, or PLAM with pre-specification of linear components. Therefore in this thesis, we try to combine the two topics PLAM and boosting by proposing a gradient boosting approach for PLAM which can perform variable selection and linear component detection with no prior knowledge on model specification. The proposed approach is able to estimate the underlying component functions with indications about how predictors affect the conditional mean or quantiles of the response. Also the proposed boosting approach can be adapted to the context of survival analysis for partially linear additive proportional hazards models. Twin boosting technique is incorporated in the proposed approach to improve variable se-

lection and model specification. Numerical studies including simulations and real data applications are conducted. The proposed approach is compared with several other approaches including standard gradient boosting, boosting for purely nonlinear model, penalization approach, and Bayesian approach. Numerical results encouragingly illustrate the power of the proposed twin boosting approach in different regression settings: mean regression, quantile regression, and survival analysis with Cox's proportional hazards model. The proposed approach is competitive with, if not better than, other approaches in terms of the estimation accuracy and the prediction accuracy. The strength of the proposed method is highlighted with respect to the ability to identify informative covariates and to detect linear components.

In this thesis, numerical results are presented with respect to averages and standard deviations over independent replicates. More visible comparisons of different approaches are provided by boxplots. However more thorough assessments, in terms of whether there are important differences among approaches, can be conducted. Numerical results of independent replicates by different approaches can be treated as independent samples from different populations, thus hypothesis testing techniques can be applied to verify whether there are significant differences among different approaches. For instance, we compare twin boosting for PLAM ("PL-Tw") with twin boosting for nonlinear additive model ("NL-Tw") in mean regression with $N = 300$, $p = 15$, and noise terms following the Gaussian distribution, in terms of $\sqrt{\widehat{ISE}_1}$. Hypothesis testing technique for paired two group comparison is applied with null hypothesis that "PL-Tw" provides the same expected error as "NL-Tw", against alternative hypothesis that "PL-Tw" provides smaller expected error than "NL-Tw". Numerical

results of 100 independent replicates by the two approaches are treated as two independent samples with size 100 from two underlying distributions. It turns out that test statistic = -5.0928 , p-value $< 10^{-6}$, “PL-Tw” provides smaller expected error than “Nl-Tw” significantly. For another example, we compare the proposed approach (“PL-Tw”) with the R package *mgcv* in survival analysis with $N = 200$, $p = 20$, and $C = 0.2$, in terms of $\sqrt{\widehat{ISE}_1}$ as well. Test statistic = -0.4816 , p-value = 0.3156 , “PL-Tw” provides smaller error on average than “mgcv” but not significantly. This kind of comparison based on hypothesis testing was not constructed throughout in our numerical studies as it would have dramatically increased the amount of numerical results to present.

4.1 Discussions on Boosting and PLAM

4.1.1 Asymptotic Theories

Generally, it could be challenging to develop asymptotic theories for gradient boosting in various regression settings, with various loss functions, or with different types of base learners. As discussed in Chapter 1, theoretical justification of gradient boosting includes several core points which are technically reasonable and with theoretical justifications to some extent respectively. However, asymptotic analyses in an overall sense for gradient boosting estimators are relatively lacking. This is an important part of future works. Some limited theoretical results for gradient boosting can be viewed in Bühlmann and Yu (2003) and Zhang and Yu (2005), all with strong assumptions on loss functions or base learners.

Gradient boosting seeks to optimize a loss function in a greedy fash-

ion by iteratively selecting one of the base learners based on the updated sample data, given the current estimate resulted from previous iterations. To numerically implement the optimization, the gradient descent method is adopted so that in each iteration, the difficult function optimization problem is replaced by a least squares minimization (Friedman (2001)). The gradient descent method works as follows.

Gradient descent algorithm

To solve multivariable minimization problem $\min_x f(x)$:

1. Initialize $m = 0$ and $x^{[0]}$.
2. Compute the negative gradient direction at the current point $x^{[m]}$:

$$u^{[m]} = -\nabla f(x^{[m]}).$$

3. Do line search (e.g. Newton's method, golden section method) to find $\nu^{[m]} = \arg \min_{\nu} f(x^{[m]} + \nu u^{[m]})$.
 4. Update $x^{[m+1]} = x^{[m]} + \nu^{[m]} u^{[m]}$. Increase m by 1.
 5. Iterate steps 2~4 until some stopping criterion is met.
-

Sometimes an approximated gradient direction is used in variants of gradient descent method for reasons, like that exact gradient may not be analytically available (Snyman (2005)). In the context of gradient boosting, $f(\eta) \triangleq \frac{1}{N} \sum_{i=1}^N L(Y_i, \eta_i)$ is minimized. In each iteration, $\eta^{[m]}$ is updated

not strictly in the negative gradient direction $u^{[m]}$, but in the direction of $g^{[m]}$, the base learner which fits $u^{[m]}$ the best in the sense of least squares. Moreover, the line searching step is replaced by a fixed small step size ν ($= 0.1$ in numerous existing literatures as well as this thesis) in gradient boosting, to reduce computational costs.

General theoretical property for twin boosting is a relatively open question so far as well. To our knowledge, only twin boosting for mean regression with purely linear additive models, using univariate least squares as base learners, has been proved to approximate the adaptive LASSO (Zou (2006), Bühlmann and Hothorn (2010)).

4.1.2 More Types of Predictor Effects

In this thesis, as we are focusing on PLAM, two types of base learners, simple linear regressions and penalized splines, are applied in the boosting approaches. However, there can be more types of predictor effects in other models, therefore it can be interesting to develop more types of base learners correspondingly for specific covariate effects. One example is the interaction of covariates, which is especially popular in geostatistics due to the existence of spatial effects. Kneib et al. (2009) focuses on the model choice in geoadditive regression models and develops boosting approaches incorporating spatial effects and varying coefficient terms. Robinson and Hothorn (2010) studies the spatio-temporal structured additive models with boosting, base learners for interactive effects are involved as well. Sobotka and Kneib (2012) develops boosting approaches for geoadditive expectile regression and also involves bivariate penalized spline base learners.

The mixed model is another model getting increasingly popular which takes random effects into consideration. Groll and Tutz (2012) studies the

generalized additive mixed models with likelihood-based boosting.

The constrained regression pays attention to prior knowledge on covariate effects which should be incorporated. Leitenstorfer and Tutz (2007) applies likelihood-based boosting to monotonic regression for generalized additive models. Hofner et al. (2016) shares a similar motivation while involves cyclic effects, monotonic effects, as well as effects with boundary constraints.

Boosting method which selects a single base learner in each iteration, like in this thesis, is commonly called component-wise boosting. In contrast, block-wise boosting updates blocks of predictors in each iteration. Tutz and Ulbricht (2009) applies block-wise boosting in penalized linear regression with correlation-based penalty, motivated by the assumption that highly correlated covariates tend to be in or out of the model together. Tutz and Gertheiss (2010) develops boosting technique in signal regression with functional data, which in each iteration selects groups of variables where the grouping is based on some metric related to the signal. Gertheiss et al. (2011) focuses on ordinal predictors and a group of dummy covariates are associated to each of the predictor. Thus block-wise boosting is applied so that covariates in a same group are selected simultaneously.

After all, measurement of model complexity (degrees of freedom in this thesis) is needed to control different types of base learners, for a “fair” selection by boosting.

4.1.3 The Degree and Internal Knots of Spline Basis

As we use penalized splines to recover nonlinear parts of covariate effects in PLAM, it is a natural question that how to choose the degree and internal knots of spline basis. And for the internal knots, the number and locations

of knots need to be decided. In this thesis, we have made some computationally convenient choices and fixed the choices throughout in all settings and scenarios. Future work includes to adopt techniques of degree and knots selection in a data-driven fashion and to compare the performances.

As for the degree of spline basis, $q = 3$ may be the most popular choice in use (Eubank et al. (2012)), resulting in piecewise cubic splines with global continuous second derivatives. Lower degrees can lead to worse fits and the loss of smoothness. However, as argued in Ruppert et al. (2003), when there are sufficient number of knots, even piecewise linear splines can be applicable. On the other hand, higher degrees may improve the fit, but maybe very modestly (Ruppert et al. (2003)), at the cost of greater computational burden. The choice of degree can also rely on the smoothness of underlying functions to be estimated, if any prior knowledge is available. An adaptive degree selection algorithm for spline smoothing is developed in Eubank et al. (2012).

The number of knots of spline basis is a value controlling, to some extent, the trade-off of model fit and model complexity. Fewer knots favor less complex models while more knots favor better fitted models (Ruppert et al. (2003)). One can treat the number of knots as a tuning parameter of the model, and apply model selection techniques like cross validations to decide (Lian et al. (2012)). A likelihood-based criterion to select the number of basic functions in penalized spline regression is proposed in Kauermann and Opsomer (2011). Algorithms are developed for the automatic selection of the number of knots in Ruppert (2012).

Besides the number of knots, the locations need to be decided as well. Other than pre-decided equidistant knots (Lian et al. (2013), Hu et al. (2015)), knots decided by sample quantiles are also popular in use, so that

there are a fixed number of observations between each pair of neighbouring knots (Ruppert et al. (2003)). A more complex knot location selection scheme for spline regression is proposed in Zhou and Shen (2001).

4.1.4 Degrees of Freedom

To recover model structure of PLAM, two types of base learners are applied in our proposed approach. To achieve a “fair” selection between the linear regression base learners and the penalized spline base learners, we control them to be comparable by enforcing equal degrees of freedom, which are seen to measure the “effective number of parameters”. In more complex regression frameworks (compared to ordinary linear regressions, like penalized linear regressions and spline regressions), when the estimated responses $\hat{Y}$ can be formularized to be linear in observed responses Y : $\hat{Y} = SY$ (S is widely known as the hat matrix), some quantities are involved in the analyses, and are seen as analogs of p (the number of covariates) in ordinary linear regressions. This is where the multiple definitions of degrees of freedom arise. For example, $\text{trace}(S)$ is involved in the so-called C_p statistic (Mallows (1973)), which corrects the residual sum of squares (RSS) to make it unbiased for the true mean squared error (MSE) (Buja et al. (1989)), with corresponding quantity in ordinary linear regression being p . Also $\text{trace}(2S - S^T S)$ is involved in the expected RSS (Buja et al. (1989)), with corresponding quantity in ordinary linear regression being p as well. Therefore, these two traces have both become definitions of degrees of freedom (Hastie and Tibshirani (1990), Ruppert et al. (2003)).

In this thesis we have used the very conventional definition of degrees of freedom $\text{df} = \text{trace}(S)$, while $\text{df} = \text{trace}(2S - S^T S)$ is used in Hofner et al. (2011) and argued to be unbiased in selection. To our knowledge, these

two are both in popular use and perform similarly in practice. Here is a quick comparison of both definitions in terms of numerical performances, in mean and quantile regressions with $N = 300$, $p = 15$, and the noise terms following the Gaussian distribution. Numerical results of two approaches “PL-Tw” and “PL-mb” are displayed in Tables 4.1 ~ 4.5. Similarities are observed from both definitions of degrees of freedom, while neither one is able to dominate empirically.

4.1.5 Stability Selection

It has been widely noticed that boosting “forever” results in the overfitting problem. Overfitting is illustrated by a quick numerical experiment here. Settings follow Chapter 2 with $N = 300$ (for both the training and testing samples), $p = 15$, and Gaussian noises for mean regression only. $RMSE$ and ADE are both calculated, and training and testing errors are displayed together against the number of iterations in Fig 4.1. For a nicer graph, error to minimum error ratios are plotted instead of the errors and the plots are restricted to iterations 500 to 5000. It is observed that, testing errors firstly decrease then increase, while training errors always decrease. Therefore early stopping techniques are commonly embedded in boosting approaches. Cross validation may be one of the mostly applied early stopping techniques, and we apply 5-fold cross validation in our simulation studies. However, it is empirically found that even with early stopping, boosting does not perform satisfying variable selection, since base learners associated with truly non-informative covariates are still selected, although occasionally. Therefore twin boosting is incorporated in our proposed approaches.

Another technique named stability selection is developed for the purpose of variable selection in Meinshausen and Bühlmann (2010), improved

Table 4.1: Comparison of two definitions of degrees of freedom. Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates, for $N = 300$, $p = 15$, and Gaussian errors.

		f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
Mean	PL-Tw	0.094	0.082	0.055	0.077	0.034	0.040	0.037	0.005
	df = trace(S)	(0.025)	(0.023)	(0.020)	(0.023)	(0.024)	(0.026)	(0.024)	(0.012)
	PL-Tw	0.102	0.084	0.052	0.083	0.036	0.033	0.036	0.007
	df = trace($2S - S^T S$)	(0.018)	(0.018)	(0.021)	(0.023)	(0.031)	(0.025)	(0.026)	(0.014)
	PL-mb	0.162	0.096	0.057	0.102	0.062	0.059	0.054	0.037
	df = trace(S)	(0.018)	(0.025)	(0.020)	(0.019)	(0.034)	(0.032)	(0.030)	(0.027)
	PL-mb	0.158	0.100	0.059	0.101	0.061	0.053	0.055	0.042
	df = trace($2S - S^T S$)	(0.017)	(0.027)	(0.022)	(0.021)	(0.032)	(0.032)	(0.030)	(0.030)
$\tau = 0.1$	PL-Tw	0.184	0.174	0.078	0.154	0.067	0.084	0.105	0.002
	df = trace(S)	(0.037)	(0.062)	(0.036)	(0.055)	(0.053)	(0.057)	(0.079)	(0.010)
	PL-Tw	0.200	0.229	0.078	0.181	0.061	0.070	0.083	0.000
	df = trace($2S - S^T S$)	(0.035)	(0.082)	(0.039)	(0.064)	(0.046)	(0.059)	(0.053)	(0.001)
	PL-mb	0.252	0.204	0.102	0.281	0.130	0.113	0.087	0.052
	df = trace(S)	(0.043)	(0.045)	(0.047)	(0.064)	(0.070)	(0.066)	(0.049)	(0.043)
	PL-mb	0.256	0.214	0.110	0.303	0.135	0.120	0.097	0.054
	df = trace($2S - S^T S$)	(0.047)	(0.048)	(0.048)	(0.067)	(0.064)	(0.085)	(0.052)	(0.050)
$\tau = 0.3$	PL-Tw	0.143	0.116	0.056	0.094	0.038	0.044	0.052	0.001
	df = trace(S)	(0.021)	(0.033)	(0.021)	(0.027)	(0.026)	(0.033)	(0.033)	(0.006)
	PL-Tw	0.166	0.133	0.058	0.110	0.039	0.039	0.044	0.000
	df = trace($2S - S^T S$)	(0.021)	(0.035)	(0.023)	(0.029)	(0.028)	(0.027)	(0.037)	(0.000)
	PL-mb	0.193	0.134	0.077	0.148	0.081	0.070	0.060	0.042
	df = trace(S)	(0.022)	(0.028)	(0.028)	(0.036)	(0.048)	(0.040)	(0.036)	(0.032)
	PL-mb	0.192	0.134	0.089	0.154	0.087	0.075	0.068	0.043
	df = trace($2S - S^T S$)	(0.018)	(0.029)	(0.032)	(0.040)	(0.045)	(0.045)	(0.042)	(0.036)

Table 4.2: Comparison of two definitions of degrees of freedom. Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates, for $N = 300$, $p = 15$, and Gaussian errors. (Continued)

		f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
$\tau = 0.5$	PL-Tw	0.136	0.112	0.054	0.086	0.038	0.042	0.048	0.000
	df = trace(S)	(0.024)	(0.030)	(0.021)	(0.026)	(0.028)	(0.033)	(0.034)	(0.002)
	PL-Tw	0.162	0.125	0.062	0.101	0.039	0.035	0.044	0.000
	df = trace($2S - S^T S$)	(0.019)	(0.027)	(0.028)	(0.026)	(0.029)	(0.025)	(0.037)	(0.000)
	PL-mb	0.189	0.131	0.073	0.121	0.069	0.062	0.056	0.040
	df = trace(S)	(0.020)	(0.027)	(0.027)	(0.026)	(0.041)	(0.038)	(0.033)	(0.029)
	PL-mb	0.187	0.131	0.081	0.120	0.071	0.064	0.056	0.040
	df = trace($2S - S^T S$)	(0.019)	(0.028)	(0.033)	(0.033)	(0.042)	(0.039)	(0.030)	(0.030)
$\tau = 0.7$	PL-Tw	0.142	0.124	0.058	0.088	0.041	0.052	0.050	0.000
	df = trace(S)	(0.028)	(0.040)	(0.025)	(0.028)	(0.034)	(0.036)	(0.042)	(0.002)
	PL-Tw	0.171	0.149	0.060	0.101	0.048	0.050	0.048	0.000
	df = trace($2S - S^T S$)	(0.021)	(0.037)	(0.023)	(0.025)	(0.030)	(0.032)	(0.039)	(0.001)
	PL-mb	0.197	0.142	0.075	0.112	0.074	0.063	0.059	0.041
	df = trace(S)	(0.024)	(0.029)	(0.030)	(0.022)	(0.044)	(0.034)	(0.038)	(0.027)
	PL-mb	0.193	0.139	0.076	0.110	0.069	0.058	0.056	0.043
	df = trace($2S - S^T S$)	(0.018)	(0.032)	(0.029)	(0.024)	(0.041)	(0.033)	(0.039)	(0.025)
$\tau = 0.9$	PL-Tw	0.185	0.204	0.091	0.116	0.068	0.079	0.086	0.001
	df = trace(S)	(0.034)	(0.071)	(0.039)	(0.047)	(0.047)	(0.059)	(0.070)	(0.009)
	PL-Tw	0.218	0.262	0.083	0.132	0.074	0.073	0.084	0.000
	df = trace($2S - S^T S$)	(0.042)	(0.081)	(0.043)	(0.042)	(0.058)	(0.049)	(0.052)	(0.000)
	PL-mb	0.258	0.223	0.129	0.142	0.136	0.111	0.093	0.047
	df = trace(S)	(0.043)	(0.050)	(0.062)	(0.030)	(0.061)	(0.065)	(0.064)	(0.034)
	PL-mb	0.268	0.242	0.145	0.146	0.131	0.121	0.096	0.050
	df = trace($2S - S^T S$)	(0.044)	(0.047)	(0.070)	(0.029)	(0.055)	(0.060)	(0.061)	(0.034)

Table 4.3: *Comparison of two definitions of degrees of freedom. Average (standard deviation) of testing errors in mean regression over 100 replicates, for $N = 300$, $p = 15$, and Gaussian errors. “RMSE” and “ADE” indicate the root mean squared errors and the absolute deviation errors, respectively.*

		df = trace(S)	df = trace($2S - S^T S$)
RMSE	PL-Tw	0.550(0.015)	0.553(0.016)
	PL-mb	0.585(0.017)	0.588(0.016)
ADE	PL-Tw	0.425(0.012)	0.427(0.012)
	PL-mb	0.458(0.014)	0.459(0.014)

Table 4.4: *Comparison of two definitions of degrees of freedom. Average (standard deviation) of check loss errors in quantile regression over 100 replicates, for $N = 300$, $p = 15$, and Gaussian errors.*

		df = trace(S)	df = trace($2S - S^T S$)
$\tau = 0.1$	PL-Tw	0.107(0.007)	0.112(0.008)
	PL-mb	0.122(0.008)	0.125(0.010)
$\tau = 0.3$	PL-Tw	0.193(0.006)	0.197(0.006)
	PL-mb	0.210(0.008)	0.212(0.009)
$\tau = 0.5$	PL-Tw	0.219(0.007)	0.225(0.008)
	PL-mb	0.237(0.008)	0.239(0.008)
$\tau = 0.7$	PL-Tw	0.194(0.007)	0.199(0.007)
	PL-mb	0.210(0.007)	0.210(0.007)
$\tau = 0.9$	PL-Tw	0.107(0.008)	0.112(0.007)
	PL-mb	0.119(0.007)	0.123(0.006)

Table 4.5: Comparison of two definitions of degrees of freedom. Average (standard deviation) of variable selection results over 100 replicates, for $N = 300$, $p = 15$, and Gaussian errors. “NV”, “NVT”, “NL”, and “NLT” indicate the number of nonzero covariates selected, the number of nonzero covariates selected which are truly informative, the number of linear covariates selected, and the number of linear covariates selected which are truly linear, respectively.

			NV	NVT	NL	NLT
Mean	PL-Tw	df = trace(S)	8.54(1.28)	7.00(0.00)	2.91(0.71)	2.58(0.62)
		df = trace($2S - S^T S$)	9.39(1.22)	7.00(0.00)	3.33(1.12)	2.44(0.77)
	PL-mb	df = trace(S)	14.6(0.82)	7.00(0.00)	0.86(0.95)	0.30(0.50)
		df = trace($2S - S^T S$)	14.5(0.82)	7.00(0.00)	1.00(1.05)	0.31(0.55)
$\tau = 0.1$	PL-Tw	df = trace(S)	7.35(0.64)	6.97(0.17)	3.14(0.51)	2.92(0.27)
		df = trace($2S - S^T S$)	7.43(0.77)	6.99(0.18)	3.31(0.48)	3.00(0.00)
	PL-mb	df = trace(S)	15.0(0.41)	7.00(0.00)	1.01(1.11)	0.23(0.47)
		df = trace($2S - S^T S$)	15.0(0.00)	7.00(0.00)	1.05(1.11)	0.20(0.41)
$\tau = 0.3$	PL-Tw	df = trace(S)	7.27(0.49)	7.00(0.00)	3.09(0.35)	2.97(0.17)
		df = trace($2S - S^T S$)	7.22(0.50)	7.00(0.00)	3.12(0.43)	2.93(0.25)
	PL-mb	df = trace(S)	14.9(0.39)	7.00(0.00)	0.89(1.10)	0.28(0.51)
		df = trace($2S - S^T S$)	14.9(0.57)	7.00(0.00)	0.68(0.84)	0.25(0.52)
$\tau = 0.5$	PL-Tw	df = trace(S)	7.17(0.43)	7.00(0.00)	3.04(0.32)	2.97(0.17)
		df = trace($2S - S^T S$)	7.19(0.40)	7.00(0.00)	3.00(0.37)	2.91(0.25)
	PL-mb	df = trace(S)	14.9(0.39)	7.00(0.00)	1.05(1.05)	0.33(0.57)
		df = trace($2S - S^T S$)	15.0(0.18)	7.00(0.00)	0.77(0.82)	0.18(0.38)
$\tau = 0.7$	PL-Tw	df = trace(S)	7.25(0.48)	7.00(0.00)	3.01(0.41)	2.91(0.29)
		df = trace($2S - S^T S$)	7.21(0.50)	7.00(0.00)	3.02(0.32)	2.97(0.18)
	PL-mb	df = trace(S)	14.9(0.33)	7.00(0.00)	0.99(1.07)	0.30(0.56)
		df = trace($2S - S^T S$)	15.0(0.00)	7.00(0.00)	0.90(1.03)	0.30(0.47)
$\tau = 0.9$	PL-Tw	df = trace(S)	7.43(0.66)	6.98(0.14)	3.00(0.53)	2.84(0.37)
		df = trace($2S - S^T S$)	7.24(0.43)	7.00(0.00)	3.20(0.48)	2.93(0.25)
	PL-mb	df = trace(S)	14.9(0.43)	7.00(0.00)	1.09(1.32)	0.28(0.53)
		df = trace($2S - S^T S$)	15.0(0.18)	7.00(0.00)	1.32(1.37)	0.30(0.54)

Figure 4.1: *Overfitting. Error to minimum error ratios over iterating. Training errors against testing errors.*

with a complementary subsampling strategy in Shah and Samworth (2013), and adapted to boosting in Hofner et al. (2014). The core idea of stability selection is quite intuitive, that by a number of subsamplings, only covariates which are selected with high frequencies should be included in the model. We associate the R package *mboost* with *stabs*, the R package implementing the stability selection method, and apply them as well in our simulation studies. It is observed that stability selection may sometimes be slightly more accurate in detecting three types of covariates in PLAM, than our proposed twin boosting approach is, but at the cost of extremely large computational burden.

Here a quick comparison between twin boosting and stability selection is provided in the case where $N = 300$ and $p = 15$ for mean regression only. The stability selection method has three tuning parameters: maximum acceptable per-family error rate ($PFER_{max}$), threshold for selection frequency (π_{thr}), and number of selected base learners per boosting run (q). References instruct users to specify $PFER_{max}$ and π_{thr} , then q will be automatically computed according to proposed results in some related literatures. However, without pre-knowledge of the underlying model, desired values of $PFER_{max}$ and π_{thr} may lead to undesired values of q . A too small q may cause the problem of missing informative variables while a too large q may make noisy variables involved by the method. Therefore, a tuning process is still necessary. We partly follow simulation settings of Hofner et al. (2014) to specify $PFER_{max} = 5$ or 10 , and $\pi_{thr} = 0.75$ or 0.9 . As for sampling types and assumptions on the distributions of the selection probabilities, three choices in Hofner et al. (2014) are exactly copied: “SS”-“unimodel”, “SS”-“r-concave”, and “MB”-“none” (also seen in manual of package *stabs*). For a single combination of noise type, sam-

pling type, assumption on the distributions of the selection probabilities, $PFER_{max}$, and π_{thr} , the average, over independent replicates, computing times for the proposed twin boosting approach, *mboost* only, and *mboost* associated with stability selection are, empirically and approximately, 45 seconds, 40 seconds, and 50 minutes respectively on our desktop. Variable selection results are displayed in Tables 4.6 and 4.7, and it needs to be mentioned that the case where sampling type=“SS”, assumption=“unimodel”, $PFER_{max} = 10$, and $\pi_{thr} = 0.75$ fails to complete due to an error encountered in calculating q thus the package rejects this combination of tuning parameters.

4.2 Discussions on Quantile Regression

Kneib (2013) highlights quantile regression, expectile regression (Newey and Powell (1987)), and the generalized additive model for location, scale, and shape (GAMLSS) (Rigby and Stasinopoulos (2005), Mayr et al. (2012a)) as important regression models that go beyond mean regression, so that lead to more complete analyses of how the response is related to the predictors. GAMLSS is more flexible than the classical generalized additive model and especially favourable in cases of heavily skewed or kurtotic distributions, while quantile regression and expectile regression are distribution-free by avoiding assumptions on conditional distributions and directly relating predictors to quantiles or expectiles (Mayr et al. (2014)). Compared to expectile regression, quantile regression has the advantage of its interpretability. Furthermore, quantile regression can be applied to construct prediction intervals (Mayr et al. (2012b)). Another model that goes beyond mean regression is the conditional transformation model which allows

Table 4.6: Comparison of twin boosting and stability selection. Average (standard deviation) of variable selection results over 100 replicates, for $N = 300$, $p = 15$, and Gaussian errors, mean regression only. “NV”, “NVT”, “NL”, and “NLT” indicate the number of nonzero covariates selected, the number of nonzero covariates selected which are truly informative, the number of linear covariates selected, and the number of linear covariates selected which are truly linear, respectively.

			NV	NVT	NL	NLT
Twin boosting			8.54(1.28)	7.00(0.00)	2.91(0.71)	2.58(0.62)
“SS” “unimodel”	$PFER_{max} = 5$	$\pi_{thr} = 0.75$	6.94(0.49)	6.85(0.38)	2.67(0.61)	2.60(0.50)
		$\pi_{thr} = 0.9$	7.30(0.60)	7.00(0.00)	2.60(0.62)	2.50(0.57)
	$PFER_{max} = 10$	$\pi_{thr} = 0.75$	-	-	-	-
		$\pi_{thr} = 0.9$	13.9(1.20)	7.00(0.00)	0.30(0.60)	0.26(0.52)
“SS” “r-concave”	$PFER_{max} = 5$	$\pi_{thr} = 0.75$	7.53(0.63)	7.00(0.00)	2.54(0.68)	2.49(0.63)
		$\pi_{thr} = 0.9$	6.97(0.41)	6.90(0.31)	2.75(0.57)	2.71(0.52)
	$PFER_{max} = 10$	$\pi_{thr} = 0.75$	7.56(0.63)	7.00(0.00)	2.53(0.68)	2.46(0.63)
		$\pi_{thr} = 0.9$	6.99(0.41)	6.90(0.31)	2.78(0.57)	2.72(0.52)
“MB” “none”	$PFER_{max} = 5$	$\pi_{thr} = 0.75$	5.53(0.73)	5.53(0.73)	2.08(0.78)	1.74(0.57)
		$\pi_{thr} = 0.9$	6.47(0.57)	6.47(0.57)	2.50(0.63)	2.47(0.57)
	$PFER_{max} = 10$	$\pi_{thr} = 0.75$	6.98(0.41)	6.87(0.35)	2.70(0.60)	2.64(0.55)
		$\pi_{thr} = 0.9$	7.00(0.37)	6.92(0.25)	2.73(0.58)	2.70(0.53)

Table 4.7: Comparison of twin boosting and stability selection. Average (standard deviation) of variable selection results over 100 replicates, for $N = 300$, $p = 15$, and Student's t errors, mean regression only. “NV”, “NVT”, “NL”, and “NLT” indicate the number of nonzero covariates selected, the number of nonzero covariates selected which are truly informative, the number of linear covariates selected, and the number of linear covariates selected which are truly linear, respectively.

			NV	NVT	NL	NLT
Twin boosting			8.71(1.42)	6.96(0.24)	2.85(0.74)	2.47(0.64)
“SS” “unimodel”	$PFER_{max} = 5$	$\pi_{thr} = 0.75$	6.80(0.55)	6.74(0.58)	2.70(0.70)	2.63(0.72)
		$\pi_{thr} = 0.9$	7.09(0.52)	6.90(0.31)	2.62(0.56)	2.58(0.57)
	$PFER_{max} = 10$	$\pi_{thr} = 0.75$	-	-	-	-
		$\pi_{thr} = 0.9$	13.7(1.36)	7.00(0.00)	0.33(0.61)	0.29(0.58)
“SS” “r-concave”	$PFER_{max} = 5$	$\pi_{thr} = 0.75$	7.26(0.64)	6.90(0.31)	2.63(0.56)	2.59(0.57)
		$\pi_{thr} = 0.9$	6.76(0.63)	6.73(0.58)	2.70(0.58)	2.62(0.72)
	$PFER_{max} = 10$	$\pi_{thr} = 0.75$	7.27(0.64)	6.90(0.31)	2.62(0.56)	2.58(0.57)
		$\pi_{thr} = 0.9$	6.78(0.63)	6.72(0.58)	2.70(0.70)	2.63(0.72)
“MB” “none”	$PFER_{max} = 5$	$\pi_{thr} = 0.75$	5.47(0.82)	5.47(0.82)	1.91(0.69)	1.70(0.65)
		$\pi_{thr} = 0.9$	6.20(0.89)	6.20(0.89)	2.40(0.56)	2.31(0.66)
	$PFER_{max} = 10$	$\pi_{thr} = 0.75$	6.76(0.63)	6.73(0.58)	2.61(0.76)	2.54(0.77)
		$\pi_{thr} = 0.9$	6.79(0.63)	6.75(0.58)	2.60(0.77)	2.55(0.78)

higher moments of the conditional distribution to depend on the explanatory variables (Hothorn et al. (2014)).

4.2.1 Quantile Crossing

One question which is not addressed in this thesis is the quantile crossing problem. When multiple quantile levels are desired, regressions are implemented by our proposed approach individually in a one-level-at-a-time fashion, and so far no technique is involved to guarantee that $\widehat{Q}_{\tau_1}[Y|\mathbf{X}] < \widehat{Q}_{\tau_2}[Y|\mathbf{X}]$ for $\tau_1 < \tau_2$ and any given $\mathbf{X}$. Other approaches including the package *mboost* and the Bayesian approach “mcmc” share the same problem, while the package *mgcv* does not since it implements quantile estimation by associating mean regression with conditional variance estimation. Illustrations can be found in the two real world examples in Chapter 2, where it is observed that estimated quantile curves at different levels may cross.

He (1997) imposes monotonicity in quantile regression by adopting the location-scale model and suggests a three-step strategy. A related work using doubly penalized kernel machine can be found in Shim et al. (2009). Takeuchi et al. (2006) deals with the quantile crossing problem by simultaneously estimating for all desired quantile levels and incorporating the non-crossing constraint into the optimization. Schnabel and Eilers (2013) defines the quantile sheet which also estimates quantile curves at multiple levels simultaneously to avoid crossings. Dette and Volgushev (2008) starts with a non-parametric initial estimate and provides estimated quantile function which is automatically monotone with respect to the quantile level. Chernozhukov et al. (2010) proposes a technique to rearrange original estimated non-monotone quantile curve. This technique is adopted in Taieb

et al. (2015) to forecast probabilistic time series.

4.2.2 Discrete Responses

At first, it might appear anomalous to consider quantile regression for discrete responses, since for discrete random variables, population quantiles may not be even well defined. However, if a continuous latent variable is introduced, things become less anomalous (Koenker (2005)). Take the binary response (Y) for instance, if a continuous latent variable (Z) is introduced as $Y = \mathbf{1}_{[Z>0]}$, quantile regression can be applied to the situation. Manski (1975) develops the maximum score estimator, which is equivalent to applying median regression (quantile regression at level 0.5) to the binary response model with latent variable. This estimator is further studied in Manski (1985) with some theoretical properties proved. Horowitz (1992) modifies the maximum score estimator with a smoothed version of score function. This approach is extended to quantiles other than only the median in Kordas (2006). Compared to conventional likelihood-based models dealing with binary responses, like the logistic regression, the strength of the above approach is its flexibility. If the covariates are assumed to act as pure location shifts in the latent variable model, conventional methods are appropriate. Meanwhile, if more complex covariate effects, like location-scale shifts, are allowed, the above approach adopting quantile regression is less restrictive (Koenker (2005)).

Another example of discrete responses to adopt quantile regression is the case of count data. Machado and Silva (2005) imposes artificial smoothness to count data by jittering then applies standard quantile regression techniques. Again, such approach has the advantage of flexibility compared to conventional models like the Poisson regression (Koenker (2005)).

4.3 Discussions on Survival Analysis

To study survival data, in this thesis we have applied the Cox's proportional hazards model. However, there are other models or frameworks dealing with survival data. For example, censored survival data can be adapted to standard L_2 loss by incorporating the inverse probability of censoring (IPC) weights (van der Laan and Robins (2003)). Hothorn et al. (2006) applies boosting in this framework. The accelerated failure time (AFT) model is a popular alternative to Cox's model, and assumes that the effect of a covariate is to accelerate or decelerate the survival time by some constant. Schmid and Hothorn (2008), Wang and Wang (2010), Johnson and Long (2011), and Ma et al. (2012) study the AFT model with boosting. Moreover, Binder et al. (2009) takes multiple causes of failure (competing risks) into consideration.

4.3.1 Baseline Hazard Function

In survival analysis, so far our proposed method has not included an approach to estimate the baseline hazard function $\lambda_0()$, since in the proportional hazards model, $\lambda_0()$ is assumed to be independent with the predictors and this thesis mainly focuses on the recovery of PLAM structure which indicates how the response is influenced by predictors. However, estimation of the baseline hazard function is needed if more complete knowledge on the distribution of survival time is desired. Breslow (1972), as a discussion paper of Cox (1972), proposes a maximum likelihood estimator for the cumulative baseline hazard function ($\Lambda_0(t) = \int_0^t \lambda_0(u)du$), which is known as the Breslow estimator and has been popular in use (Lin (2007)). Fan and Li (2002), van Houwelingen et al. (2006), and Leng and Zhang (2007) all apply

this estimator in their approaches. However, Breslow's approach results in a piecewise constant estimated cumulative baseline hazard function. There are also works developing techniques to obtain smoothed estimated baseline hazard functions, for example, Gray (1990) and Betensky et al. (2002).

4.3.2 Tied Times

Another question which is not addressed in this thesis is that, in survival analysis, it is usually but not necessarily the case to assume that the survival time follows a continuous distribution. However the problem may arise, more likely in the case where the survival time follows a discrete distribution, that there can be tied survival times among observations. Recalling the interpretation of partial likelihood (3.1.1) as the product of a sequence of conditional probabilities, it is argued that the partial likelihood is not strictly appropriate when tied survival times exist. There are approaches proposed by literatures to handle ties in Cox's model, for example, the discrete hazard logistic model (Cox and Oakes (1984)), the Breslow approximation (Breslow (1972), Breslow (1974)) and the Efron approximation (Efron (1977)) of exact partial likelihood, as well as the averaged likelihood approach (Therneau and Grambsch (2000)).

4.3.3 Time-dependent Covariates

Since the survival time measures the length of a time period, it is natural to consider whether the covariates can be time-dependent. In literatures, there are commonly two types to formularize a time-dependent covariate. Take the j -th covariate X_j as an example, one type is that the covariate has a constant value (for each subject) but a time-varying effect on the

response, $f_j(X_j; t) = \beta_j(t) \cdot X_j$. The other type is that the covariate itself is time-varying, $\mathbf{X}_j = X_j(t)$. Zucker and Karr (1990) develops a penalized partial likelihood approach for the first type of time-dependent covariates. Hofner et al. (2013) also considers the first type and applies boosting with base learners designed for time-dependent covariate effects. In this work, a selection bias by boosting in favor of time-varying effects is observed. It would be interesting to develop techniques to deal with this bias. As for the second type of time-dependent covariate, Therneau and Grambsch (2000) mentions a trick that, if the covariate values only vary occasionally, one subject can be separated into a few observations. For instance, a subject with observed survival time T_1 has original covariate value $X_j = 0$, but a change happens to this subject at time T_0 with $X_j = 1$ ever since. Then two observations, one with $X_j = 0$ and censoring time T_0 , and the other one with $X_j = 1$ and survival time $T_1 - T_0$, are added to the sample. Cox and Oakes (1984) introduces likelihood function for the linear additive proportional hazards model with a general setting where covariate values can be “always” varying. However the optimization of this likelihood function with respect to the linear coefficients can be computationally expensive. It is also discussed in Fisher and Lin (1999).

Appendix

A.1 Proof of Proposition 2.3

Let Y be any random variable.

$$\begin{aligned} & \mathbb{E}_Y[(Y - \eta)^2] \\ = & \mathbb{E}_Y[Y^2 - 2\eta Y + \eta^2] \\ = & \mathbb{E}[Y^2] - 2\eta\mathbb{E}[Y] + \eta^2. \end{aligned}$$

$$\begin{aligned} & \frac{\partial}{\partial \eta} \mathbb{E}_Y[(Y - \eta)^2] \\ = & 0 - 2\mathbb{E}[Y] + 2\eta \\ = & 2(\eta - \mathbb{E}[Y]). \end{aligned}$$

Set $\frac{\partial}{\partial \eta} \mathbb{E}_Y[(Y - \eta)^2]|_{\eta=\hat{\eta}} = 0$. We obtain

$$\hat{\eta} = \mathbb{E}[Y].$$

With $\frac{\partial^2}{\partial \eta^2} \mathbb{E}_Y[(Y - \eta)^2] = 2 \geq 0$, it is proved that $\arg \min_{\eta} \mathbb{E}_Y[(Y - \eta)^2] = \mathbb{E}[Y]$.

If Y is continuous with the cumulative distribution function (CDF) denoted as $F_Y()$ and an arbitrary quantile level τ ($0 < \tau < 1$), denote the

probability density function (PDF) of Y as $f_Y()$ and the support of Y as (a, b) . Then

$$\begin{aligned}
 & \mathbb{E}_Y[\rho_\tau(Y, \eta)] \\
 = & \int_a^b (y - \eta)(\tau - \mathbf{1}_{[y < \eta]}) f_Y(y) dy \\
 = & \tau \int_a^b (y - \eta) f_Y(y) dy - \int_a^\eta (y - \eta) f_Y(y) dy \\
 = & \tau \int_a^b y f_Y(y) dy - \tau \eta \int_a^b f_Y(y) dy - \int_a^\eta y f_Y(y) dy + \eta \int_a^\eta f_Y(y) dy \\
 = & \tau \mathbb{E}[Y] - \tau \eta - \int_a^\eta y f_Y(y) dy + \eta F_Y(\eta).
 \end{aligned}$$

$$\begin{aligned}
 & \frac{\partial}{\partial \eta} \mathbb{E}_Y[\rho_\tau(Y, \eta)] \\
 = & 0 - \tau - \eta f_Y(\eta) + [F_Y(\eta) + \eta f_Y(\eta)] \\
 = & F_Y(\eta) - \tau.
 \end{aligned}$$

Set $\frac{\partial}{\partial \eta} \mathbb{E}_Y[\rho_\tau(Y, \eta)]|_{\eta=\hat{\eta}} = 0$. We obtain

$$\hat{\eta} = F_Y^{-1}(\tau).$$

With $\frac{\partial^2}{\partial \eta^2} \mathbb{E}_Y[\rho_\tau(Y, \eta)] = f_Y(\eta) \geq 0$, it is proved that $\arg \min_\eta \mathbb{E}_Y[\rho_\tau(Y, \eta)] = F_Y^{-1}(\tau)$.

A.2 Supplementary Numerical Results of Chapter 2

A.2.1 Estimation Accuracy

Tables A.1 ~ A.14 present $\sqrt{\widehat{ISE}}$ of all the 7 nonzero functions ($f_1, \dots, f_7$) and the first zero function (f_8) for all different combinations of $N = 300$ or 500, $p = 15$ or 30, and normally or Student's t distributed noise terms except for the scenario displayed in the main text where $N = 300$, $p = 15$, and the noises follow the Gaussian distribution.

A.2.2 Prediction Accuracy

Test errors for different scenarios with Student's t distributed noises are displayed in Tables A.15 and A.16. Those for cases with Gaussian errors are displayed in the main text.

A.2.3 Variable Selection

Tables A.17 ~ A.23 present numerical results in terms of variable selection for all different combinations of $N = 300$ or 500, $p = 15$ or 30, and normally or Student's t distributed noise terms except for the scenario displayed in the main text where $N = 300$, $p = 15$, and the noises follow the Gaussian distribution. In all these tables, "NV", "NVT", "NL", and "NLT" indicate the number of nonzero covariates selected, the number of nonzero covariates selected which are truly informative, the number of linear covariates selected, and the number of linear covariates selected which are truly linear, respectively.

Table A.1: Average (standard deviation) of $\sqrt{ISE}$ over 100 replicates by different approaches, for $N = 300$, $p = 30$, and Gaussian errors.

		f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
Mean	PL-Tw	0.100	0.088	0.056	0.075	0.039	0.035	0.036	0.003
		(0.023)	(0.025)	(0.022)	(0.023)	(0.029)	(0.027)	(0.027)	(0.010)
	PL-St	0.171	0.118	0.068	0.113	0.091	0.082	0.065	0.026
		(0.020)	(0.030)	(0.027)	(0.028)	(0.045)	(0.038)	(0.034)	(0.024)
	NL-Tw	0.104	0.084	0.066	0.077	0.067	0.066	0.054	0.002
		(0.024)	(0.022)	(0.024)	(0.025)	(0.029)	(0.021)	(0.022)	(0.007)
	PL-mb	0.180	0.116	0.059	0.125	0.077	0.069	0.058	0.027
		(0.017)	(0.022)	(0.021)	(0.027)	(0.042)	(0.034)	(0.034)	(0.023)
$\tau = 0.1$	mgcv	0.103	0.090	0.062	0.084	0.059	0.045	0.048	0.030
		(0.021)	(0.022)	(0.018)	(0.031)	(0.035)	(0.028)	(0.040)	(0.042)
	mcmc	0.108	0.104	0.083	0.111	0.066	0.052	0.055	0.003
		(0.019)	(0.022)	(0.018)	(0.021)	(0.030)	(0.027)	(0.026)	(0.010)
	PL-Tw	0.190	0.191	0.089	0.186	0.084	0.097	0.126	0.001
		(0.038)	(0.068)	(0.039)	(0.069)	(0.073)	(0.078)	(0.088)	(0.007)
	PL-St	0.261	0.208	0.116	0.313	0.206	0.186	0.137	0.039
		(0.047)	(0.059)	(0.045)	(0.081)	(0.084)	(0.079)	(0.066)	(0.032)
$\tau = 0.3$	NL-Tw	0.155	0.134	0.105	0.138	0.110	0.124	0.143	0.001
		(0.033)	(0.036)	(0.041)	(0.045)	(0.058)	(0.056)	(0.075)	(0.004)
	PL-mb	0.306	0.247	0.125	0.364	0.195	0.180	0.129	0.041
		(0.055)	(0.047)	(0.058)	(0.082)	(0.090)	(0.081)	(0.064)	(0.032)
	mcmc	0.167	0.190	0.134	0.242	0.092	0.084	0.084	0.018
		(0.039)	(0.094)	(0.059)	(0.115)	(0.063)	(0.053)	(0.067)	(0.038)
	PL-Tw	0.147	0.119	0.060	0.094	0.046	0.050	0.052	0.000
		(0.026)	(0.034)	(0.025)	(0.026)	(0.036)	(0.036)	(0.041)	(0.000)
	PL-St	0.195	0.139	0.075	0.146	0.105	0.095	0.074	0.035
		(0.024)	(0.035)	(0.029)	(0.039)	(0.052)	(0.045)	(0.043)	(0.023)
	NL-Tw	0.122	0.092	0.071	0.093	0.072	0.074	0.067	0.000
		(0.026)	(0.028)	(0.028)	(0.027)	(0.031)	(0.028)	(0.037)	(0.001)
	PL-mb	0.217	0.160	0.089	0.184	0.103	0.091	0.071	0.033
		(0.028)	(0.030)	(0.036)	(0.037)	(0.056)	(0.049)	(0.045)	(0.024)
	mcmc	0.118	0.109	0.090	0.113	0.059	0.062	0.056	0.006
		(0.029)	(0.023)	(0.022)	(0.035)	(0.040)	(0.037)	(0.033)	(0.019)

Table A.2: Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates by different approaches, for $N = 300$, $p = 30$, and Gaussian errors. (Continued)

		f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
$\tau = 0.5$	PL-Tw	0.141	0.113	0.056	0.086	0.045	0.044	0.052	0.001
		(0.026)	(0.037)	(0.024)	(0.025)	(0.034)	(0.035)	(0.043)	(0.005)
	PL-St	0.189	0.135	0.070	0.116	0.098	0.086	0.069	0.030
		(0.020)	(0.030)	(0.032)	(0.032)	(0.049)	(0.042)	(0.045)	(0.024)
	NL-Tw	0.114	0.091	0.068	0.092	0.073	0.075	0.067	0.001
		(0.027)	(0.026)	(0.028)	(0.028)	(0.031)	(0.026)	(0.036)	(0.010)
	PL-mb	0.209	0.156	0.081	0.149	0.088	0.075	0.061	0.034
		(0.023)	(0.029)	(0.032)	(0.025)	(0.051)	(0.039)	(0.039)	(0.023)
	mcmc	0.117	0.102	0.093	0.110	0.061	0.054	0.054	0.014
		(0.031)	(0.025)	(0.027)	(0.032)	(0.041)	(0.028)	(0.031)	(0.032)
$\tau = 0.7$	PL-Tw	0.144	0.126	0.058	0.089	0.049	0.047	0.051	0.000
		(0.024)	(0.039)	(0.029)	(0.029)	(0.037)	(0.036)	(0.040)	(0.003)
	PL-St	0.193	0.144	0.074	0.114	0.109	0.094	0.077	0.034
		(0.022)	(0.035)	(0.033)	(0.036)	(0.057)	(0.049)	(0.042)	(0.024)
	NL-Tw	0.119	0.097	0.070	0.098	0.073	0.079	0.067	0.000
		(0.027)	(0.026)	(0.032)	(0.031)	(0.034)	(0.030)	(0.039)	(0.002)
	PL-mb	0.222	0.170	0.088	0.140	0.098	0.086	0.070	0.031
		(0.028)	(0.031)	(0.034)	(0.021)	(0.054)	(0.044)	(0.044)	(0.022)
	mcmc	0.117	0.099	0.089	0.107	0.063	0.049	0.052	0.017
		(0.021)	(0.027)	(0.026)	(0.024)	(0.037)	(0.026)	(0.036)	(0.033)
$\tau = 0.9$	PL-Tw	0.192	0.232	0.094	0.115	0.084	0.097	0.103	0.005
		(0.039)	(0.087)	(0.050)	(0.049)	(0.059)	(0.075)	(0.087)	(0.017)
	PL-St	0.269	0.233	0.122	0.142	0.194	0.180	0.122	0.047
		(0.044)	(0.059)	(0.049)	(0.051)	(0.078)	(0.073)	(0.060)	(0.039)
	NL-Tw	0.156	0.151	0.120	0.118	0.101	0.124	0.122	0.002
		(0.034)	(0.054)	(0.055)	(0.039)	(0.049)	(0.058)	(0.079)	(0.008)
	PL-mb	0.311	0.278	0.163	0.169	0.186	0.167	0.122	0.033
		(0.051)	(0.048)	(0.060)	(0.038)	(0.076)	(0.077)	(0.071)	(0.031)
	mcmc	0.149	0.187	0.184	0.164	0.095	0.081	0.090	0.005
		(0.039)	(0.084)	(0.095)	(0.086)	(0.100)	(0.102)	(0.082)	(0.024)

Table A.3: Average (standard deviation) of $\sqrt{ISE}$ over 100 replicates by different approaches, for $N = 500$, $p = 15$, and Gaussian errors.

		f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
Mean	PL-Tw	0.075	0.061	0.043	0.061	0.026	0.025	0.028	0.003
		(0.021)	(0.016)	(0.015)	(0.017)	(0.020)	(0.021)	(0.020)	(0.008)
	PL-St	0.107	0.067	0.050	0.070	0.053	0.049	0.049	0.034
		(0.025)	(0.017)	(0.015)	(0.020)	(0.022)	(0.022)	(0.022)	(0.019)
	NL-Tw	0.078	0.061	0.052	0.065	0.049	0.049	0.044	0.002
		(0.020)	(0.015)	(0.016)	(0.018)	(0.021)	(0.019)	(0.020)	(0.008)
	PL-mb	0.142	0.075	0.043	0.079	0.043	0.044	0.040	0.033
		(0.014)	(0.020)	(0.016)	(0.018)	(0.023)	(0.021)	(0.021)	(0.019)
$\tau = 0.1$	mgcv	0.072	0.072	0.056	0.064	0.032	0.031	0.029	0.028
		(0.013)	(0.020)	(0.018)	(0.018)	(0.019)	(0.023)	(0.019)	(0.029)
	mcmc	0.086	0.083	0.077	0.092	0.050	0.047	0.044	0.004
		(0.013)	(0.021)	(0.021)	(0.016)	(0.019)	(0.021)	(0.018)	(0.014)
	PL-Tw	0.153	0.116	0.059	0.111	0.045	0.055	0.061	0.000
		(0.024)	(0.038)	(0.025)	(0.034)	(0.037)	(0.046)	(0.052)	(0.001)
	PL-St	0.179	0.111	0.073	0.144	0.087	0.079	0.069	0.043
		(0.018)	(0.031)	(0.032)	(0.036)	(0.042)	(0.053)	(0.041)	(0.026)
$\tau = 0.3$	NL-Tw	0.135	0.097	0.077	0.101	0.074	0.080	0.075	0.001
		(0.027)	(0.027)	(0.032)	(0.029)	(0.034)	(0.038)	(0.041)	(0.004)
	PL-mb	0.205	0.154	0.076	0.186	0.080	0.067	0.066	0.046
		(0.028)	(0.033)	(0.032)	(0.040)	(0.042)	(0.039)	(0.036)	(0.027)
	mcmc	0.108	0.100	0.099	0.116	0.046	0.045	0.044	0.014
		(0.026)	(0.026)	(0.026)	(0.026)	(0.024)	(0.021)	(0.027)	(0.032)
	PL-Tw	0.111	0.084	0.047	0.076	0.029	0.030	0.038	0.001
		(0.024)	(0.022)	(0.016)	(0.022)	(0.023)	(0.025)	(0.026)	(0.005)
	PL-St	0.152	0.085	0.054	0.092	0.060	0.056	0.053	0.036
		(0.017)	(0.022)	(0.018)	(0.026)	(0.031)	(0.031)	(0.028)	(0.025)
	NL-Tw	0.097	0.075	0.057	0.082	0.056	0.056	0.049	0.002
		(0.023)	(0.018)	(0.021)	(0.025)	(0.024)	(0.021)	(0.024)	(0.008)
	PL-mb	0.176	0.105	0.058	0.105	0.053	0.048	0.045	0.032
		(0.013)	(0.019)	(0.020)	(0.025)	(0.030)	(0.028)	(0.027)	(0.022)
	mcmc	0.088	0.084	0.083	0.095	0.045	0.035	0.031	0.006
		(0.017)	(0.019)	(0.022)	(0.021)	(0.024)	(0.017)	(0.019)	(0.019)

Table A.4: Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates by different approaches, for $N = 500$, $p = 15$, and Gaussian errors. (Continued)

		f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
$\tau = 0.5$	PL-Tw	0.104	0.080	0.046	0.069	0.028	0.027	0.032	0.001
		(0.022)	(0.020)	(0.018)	(0.017)	(0.020)	(0.021)	(0.023)	(0.005)
	PL-St	0.149	0.083	0.056	0.082	0.058	0.051	0.049	0.032
		(0.019)	(0.020)	(0.020)	(0.023)	(0.025)	(0.023)	(0.023)	(0.022)
	NL-Tw	0.091	0.070	0.057	0.075	0.053	0.052	0.046	0.001
		(0.022)	(0.017)	(0.022)	(0.021)	(0.020)	(0.018)	(0.020)	(0.004)
	PL-mb	0.173	0.102	0.052	0.096	0.048	0.047	0.038	0.035
		(0.011)	(0.018)	(0.021)	(0.022)	(0.026)	(0.026)	(0.021)	(0.019)
	mcmc	0.087	0.085	0.076	0.093	0.039	0.037	0.036	0.012
		(0.015)	(0.023)	(0.020)	(0.018)	(0.020)	(0.022)	(0.020)	(0.027)
$\tau = 0.7$	PL-Tw	0.111	0.085	0.047	0.068	0.031	0.031	0.036	0.000
		(0.022)	(0.019)	(0.018)	(0.021)	(0.022)	(0.025)	(0.022)	(0.002)
	PL-St	0.155	0.089	0.058	0.076	0.060	0.055	0.053	0.035
		(0.015)	(0.023)	(0.023)	(0.020)	(0.026)	(0.029)	(0.023)	(0.020)
	NL-Tw	0.096	0.073	0.060	0.079	0.058	0.058	0.049	0.000
		(0.021)	(0.019)	(0.023)	(0.024)	(0.023)	(0.023)	(0.020)	(0.003)
	PL-mb	0.180	0.110	0.051	0.095	0.049	0.045	0.042	0.032
		(0.011)	(0.021)	(0.021)	(0.023)	(0.027)	(0.025)	(0.026)	(0.020)
	mcmc	0.090	0.087	0.082	0.093	0.040	0.042	0.038	0.004
		(0.018)	(0.019)	(0.025)	(0.014)	(0.022)	(0.024)	(0.016)	(0.016)
$\tau = 0.9$	PL-Tw	0.153	0.134	0.064	0.086	0.040	0.047	0.058	0.001
		(0.023)	(0.040)	(0.028)	(0.028)	(0.030)	(0.039)	(0.038)	(0.007)
	PL-St	0.187	0.128	0.072	0.089	0.081	0.078	0.073	0.046
		(0.023)	(0.034)	(0.026)	(0.027)	(0.041)	(0.041)	(0.035)	(0.026)
	NL-Tw	0.128	0.103	0.081	0.098	0.070	0.077	0.072	0.001
		(0.025)	(0.029)	(0.031)	(0.032)	(0.029)	(0.028)	(0.034)	(0.007)
	PL-mb	0.213	0.151	0.084	0.118	0.090	0.069	0.059	0.041
		(0.026)	(0.035)	(0.036)	(0.031)	(0.045)	(0.041)	(0.039)	(0.029)
	mcmc	0.108	0.104	0.097	0.117	0.050	0.051	0.049	0.012
		(0.021)	(0.021)	(0.027)	(0.023)	(0.028)	(0.034)	(0.029)	(0.029)

Table A.5: Average (standard deviation) of $\sqrt{ISE}$ over 100 replicates by different approaches, for $N = 500$, $p = 30$, and Gaussian errors.

		f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
Mean	PL-Tw	0.071	0.064	0.044	0.057	0.026	0.028	0.027	0.002
		(0.020)	(0.016)	(0.018)	(0.018)	(0.020)	(0.019)	(0.021)	(0.007)
	PL-St	0.138	0.086	0.051	0.082	0.057	0.056	0.050	0.026
		(0.021)	(0.021)	(0.020)	(0.020)	(0.028)	(0.027)	(0.023)	(0.019)
	NL-Tw	0.074	0.065	0.052	0.059	0.051	0.051	0.044	0.002
		(0.019)	(0.015)	(0.017)	(0.017)	(0.019)	(0.019)	(0.019)	(0.007)
	PL-mb	0.162	0.097	0.048	0.096	0.053	0.050	0.042	0.023
		(0.014)	(0.021)	(0.017)	(0.022)	(0.028)	(0.026)	(0.024)	(0.019)
	mgcv	0.075	0.072	0.057	0.068	0.035	0.031	0.035	0.022
		(0.019)	(0.021)	(0.021)	(0.020)	(0.028)	(0.026)	(0.025)	(0.026)
	mcmc	0.084	0.081	0.077	0.092	0.054	0.045	0.044	0.003
		(0.019)	(0.019)	(0.016)	(0.015)	(0.025)	(0.021)	(0.021)	(0.013)
$\tau = 0.1$	PL-Tw	0.152	0.126	0.061	0.114	0.051	0.059	0.069	0.001
		(0.026)	(0.037)	(0.025)	(0.033)	(0.035)	(0.044)	(0.051)	(0.005)
	PL-St	0.197	0.140	0.075	0.195	0.118	0.108	0.088	0.040
		(0.027)	(0.034)	(0.029)	(0.046)	(0.056)	(0.049)	(0.045)	(0.027)
	NL-Tw	0.129	0.098	0.074	0.100	0.077	0.083	0.082	0.001
		(0.026)	(0.025)	(0.032)	(0.028)	(0.033)	(0.034)	(0.043)	(0.005)
	PL-mb	0.228	0.183	0.087	0.246	0.111	0.107	0.079	0.033
		(0.031)	(0.038)	(0.035)	(0.052)	(0.054)	(0.053)	(0.040)	(0.026)
	mcmc	0.110	0.101	0.094	0.119	0.044	0.052	0.043	0.008
		(0.023)	(0.022)	(0.024)	(0.029)	(0.027)	(0.032)	(0.026)	(0.022)
$\tau = 0.3$	PL-Tw	0.111	0.089	0.047	0.076	0.029	0.035	0.042	0.001
		(0.020)	(0.023)	(0.018)	(0.019)	(0.023)	(0.024)	(0.033)	(0.004)
	PL-St	0.166	0.106	0.055	0.108	0.072	0.069	0.061	0.029
		(0.012)	(0.024)	(0.021)	(0.027)	(0.032)	(0.033)	(0.034)	(0.022)
	NL-Tw	0.095	0.078	0.057	0.076	0.055	0.058	0.053	0.000
		(0.020)	(0.020)	(0.021)	(0.022)	(0.021)	(0.023)	(0.027)	(0.002)
	PL-mb	0.185	0.127	0.066	0.136	0.068	0.061	0.048	0.026
		(0.013)	(0.023)	(0.023)	(0.028)	(0.033)	(0.034)	(0.033)	(0.020)
	mcmc	0.090	0.087	0.079	0.092	0.046	0.049	0.048	0.014
		(0.034)	(0.024)	(0.019)	(0.017)	(0.033)	(0.025)	(0.024)	(0.025)

Table A.6: Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates by different approaches, for $N = 500$, $p = 30$, and Gaussian errors. (Continued)

		f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
$\tau = 0.5$	PL-Tw	0.103	0.080	0.046	0.067	0.026	0.026	0.035	0.001
		(0.020)	(0.018)	(0.018)	(0.020)	(0.020)	(0.022)	(0.028)	(0.006)
	PL-St	0.162	0.101	0.054	0.089	0.065	0.062	0.056	0.028
		(0.012)	(0.022)	(0.018)	(0.024)	(0.030)	(0.028)	(0.023)	(0.023)
	NL-Tw	0.088	0.073	0.053	0.072	0.053	0.052	0.050	0.000
		(0.018)	(0.017)	(0.020)	(0.020)	(0.018)	(0.022)	(0.023)	(0.001)
	PL-mb	0.182	0.125	0.057	0.111	0.063	0.052	0.041	0.028
		(0.012)	(0.019)	(0.021)	(0.023)	(0.031)	(0.031)	(0.025)	(0.021)
	mcmc	0.090	0.090	0.080	0.089	0.043	0.046	0.037	0.005
		(0.020)	(0.024)	(0.019)	(0.016)	(0.030)	(0.023)	(0.020)	(0.016)
$\tau = 0.7$	PL-Tw	0.110	0.086	0.048	0.070	0.032	0.032	0.038	0.000
		(0.020)	(0.022)	(0.017)	(0.019)	(0.024)	(0.026)	(0.029)	(0.003)
	PL-St	0.168	0.106	0.056	0.085	0.069	0.069	0.058	0.029
		(0.014)	(0.023)	(0.023)	(0.023)	(0.033)	(0.032)	(0.028)	(0.021)
	NL-Tw	0.093	0.074	0.058	0.075	0.058	0.058	0.051	0.000
		(0.020)	(0.017)	(0.022)	(0.023)	(0.021)	(0.024)	(0.023)	(0.001)
	PL-mb	0.190	0.135	0.064	0.111	0.067	0.058	0.046	0.025
		(0.015)	(0.022)	(0.025)	(0.022)	(0.033)	(0.033)	(0.028)	(0.020)
	mcmc	0.094	0.087	0.079	0.104	0.045	0.039	0.041	0.009
		(0.018)	(0.019)	(0.017)	(0.020)	(0.023)	(0.021)	(0.024)	(0.026)
$\tau = 0.9$	PL-Tw	0.147	0.136	0.061	0.086	0.039	0.053	0.059	0.000
		(0.021)	(0.041)	(0.026)	(0.024)	(0.031)	(0.038)	(0.044)	(0.001)
	PL-St	0.207	0.160	0.077	0.097	0.112	0.112	0.087	0.037
		(0.026)	(0.038)	(0.032)	(0.028)	(0.044)	(0.052)	(0.043)	(0.027)
	NL-Tw	0.123	0.104	0.078	0.090	0.071	0.085	0.082	0.000
		(0.023)	(0.028)	(0.035)	(0.028)	(0.027)	(0.030)	(0.041)	(0.004)
	PL-mb	0.239	0.193	0.111	0.132	0.112	0.100	0.073	0.029
		(0.030)	(0.040)	(0.042)	(0.020)	(0.049)	(0.056)	(0.040)	(0.023)
	mcmc	0.122	0.101	0.098	0.125	0.042	0.043	0.043	0.022
		(0.022)	(0.037)	(0.021)	(0.031)	(0.030)	(0.023)	(0.023)	(0.039)

Table A.7: Average (standard deviation) of $\sqrt{ISE}$ over 100 replicates by different approaches, for $N = 300$, $p = 15$, and Student's t errors.

		f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
Mean	PL-Tw	0.113	0.097	0.064	0.086	0.047	0.051	0.048	0.007
		(0.058)	(0.058)	(0.043)	(0.038)	(0.040)	(0.049)	(0.047)	(0.016)
	PL-St	0.154	0.108	0.069	0.108	0.087	0.087	0.070	0.043
		(0.055)	(0.050)	(0.047)	(0.055)	(0.046)	(0.053)	(0.049)	(0.029)
	NL-Tw	0.115	0.092	0.075	0.088	0.077	0.083	0.067	0.006
		(0.050)	(0.040)	(0.048)	(0.036)	(0.044)	(0.050)	(0.047)	(0.015)
	PL-mb	0.166	0.101	0.056	0.110	0.068	0.064	0.060	0.037
		(0.027)	(0.035)	(0.023)	(0.044)	(0.038)	(0.044)	(0.040)	(0.024)
$\tau = 0.1$	mgcv	0.114	0.086	0.076	0.098	0.047	0.047	0.053	0.029
		(0.040)	(0.035)	(0.042)	(0.045)	(0.037)	(0.036)	(0.048)	(0.042)
	mcmc	0.123	0.102	0.092	0.120	0.063	0.067	0.056	0.004
		(0.043)	(0.039)	(0.030)	(0.035)	(0.037)	(0.040)	(0.048)	(0.008)
	PL-Tw	0.175	0.146	0.071	0.142	0.047	0.067	0.078	0.002
		(0.038)	(0.059)	(0.042)	(0.055)	(0.040)	(0.048)	(0.058)	(0.012)
	PL-St	0.205	0.133	0.082	0.190	0.115	0.102	0.088	0.051
		(0.030)	(0.040)	(0.034)	(0.060)	(0.054)	(0.047)	(0.044)	(0.035)
$\tau = 0.3$	NL-Tw	0.145	0.111	0.080	0.112	0.082	0.093	0.095	0.000
		(0.035)	(0.034)	(0.041)	(0.038)	(0.037)	(0.040)	(0.058)	(0.002)
	PL-mb	0.249	0.195	0.095	0.266	0.119	0.111	0.087	0.054
		(0.052)	(0.045)	(0.044)	(0.078)	(0.064)	(0.052)	(0.048)	(0.037)
	mcmc	0.150	0.158	0.137	0.212	0.063	0.068	0.044	0.015
		(0.060)	(0.080)	(0.079)	(0.116)	(0.039)	(0.039)	(0.035)	(0.030)
	PL-Tw	0.106	0.081	0.042	0.074	0.023	0.028	0.030	0.000
		(0.024)	(0.022)	(0.016)	(0.018)	(0.018)	(0.022)	(0.021)	(0.000)
	PL-St	0.151	0.082	0.050	0.092	0.051	0.046	0.045	0.031
		(0.012)	(0.022)	(0.018)	(0.026)	(0.025)	(0.024)	(0.021)	(0.017)
	NL-Tw	0.074	0.064	0.046	0.060	0.042	0.043	0.037	0.000
		(0.020)	(0.015)	(0.016)	(0.018)	(0.016)	(0.014)	(0.016)	(0.001)
	PL-mb	0.176	0.113	0.057	0.126	0.054	0.047	0.043	0.032
		(0.012)	(0.022)	(0.018)	(0.029)	(0.029)	(0.028)	(0.024)	(0.020)
	mcmc	0.078	0.070	0.067	0.089	0.033	0.040	0.032	0.002
		(0.018)	(0.016)	(0.011)	(0.028)	(0.022)	(0.017)	(0.019)	(0.009)

Table A.8: Average (standard deviation) of $\sqrt{\widehat{ISE}}$ over 100 replicates by different approaches, for $N = 300$, $p = 15$, and Student's t errors. (Continued)

		f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
$\tau = 0.5$	PL-Tw	0.094	0.073	0.037	0.061	0.023	0.024	0.025	0.000
		(0.019)	(0.019)	(0.014)	(0.013)	(0.016)	(0.018)	(0.020)	(0.000)
	PL-St	0.142	0.078	0.048	0.074	0.047	0.042	0.038	0.031
		(0.014)	(0.018)	(0.016)	(0.017)	(0.022)	(0.023)	(0.023)	(0.018)
	NI-Tw	0.064	0.057	0.042	0.055	0.037	0.039	0.032	0.000
		(0.015)	(0.015)	(0.014)	(0.013)	(0.013)	(0.013)	(0.014)	(0.000)
	PL-mb	0.172	0.108	0.051	0.100	0.047	0.042	0.040	0.031
		(0.011)	(0.017)	(0.018)	(0.020)	(0.024)	(0.026)	(0.024)	(0.019)
	mcmc	0.075	0.065	0.065	0.084	0.040	0.041	0.032	0.004
		(0.027)	(0.018)	(0.018)	(0.026)	(0.026)	(0.021)	(0.021)	(0.012)
$\tau = 0.7$	PL-Tw	0.108	0.087	0.041	0.066	0.022	0.029	0.033	0.000
		(0.020)	(0.027)	(0.016)	(0.017)	(0.018)	(0.023)	(0.029)	(0.001)
	PL-St	0.155	0.088	0.051	0.074	0.053	0.050	0.044	0.033
		(0.013)	(0.024)	(0.016)	(0.018)	(0.029)	(0.027)	(0.028)	(0.016)
	NI-Tw	0.074	0.068	0.047	0.063	0.044	0.045	0.038	0.000
		(0.019)	(0.020)	(0.017)	(0.019)	(0.016)	(0.015)	(0.019)	(0.002)
	PL-mb	0.184	0.116	0.057	0.098	0.054	0.047	0.042	0.028
		(0.014)	(0.021)	(0.023)	(0.019)	(0.029)	(0.029)	(0.026)	(0.020)
	mcmc	0.077	0.065	0.069	0.085	0.040	0.038	0.033	0.003
		(0.017)	(0.013)	(0.016)	(0.018)	(0.019)	(0.021)	(0.023)	(0.012)
$\tau = 0.9$	PL-Tw	0.180	0.201	0.078	0.105	0.054	0.078	0.096	0.004
		(0.031)	(0.071)	(0.041)	(0.036)	(0.047)	(0.070)	(0.073)	(0.016)
	PL-St	0.223	0.172	0.088	0.119	0.124	0.121	0.098	0.056
		(0.037)	(0.047)	(0.039)	(0.042)	(0.062)	(0.069)	(0.051)	(0.033)
	NI-Tw	0.143	0.137	0.095	0.107	0.089	0.102	0.103	0.002
		(0.032)	(0.046)	(0.044)	(0.034)	(0.041)	(0.051)	(0.061)	(0.008)
	PL-mb	0.254	0.211	0.115	0.149	0.122	0.098	0.081	0.049
		(0.041)	(0.049)	(0.052)	(0.043)	(0.057)	(0.060)	(0.057)	(0.031)
	mcmc	0.147	0.186	0.156	0.156	0.060	0.063	0.070	0.015
		(0.073)	(0.127)	(0.102)	(0.067)	(0.041)	(0.053)	(0.064)	(0.033)

Table A.9: Average (standard deviation) of $\sqrt{ISE}$ over 100 replicates by different approaches, for $N = 300$, $p = 30$, and Student's t errors.

		f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
Mean	PL-Tw	0.103	0.092	0.060	0.083	0.036	0.042	0.046	0.004
		(0.046)	(0.058)	(0.040)	(0.041)	(0.039)	(0.044)	(0.046)	(0.013)
	PL-St	0.169	0.120	0.070	0.121	0.087	0.088	0.072	0.029
		(0.038)	(0.052)	(0.033)	(0.045)	(0.056)	(0.051)	(0.045)	(0.025)
	NL-Tw	0.103	0.086	0.069	0.082	0.064	0.069	0.062	0.003
		(0.034)	(0.043)	(0.030)	(0.032)	(0.030)	(0.036)	(0.038)	(0.011)
	PL-mb	0.184	0.125	0.071	0.139	0.090	0.076	0.060	0.029
		(0.035)	(0.043)	(0.039)	(0.047)	(0.055)	(0.048)	(0.042)	(0.031)
$\tau = 0.1$	mgcv	0.105	0.104	0.089	0.094	0.056	0.049	0.046	0.029
		(0.063)	(0.048)	(0.070)	(0.040)	(0.047)	(0.033)	(0.044)	(0.045)
	mcmc	0.117	0.115	0.104	0.120	0.074	0.067	0.047	0.002
		(0.073)	(0.064)	(0.062)	(0.043)	(0.041)	(0.040)	(0.028)	(0.006)
	PL-Tw	0.172	0.166	0.078	0.180	0.060	0.082	0.106	0.004
		(0.038)	(0.068)	(0.039)	(0.074)	(0.049)	(0.068)	(0.087)	(0.018)
	PL-St	0.248	0.191	0.114	0.299	0.177	0.166	0.131	0.046
		(0.049)	(0.056)	(0.053)	(0.079)	(0.070)	(0.075)	(0.062)	(0.034)
$\tau = 0.3$	NL-Tw	0.138	0.114	0.090	0.118	0.078	0.096	0.114	0.001
		(0.033)	(0.036)	(0.042)	(0.044)	(0.036)	(0.036)	(0.069)	(0.007)
	PL-mb	0.288	0.233	0.124	0.361	0.176	0.154	0.115	0.043
		(0.057)	(0.044)	(0.058)	(0.078)	(0.074)	(0.069)	(0.063)	(0.038)
	mcmc	0.161	0.187	0.180	0.289	0.065	0.073	0.053	0.018
		(0.104)	(0.099)	(0.118)	(0.127)	(0.040)	(0.044)	(0.034)	(0.043)
	PL-Tw	0.106	0.081	0.044	0.078	0.024	0.027	0.033	0.000
		(0.023)	(0.020)	(0.018)	(0.020)	(0.018)	(0.022)	(0.024)	(0.001)
	PL-St	0.166	0.101	0.062	0.118	0.065	0.064	0.052	0.025
		(0.012)	(0.021)	(0.022)	(0.028)	(0.029)	(0.031)	(0.030)	(0.016)
	NL-Tw	0.073	0.063	0.046	0.059	0.042	0.046	0.038	0.000
		(0.018)	(0.015)	(0.015)	(0.018)	(0.016)	(0.015)	(0.018)	(0.000)
	PL-mb	0.190	0.128	0.072	0.159	0.071	0.060	0.047	0.024
		(0.015)	(0.023)	(0.022)	(0.029)	(0.033)	(0.032)	(0.030)	(0.021)
	mcmc	0.072	0.068	0.066	0.083	0.035	0.035	0.024	0.004
		(0.017)	(0.016)	(0.017)	(0.017)	(0.020)	(0.020)	(0.012)	(0.015)

Table A.10: Average (standard deviation) of $\sqrt{ISE}$ over 100 replicates by different approaches, for $N = 300$, $p = 30$, and Student's t errors. (Continued)

		f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
$\tau = 0.5$	PL-Tw	0.095	0.073	0.042	0.064	0.018	0.022	0.029	0.000
		(0.019)	(0.016)	(0.015)	(0.016)	(0.014)	(0.021)	(0.021)	(0.000)
	PL-St	0.160	0.096	0.058	0.093	0.059	0.054	0.044	0.022
		(0.012)	(0.019)	(0.018)	(0.021)	(0.027)	(0.028)	(0.027)	(0.015)
	NL-Tw	0.062	0.055	0.043	0.054	0.037	0.040	0.031	0.000
		(0.016)	(0.012)	(0.013)	(0.017)	(0.014)	(0.014)	(0.014)	(0.000)
	PL-mb	0.184	0.126	0.065	0.130	0.062	0.050	0.041	0.023
		(0.013)	(0.022)	(0.020)	(0.024)	(0.035)	(0.029)	(0.023)	(0.018)
	mcmc	0.068	0.062	0.062	0.076	0.033	0.029	0.031	0.005
		(0.016)	(0.018)	(0.015)	(0.017)	(0.018)	(0.012)	(0.016)	(0.015)
$\tau = 0.7$	PL-Tw	0.108	0.090	0.046	0.065	0.026	0.030	0.033	0.000
		(0.021)	(0.026)	(0.019)	(0.016)	(0.018)	(0.025)	(0.026)	(0.000)
	PL-St	0.173	0.109	0.064	0.089	0.066	0.066	0.053	0.025
		(0.016)	(0.026)	(0.022)	(0.019)	(0.035)	(0.032)	(0.032)	(0.018)
	NL-Tw	0.073	0.065	0.051	0.060	0.044	0.045	0.037	0.000
		(0.020)	(0.017)	(0.017)	(0.018)	(0.016)	(0.015)	(0.020)	(0.000)
	PL-mb	0.199	0.139	0.073	0.123	0.074	0.058	0.048	0.024
		(0.016)	(0.024)	(0.029)	(0.023)	(0.041)	(0.033)	(0.029)	(0.018)
	mcmc	0.076	0.067	0.068	0.081	0.038	0.040	0.032	0.005
		(0.027)	(0.017)	(0.020)	(0.019)	(0.034)	(0.027)	(0.016)	(0.015)
$\tau = 0.9$	PL-Tw	0.189	0.235	0.084	0.115	0.060	0.086	0.107	0.001
		(0.042)	(0.088)	(0.044)	(0.043)	(0.045)	(0.068)	(0.087)	(0.006)
	PL-St	0.264	0.228	0.108	0.136	0.163	0.166	0.117	0.043
		(0.049)	(0.062)	(0.050)	(0.048)	(0.068)	(0.078)	(0.064)	(0.031)
	NL-Tw	0.143	0.142	0.103	0.113	0.078	0.107	0.108	0.001
		(0.034)	(0.059)	(0.044)	(0.042)	(0.036)	(0.060)	(0.068)	(0.011)
	PL-mb	0.297	0.277	0.150	0.176	0.178	0.153	0.115	0.042
		(0.052)	(0.054)	(0.064)	(0.050)	(0.087)	(0.073)	(0.063)	(0.035)
	mcmc	0.150	0.219	0.198	0.158	0.070	0.074	0.098	0.021
		(0.050)	(0.108)	(0.116)	(0.084)	(0.041)	(0.056)	(0.081)	(0.047)

Table A.11: Average (standard deviation) of $\sqrt{ISE}$ over 100 replicates by different approaches, for $N = 500$, $p = 15$, and Student's t errors.

		f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
Mean	PL-Tw	0.095	0.081	0.048	0.075	0.043	0.047	0.037	0.004
		(0.070)	(0.063)	(0.030)	(0.068)	(0.069)	(0.066)	(0.035)	(0.017)
	PL-St	0.130	0.088	0.054	0.087	0.072	0.070	0.056	0.036
		(0.069)	(0.062)	(0.028)	(0.063)	(0.066)	(0.070)	(0.034)	(0.027)
	NL-Tw	0.096	0.079	0.059	0.074	0.066	0.068	0.054	0.005
		(0.062)	(0.054)	(0.034)	(0.044)	(0.060)	(0.063)	(0.032)	(0.018)
	PL-mb	0.150	0.084	0.050	0.087	0.052	0.046	0.044	0.031
		(0.023)	(0.028)	(0.024)	(0.027)	(0.031)	(0.027)	(0.027)	(0.021)
$\tau = 0.1$	mgcv	0.083	0.068	0.060	0.065	0.048	0.036	0.033	0.018
		(0.034)	(0.024)	(0.019)	(0.022)	(0.029)	(0.025)	(0.025)	(0.034)
	mcmc	0.092	0.082	0.079	0.093	0.057	0.052	0.044	0.004
		(0.033)	(0.027)	(0.018)	(0.021)	(0.031)	(0.025)	(0.024)	(0.009)
	PL-Tw	0.141	0.104	0.049	0.106	0.035	0.049	0.054	0.000
		(0.023)	(0.034)	(0.018)	(0.034)	(0.028)	(0.033)	(0.042)	(0.000)
	PL-St	0.177	0.105	0.059	0.137	0.074	0.072	0.064	0.039
		(0.019)	(0.028)	(0.022)	(0.041)	(0.038)	(0.038)	(0.035)	(0.022)
$\tau = 0.3$	NL-Tw	0.125	0.088	0.059	0.089	0.064	0.071	0.069	0.000
		(0.026)	(0.026)	(0.024)	(0.027)	(0.026)	(0.029)	(0.040)	(0.001)
	PL-mb	0.205	0.145	0.066	0.194	0.072	0.063	0.056	0.036
		(0.031)	(0.042)	(0.026)	(0.041)	(0.046)	(0.034)	(0.032)	(0.025)
	mcmc	0.099	0.082	0.086	0.114	0.056	0.049	0.046	0.006
		(0.024)	(0.021)	(0.023)	(0.024)	(0.024)	(0.033)	(0.022)	(0.018)
	PL-Tw	0.077	0.061	0.031	0.056	0.017	0.021	0.021	0.000
		(0.016)	(0.012)	(0.010)	(0.014)	(0.012)	(0.015)	(0.017)	(0.000)
	PL-St	0.118	0.062	0.037	0.065	0.033	0.033	0.030	0.023
		(0.015)	(0.013)	(0.012)	(0.016)	(0.018)	(0.016)	(0.015)	(0.011)
	NL-Tw	0.055	0.051	0.035	0.045	0.034	0.034	0.027	0.000
		(0.013)	(0.010)	(0.011)	(0.013)	(0.012)	(0.011)	(0.012)	(0.000)
	PL-mb	0.164	0.087	0.044	0.092	0.037	0.030	0.028	0.024
		(0.007)	(0.015)	(0.014)	(0.020)	(0.018)	(0.015)	(0.015)	(0.015)
	mcmc	0.056	0.056	0.051	0.073	0.037	0.028	0.024	0.006
		(0.011)	(0.012)	(0.009)	(0.020)	(0.028)	(0.019)	(0.011)	(0.019)

Table A.12: Average (standard deviation) of $\sqrt{ISE}$ over 100 replicates by different approaches, for $N = 500$, $p = 15$, and Student's t errors. (Continued)

		f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
$\tau = 0.5$	PL-Tw	0.068	0.056	0.029	0.045	0.014	0.019	0.018	0.000
		(0.017)	(0.012)	(0.010)	(0.010)	(0.010)	(0.013)	(0.013)	(0.000)
	PL-St	0.107	0.058	0.036	0.051	0.030	0.027	0.026	0.022
		(0.014)	(0.012)	(0.011)	(0.012)	(0.014)	(0.014)	(0.014)	(0.011)
	NL-Tw	0.047	0.045	0.033	0.040	0.029	0.028	0.023	0.000
		(0.013)	(0.010)	(0.010)	(0.009)	(0.009)	(0.009)	(0.010)	(0.000)
	PL-mb	0.157	0.082	0.041	0.075	0.032	0.027	0.026	0.021
		(0.007)	(0.014)	(0.012)	(0.014)	(0.018)	(0.015)	(0.014)	(0.014)
	mcmc	0.052	0.048	0.050	0.060	0.034	0.026	0.026	0.003
		(0.012)	(0.011)	(0.011)	(0.008)	(0.014)	(0.014)	(0.011)	(0.010)
$\tau = 0.7$	PL-Tw	0.081	0.063	0.033	0.048	0.018	0.024	0.019	0.000
		(0.020)	(0.016)	(0.010)	(0.010)	(0.014)	(0.017)	(0.014)	(0.000)
	PL-St	0.123	0.063	0.040	0.054	0.033	0.032	0.028	0.027
		(0.015)	(0.014)	(0.012)	(0.012)	(0.015)	(0.017)	(0.015)	(0.013)
	NL-Tw	0.057	0.050	0.038	0.045	0.033	0.034	0.024	0.000
		(0.015)	(0.011)	(0.012)	(0.011)	(0.012)	(0.011)	(0.010)	(0.000)
	PL-mb	0.165	0.088	0.041	0.073	0.034	0.029	0.028	0.023
		(0.008)	(0.016)	(0.014)	(0.014)	(0.021)	(0.017)	(0.015)	(0.012)
	mcmc	0.057	0.055	0.055	0.062	0.033	0.027	0.024	0.005
		(0.012)	(0.013)	(0.009)	(0.011)	(0.017)	(0.012)	(0.010)	(0.021)
$\tau = 0.9$	PL-Tw	0.149	0.130	0.058	0.080	0.038	0.048	0.047	0.000
		(0.025)	(0.048)	(0.027)	(0.023)	(0.029)	(0.033)	(0.037)	(0.000)
	PL-St	0.187	0.118	0.064	0.083	0.073	0.068	0.061	0.040
		(0.023)	(0.034)	(0.024)	(0.022)	(0.035)	(0.037)	(0.034)	(0.024)
	NL-Tw	0.120	0.096	0.069	0.080	0.060	0.070	0.060	0.000
		(0.025)	(0.028)	(0.028)	(0.028)	(0.025)	(0.027)	(0.034)	(0.002)
	PL-mb	0.212	0.146	0.080	0.120	0.073	0.069	0.055	0.037
		(0.025)	(0.034)	(0.031)	(0.027)	(0.039)	(0.039)	(0.032)	(0.026)
	mcmc	0.094	0.094	0.103	0.100	0.051	0.040	0.040	0.005
		(0.019)	(0.032)	(0.059)	(0.024)	(0.037)	(0.020)	(0.025)	(0.020)

Table A.13: Average (standard deviation) of $\sqrt{ISE}$ over 100 replicates by different approaches, for $N = 500$, $p = 30$, and Student's t errors.

		f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
Mean	PL-Tw	0.100	0.092	0.054	0.085	0.044	0.047	0.044	0.003
		(0.070)	(0.077)	(0.040)	(0.101)	(0.083)	(0.080)	(0.058)	(0.010)
	PL-St	0.154	0.108	0.062	0.105	0.081	0.078	0.065	0.022
		(0.053)	(0.071)	(0.041)	(0.086)	(0.082)	(0.074)	(0.050)	(0.021)
	NL-Tw	0.094	0.083	0.064	0.080	0.067	0.068	0.058	0.002
		(0.043)	(0.064)	(0.046)	(0.066)	(0.079)	(0.074)	(0.052)	(0.009)
	PL-mb	0.173	0.103	0.052	0.106	0.067	0.063	0.054	0.023
		(0.065)	(0.056)	(0.038)	(0.052)	(0.073)	(0.069)	(0.041)	(0.020)
$\tau = 0.1$	mgcv	0.080	0.064	0.060	0.073	0.046	0.039	0.042	0.029
		(0.026)	(0.023)	(0.021)	(0.029)	(0.024)	(0.028)	(0.034)	(0.029)
	mcmc	0.090	0.077	0.079	0.095	0.061	0.052	0.042	0.001
		(0.027)	(0.030)	(0.023)	(0.022)	(0.027)	(0.020)	(0.022)	(0.003)
	PL-Tw	0.142	0.108	0.052	0.107	0.037	0.045	0.055	0.000
		(0.030)	(0.034)	(0.021)	(0.032)	(0.033)	(0.039)	(0.039)	(0.002)
	PL-St	0.193	0.126	0.069	0.181	0.106	0.103	0.076	0.033
		(0.021)	(0.032)	(0.027)	(0.048)	(0.045)	(0.047)	(0.035)	(0.027)
$\tau = 0.3$	NL-Tw	0.117	0.089	0.059	0.090	0.061	0.070	0.069	0.000
		(0.028)	(0.025)	(0.023)	(0.031)	(0.026)	(0.031)	(0.037)	(0.000)
	PL-mb	0.226	0.165	0.072	0.237	0.094	0.086	0.074	0.029
		(0.039)	(0.035)	(0.029)	(0.048)	(0.050)	(0.048)	(0.042)	(0.023)
	mcmc	0.096	0.099	0.074	0.134	0.052	0.040	0.043	0.004
		(0.023)	(0.055)	(0.017)	(0.075)	(0.028)	(0.026)	(0.021)	(0.013)
	PL-Tw	0.076	0.061	0.032	0.055	0.016	0.021	0.021	0.000
		(0.017)	(0.012)	(0.009)	(0.012)	(0.011)	(0.016)	(0.015)	(0.000)
$\tau = 0.3$	PL-St	0.146	0.077	0.043	0.082	0.041	0.044	0.036	0.019
		(0.012)	(0.014)	(0.012)	(0.018)	(0.021)	(0.022)	(0.019)	(0.015)
	NL-Tw	0.054	0.050	0.035	0.045	0.032	0.034	0.027	0.000
		(0.015)	(0.011)	(0.009)	(0.012)	(0.010)	(0.012)	(0.011)	(0.002)
	PL-mb	0.171	0.103	0.049	0.112	0.043	0.038	0.036	0.020
		(0.008)	(0.015)	(0.014)	(0.021)	(0.023)	(0.024)	(0.020)	(0.016)
	mcmc	0.054	0.054	0.053	0.068	0.035	0.034	0.030	0.002
		(0.013)	(0.010)	(0.009)	(0.011)	(0.018)	(0.017)	(0.013)	(0.009)

Table A.14: Average (standard deviation) of $\sqrt{ISE}$ over 100 replicates by different approaches, for $N = 500$, $p = 30$, and Student's t errors. (Continued)

		f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
$\tau = 0.5$	PL-Tw	0.068	0.055	0.031	0.045	0.015	0.018	0.018	0.000
		(0.017)	(0.012)	(0.010)	(0.010)	(0.011)	(0.014)	(0.013)	(0.000)
	PL-St	0.133	0.070	0.041	0.064	0.035	0.034	0.030	0.018
		(0.011)	(0.013)	(0.012)	(0.014)	(0.017)	(0.020)	(0.017)	(0.012)
	NL-Tw	0.046	0.044	0.032	0.038	0.028	0.029	0.023	0.000
		(0.011)	(0.009)	(0.009)	(0.011)	(0.010)	(0.009)	(0.008)	(0.001)
	PL-mb	0.166	0.099	0.042	0.092	0.038	0.034	0.031	0.019
		(0.006)	(0.014)	(0.014)	(0.016)	(0.021)	(0.021)	(0.019)	(0.014)
	mcmc	0.049	0.048	0.048	0.060	0.029	0.025	0.028	0.003
		(0.010)	(0.009)	(0.010)	(0.008)	(0.014)	(0.012)	(0.012)	(0.012)
$\tau = 0.7$	PL-Tw	0.080	0.064	0.034	0.048	0.019	0.019	0.020	0.000
		(0.018)	(0.014)	(0.012)	(0.011)	(0.015)	(0.015)	(0.016)	(0.003)
	PL-St	0.148	0.079	0.043	0.062	0.043	0.043	0.034	0.020
		(0.013)	(0.013)	(0.014)	(0.012)	(0.021)	(0.021)	(0.018)	(0.016)
	NL-Tw	0.056	0.050	0.035	0.044	0.034	0.032	0.027	0.000
		(0.014)	(0.010)	(0.012)	(0.013)	(0.012)	(0.011)	(0.011)	(0.001)
	PL-mb	0.175	0.106	0.046	0.088	0.044	0.037	0.034	0.021
		(0.009)	(0.016)	(0.017)	(0.017)	(0.022)	(0.023)	(0.021)	(0.015)
	mcmc	0.055	0.052	0.050	0.062	0.029	0.029	0.029	0.003
		(0.011)	(0.007)	(0.012)	(0.011)	(0.013)	(0.012)	(0.014)	(0.010)
$\tau = 0.9$	PL-Tw	0.151	0.137	0.058	0.080	0.043	0.047	0.054	0.001
		(0.023)	(0.048)	(0.023)	(0.022)	(0.030)	(0.041)	(0.045)	(0.008)
	PL-St	0.212	0.150	0.071	0.090	0.102	0.095	0.073	0.034
		(0.029)	(0.037)	(0.029)	(0.026)	(0.046)	(0.046)	(0.038)	(0.026)
	NL-Tw	0.123	0.097	0.071	0.080	0.063	0.074	0.067	0.000
		(0.028)	(0.026)	(0.027)	(0.018)	(0.027)	(0.033)	(0.038)	(0.004)
	PL-mb	0.233	0.180	0.100	0.130	0.091	0.084	0.075	0.031
		(0.029)	(0.038)	(0.034)	(0.028)	(0.054)	(0.054)	(0.048)	(0.034)
	mcmc	0.109	0.145	0.132	0.115	0.063	0.056	0.061	0.005
		(0.022)	(0.087)	(0.083)	(0.037)	(0.028)	(0.030)	(0.040)	(0.014)

Table A.15: *Average (standard deviation) of testing errors in mean regression over 100 replicates by different approaches with Student's t errors. "RMSE" and "ADE" indicate the root mean squared errors and the absolute deviation errors, respectively.*

		$N = 300$	$N = 300$	$N = 500$	$N = 500$
		$p = 15$	$p = 30$	$p = 15$	$p = 30$
RMSE	PL-Tw	0.719(0.245)	0.733(0.492)	0.708(0.467)	0.775(0.544)
	PL-St	0.747(0.240)	0.777(0.484)	0.729(0.465)	0.805(0.538)
	Nl-Tw	0.725(0.243)	0.735(0.490)	0.712(0.466)	0.776(0.542)
	PL-mb	0.728(0.263)	0.774(0.376)	0.706(0.391)	0.845(1.032)
	mgcv	0.803(0.474)	0.749(0.240)	0.678(0.217)	0.682(0.234)
	mcmc	0.692(0.214)	0.734(0.426)	0.678(0.210)	0.711(0.432)
ADE	PL-Tw	0.351(0.064)	0.342(0.059)	0.334(0.077)	0.343(0.091)
	PL-St	0.393(0.065)	0.407(0.055)	0.367(0.081)	0.391(0.084)
	Nl-Tw	0.360(0.059)	0.347(0.043)	0.341(0.075)	0.346(0.080)
	PL-mb	0.385(0.037)	0.417(0.052)	0.362(0.035)	0.395(0.084)
	mgcv	0.369(0.049)	0.410(0.104)	0.332(0.030)	0.356(0.039)
	mcmc	0.351(0.046)	0.361(0.057)	0.328(0.028)	0.331(0.030)

Table A.16: *Average (standard deviation) of check loss errors in quantile regression over 100 replicates by different approaches with Student's t errors.*

		$N = 300$	$N = 300$	$N = 500$	$N = 500$
		$p = 15$	$p = 30$	$p = 15$	$p = 30$
$\tau = 0.1$	PL-Tw	0.102(0.012)	0.106(0.017)	0.095(0.015)	0.096(0.012)
	PL-St	0.111(0.013)	0.130(0.018)	0.101(0.015)	0.108(0.014)
	Nl-Tw	0.101(0.012)	0.102(0.016)	0.095(0.015)	0.095(0.013)
	PL-mb	0.119(0.013)	0.136(0.017)	0.105(0.014)	0.116(0.028)
	mgcv	0.121(0.021)	0.123(0.021)	0.110(0.011)	0.112(0.014)
	mcmc	0.104(0.014)	0.120(0.026)	0.092(0.009)	0.096(0.017)
$\tau = 0.3$	PL-Tw	0.147(0.011)	0.148(0.012)	0.140(0.013)	0.140(0.012)
	PL-St	0.160(0.011)	0.171(0.013)	0.149(0.013)	0.158(0.012)
	Nl-Tw	0.143(0.010)	0.143(0.012)	0.137(0.012)	0.138(0.011)
	PL-mb	0.169(0.011)	0.181(0.012)	0.159(0.012)	0.168(0.022)
	mgcv	0.172(0.024)	0.187(0.042)	0.154(0.013)	0.163(0.018)
	mcmc	0.143(0.010)	0.146(0.014)	0.138(0.011)	0.144(0.016)
$\tau = 0.5$	PL-Tw	0.159(0.010)	0.160(0.011)	0.152(0.010)	0.153(0.012)
	PL-St	0.174(0.010)	0.185(0.011)	0.161(0.011)	0.172(0.012)
	Nl-Tw	0.155(0.009)	0.154(0.010)	0.149(0.010)	0.150(0.012)
	PL-mb	0.186(0.010)	0.199(0.011)	0.174(0.010)	0.184(0.017)
	mgcv	0.185(0.025)	0.205(0.052)	0.166(0.015)	0.178(0.020)
	mcmc	0.199(0.018)	0.159(0.013)	0.152(0.009)	0.153(0.011)
$\tau = 0.7$	PL-Tw	0.148(0.010)	0.148(0.011)	0.140(0.009)	0.141(0.013)
	PL-St	0.161(0.010)	0.171(0.012)	0.149(0.009)	0.158(0.013)
	Nl-Tw	0.144(0.010)	0.143(0.009)	0.138(0.009)	0.139(0.013)
	PL-mb	0.170(0.011)	0.182(0.012)	0.159(0.009)	0.167(0.013)
	mgcv	0.168(0.020)	0.148(0.015)	0.154(0.014)	0.164(0.017)
	mcmc	0.144(0.008)	0.195(0.009)	0.141(0.020)	0.139(0.011)
$\tau = 0.9$	PL-Tw	0.106(0.011)	0.107(0.013)	0.095(0.010)	0.097(0.015)
	PL-St	0.114(0.011)	0.126(0.014)	0.100(0.010)	0.109(0.015)
	Nl-Tw	0.104(0.010)	0.102(0.012)	0.094(0.010)	0.096(0.015)
	PL-mb	0.118(0.012)	0.132(0.015)	0.105(0.009)	0.113(0.013)
	mgcv	0.114(0.011)	0.124(0.031)	0.110(0.011)	0.113(0.011)
	mcmc	0.105(0.018)	0.112(0.019)	0.092(0.008)	0.100(0.012)

Table A.17: *Average (standard deviation) of variable selection results over 100 replicates by different approaches, for $N = 300$, $p = 30$, and Gaussian errors.*

		NV	NVT	NL	NLT
Mean	PL-Tw	9.81(2.01)	7.00(0.00)	3.41(0.96)	2.72(0.47)
	PL-St	25.3(3.20)	7.00(0.00)	3.15(1.84)	0.78(0.77)
	Nl-Tw	8.77(1.67)	7.00(0.00)	-	-
	PL-mb	26.1(2.91)	7.00(0.00)	4.40(2.27)	0.91(0.88)
	mgcv	30.0(0.00)	7.00(0.00)	0.00(0.00)	0.00(0.00)
	mcmc	7.23(0.50)	7.00(0.00)	1.63(0.85)	1.60(0.81)
$\tau = 0.1$	PL-Tw	8.32(1.21)	6.89(0.31)	3.52(0.87)	2.85(0.36)
	PL-St	29.6(0.89)	7.00(0.00)	1.03(1.55)	0.20(0.49)
	Nl-Tw	7.71(0.90)	6.91(0.29)	-	-
	PL-mb	29.4(0.93)	7.00(0.00)	4.98(2.50)	0.66(0.74)
	mcmc	8.58(1.43)	6.97(0.18)	4.01(1.47)	2.48(0.63)
$\tau = 0.3$	PL-Tw	7.69(0.95)	7.00(0.00)	3.19(0.68)	2.92(0.27)
	PL-St	29.2(1.56)	7.00(0.00)	1.55(1.57)	0.27(0.55)
	Nl-Tw	7.64(0.93)	7.00(0.00)	-	-
	PL-mb	28.9(1.77)	7.00(0.00)	4.90(2.58)	0.53(0.78)
	mcmc	8.23(1.52)	7.00(0.00)	2.61(0.89)	2.40(0.72)
$\tau = 0.5$	PL-Tw	7.55(0.76)	7.00(0.00)	3.19(0.51)	2.98(0.14)
	PL-St	28.9(1.63)	7.00(0.00)	1.88(1.72)	0.32(0.66)
	Nl-Tw	7.70(0.90)	7.00(0.00)	-	-
	PL-mb	29.2(1.23)	7.00(0.00)	4.64(2.50)	0.75(0.80)
	mcmc	8.93(2.26)	7.00(0.00)	2.79(0.73)	2.60(0.62)
$\tau = 0.7$	PL-Tw	7.69(0.76)	7.00(0.00)	3.26(0.54)	2.97(0.17)
	PL-St	29.2(1.43)	7.00(0.00)	1.87(1.76)	0.32(0.63)
	Nl-Tw	7.70(0.83)	6.99(0.10)	-	-
	PL-mb	28.8(1.64)	7.00(0.00)	5.13(2.65)	0.66(0.74)
	mcmc	8.72(2.02)	7.00(0.00)	2.90(0.99)	2.48(0.63)
$\tau = 0.9$	PL-Tw	8.05(1.12)	6.89(0.31)	3.41(0.82)	2.82(0.41)
	PL-St	29.8(0.88)	7.00(0.00)	0.76(1.05)	0.08(0.27)
	Nl-Tw	7.68(0.98)	6.93(0.26)	-	-
	PL-mb	28.5(2.29)	6.99(0.10)	5.64(2.25)	0.67(0.78)
	mcmc	8.34(1.97)	6.63(0.85)	3.62(1.63)	2.47(0.73)

Table A.18: *Average (standard deviation) of variable selection results over 100 replicates by different approaches, for $N = 500$, $p = 15$, and Gaussian errors.*

		NV	NVT	NL	NLT
Mean	PL-Tw	8.61(1.56)	7.00(0.00)	2.89(0.82)	2.55(0.66)
	PL-St	14.7(0.71)	7.00(0.00)	0.43(0.84)	0.17(0.45)
	Nl-Tw	8.09(1.41)	7.00(0.00)	-	-
	PL-mb	14.8(0.43)	7.00(0.00)	0.66(0.82)	0.29(0.52)
	mgcv	15.0(0.00)	7.00(0.00)	0.00(0.00)	0.00(0.00)
	mcmc	7.28(0.52)	7.00(0.00)	1.46(1.31)	1.27(1.05)
$\tau = 0.1$	PL-Tw	7.17(0.43)	7.00(0.00)	2.99(0.36)	2.93(0.26)
	PL-St	15.0(0.30)	7.00(0.00)	0.13(0.42)	0.05(0.22)
	Nl-Tw	7.13(0.34)	7.00(0.00)	-	-
	PL-mb	15.0(0.22)	7.00(0.00)	0.81(0.97)	0.18(0.41)
	mcmc	7.80(0.89)	7.00(0.00)	3.18(0.83)	2.74(0.45)
$\tau = 0.3$	PL-Tw	7.18(0.44)	7.00(0.00)	3.01(0.30)	2.94(0.24)
	PL-St	14.9(0.44)	7.00(0.00)	0.29(0.64)	0.14(0.40)
	Nl-Tw	7.18(0.41)	7.00(0.00)	-	-
	PL-mb	14.9(0.45)	7.00(0.00)	0.76(1.13)	0.20(0.47)
	mcmc	7.70(0.99)	7.00(0.00)	2.89(1.01)	2.52(0.63)
$\tau = 0.5$	PL-Tw	7.22(0.46)	7.00(0.00)	3.06(0.40)	2.95(0.22)
	PL-St	14.8(0.53)	7.00(0.00)	0.48(0.96)	0.18(0.44)
	Nl-Tw	7.28(0.53)	7.00(0.00)	-	-
	PL-mb	14.9(0.52)	7.00(0.00)	0.70(0.86)	0.26(0.46)
	mcmc	8.11(1.46)	7.00(0.00)	2.54(1.07)	2.37(0.93)
$\tau = 0.7$	PL-Tw	7.20(0.45)	7.00(0.00)	3.06(0.34)	2.97(0.17)
	PL-St	14.9(0.37)	7.00(0.00)	0.24(0.64)	0.09(0.40)
	Nl-Tw	7.33(0.57)	7.00(0.00)	-	-
	PL-mb	14.9(0.40)	7.00(0.00)	0.76(1.13)	0.18(0.41)
	mcmc	8.08(1.11)	7.00(0.00)	2.91(0.91)	2.54(0.68)
$\tau = 0.9$	PL-Tw	7.23(0.47)	7.00(0.00)	3.02(0.45)	2.91(0.32)
	PL-St	15.0(0.17)	7.00(0.00)	0.18(0.52)	0.09(0.32)
	Nl-Tw	7.24(0.45)	7.00(0.00)	-	-
	PL-mb	14.9(0.35)	7.00(0.00)	0.78(1.11)	0.20(0.47)
	mcmc	7.89(1.01)	7.00(0.00)	3.10(0.80)	2.57(0.82)

Table A.19: *Average (standard deviation) of variable selection results over 100 replicates by different approaches, for $N = 500$, $p = 30$, and Gaussian errors.*

		NV	NVT	NL	NLT
Mean	PL-Tw	9.04(1.90)	7.00(0.00)	3.48(0.87)	2.83(0.38)
	PL-St	26.8(2.93)	7.00(0.00)	2.70(1.73)	0.65(0.77)
	Nl-Tw	8.29(1.48)	7.00(0.00)	-	-
	PL-mb	27.4(2.55)	7.00(0.00)	3.71(2.11)	0.66(0.84)
	mgcv	30.0(0.00)	7.00(0.00)	0.00(0.00)	0.00(0.00)
	mcmc	7.30(2.79)	7.00(0.00)	1.43(1.01)	1.30(0.95)
$\tau = 0.1$	PL-Tw	7.29(0.48)	7.00(0.00)	3.12(0.41)	2.98(0.14)
	PL-St	29.3(1.65)	7.00(0.00)	1.22(1.62)	0.18(0.41)
	Nl-Tw	7.26(0.44)	7.00(0.00)	-	-
	PL-mb	29.4(1.20)	7.00(0.00)	3.37(1.99)	0.42(0.67)
	mcmc	8.83(1.46)	7.00(0.00)	3.28(1.31)	2.54(0.57)
$\tau = 0.3$	PL-Tw	7.23(0.55)	7.00(0.00)	3.09(0.40)	2.98(0.14)
	PL-St	28.6(1.78)	7.00(0.00)	1.87(1.89)	0.41(0.64)
	Nl-Tw	7.22(0.48)	7.00(0.00)	-	-
	PL-mb	28.1(2.54)	7.00(0.00)	4.31(2.68)	0.69(0.80)
	mcmc	9.21(2.08)	7.00(0.00)	2.51(1.36)	2.20(0.96)
$\tau = 0.5$	PL-Tw	7.34(0.64)	7.00(0.00)	3.10(0.46)	2.97(0.17)
	PL-St	28.1(2.69)	7.00(0.00)	2.20(1.97)	0.51(0.69)
	Nl-Tw	7.35(0.61)	7.00(0.00)	-	-
	PL-mb	29.0(1.63)	7.00(0.00)	4.04(2.83)	0.65(0.77)
	mcmc	8.94(1.48)	7.00(0.00)	2.73(1.44)	2.22(0.90)
$\tau = 0.7$	PL-Tw	7.27(0.49)	7.00(0.00)	3.04(0.32)	2.98(0.14)
	PL-St	28.7(2.07)	7.00(0.00)	2.14(1.83)	0.39(0.63)
	Nl-Tw	7.32(0.58)	7.00(0.00)	-	-
	PL-mb	28.6(2.41)	7.00(0.00)	4.93(2.72)	0.65(0.80)
	mcmc	8.66(1.75)	7.00(0.00)	2.77(0.97)	2.44(0.77)
$\tau = 0.9$	PL-Tw	7.37(0.60)	7.00(0.00)	3.14(0.43)	2.98(0.14)
	PL-St	29.3(1.49)	7.00(0.00)	1.38(1.73)	0.25(0.54)
	Nl-Tw	7.29(0.56)	7.00(0.00)	-	-
	PL-mb	29.2(1.27)	7.00(0.00)	4.90(2.62)	0.44(0.67)
	mcmc	9.20(2.70)	7.00(0.00)	3.59(1.61)	2.59(0.73)

Table A.20: *Average (standard deviation) of variable selection results over 100 replicates by different approaches, for $N = 300$, $p = 15$, and Student's t errors.*

		NV	NVT	NL	NLT
Mean	PL-Tw	8.71(1.42)	6.96(0.24)	2.85(0.74)	2.47(0.64)
	PL-St	14.3(1.18)	7.00(0.00)	0.72(0.98)	0.40(0.70)
	Nl-Tw	8.37(1.33)	6.97(0.22)	-	-
	PL-mb	14.6(0.79)	7.00(0.00)	0.93(1.03)	0.33(0.59)
	mgcv	15.0(0.00)	7.00(0.00)	0.00(0.00)	0.00(0.00)
	mcmc	7.13(0.51)	6.97(0.18)	1.40(0.93)	1.30(0.95)
$\tau = 0.1$	PL-Tw	7.32(0.58)	7.00(0.00)	3.06(0.45)	2.92(0.27)
	PL-St	15.0(0.26)	7.00(0.00)	0.12(0.46)	0.04(0.24)
	Nl-Tw	7.21(0.48)	6.99(0.10)	-	-
	PL-mb	14.9(0.26)	7.00(0.00)	0.81(0.93)	0.17(0.45)
	mcmc	8.00(1.34)	6.90(0.31)	3.30(1.44)	2.40(0.81)
$\tau = 0.3$	PL-Tw	7.02(0.14)	7.00(0.00)	3.02(0.14)	3.00(0.00)
	PL-St	15.0(0.17)	7.00(0.00)	0.24(0.68)	0.06(0.31)
	Nl-Tw	7.04(0.20)	7.00(0.00)	-	-
	PL-mb	14.9(0.26)	7.00(0.00)	0.94(1.01)	0.30(0.52)
	mcmc	7.34(0.61)	7.00(0.00)	2.50(0.68)	2.47(0.73)
$\tau = 0.5$	PL-Tw	7.02(0.14)	7.00(0.00)	3.02(0.14)	3.00(0.00)
	PL-St	15.0(0.24)	7.00(0.00)	0.22(0.58)	0.09(0.32)
	Nl-Tw	7.06(0.24)	7.00(0.00)	-	-
	PL-mb	15.0(0.17)	7.00(0.00)	0.71(0.84)	0.20(0.43)
	mcmc	7.42(0.90)	7.00(0.00)	2.32(0.96)	2.30(0.95)
$\tau = 0.7$	PL-Tw	7.05(0.26)	7.00(0.00)	3.03(0.17)	3.00(0.00)
	PL-St	15.0(0.10)	7.00(0.00)	0.19(0.42)	0.03(0.17)
	Nl-Tw	7.07(0.36)	7.00(0.00)	-	-
	PL-mb	14.9(0.67)	7.00(0.00)	1.05(1.18)	0.33(0.59)
	mcmc	7.51(1.20)	7.00(0.00)	2.56(0.90)	2.47(0.82)
$\tau = 0.9$	PL-Tw	7.32(0.60)	6.98(0.14)	3.02(0.53)	2.87(0.37)
	PL-St	15.0(0.17)	7.00(0.00)	0.18(0.48)	0.03(0.17)
	Nl-Tw	7.27(0.60)	6.98(0.14)	-	-
	PL-mb	14.9(0.29)	7.00(0.00)	1.01(1.11)	0.23(0.47)
	mcmc	8.11(1.53)	6.83(0.46)	3.50(1.66)	2.32(0.96)

Table A.21: *Average (standard deviation) of variable selection results over 100 replicates by different approaches, for $N = 300$, $p = 30$, and Student's t errors.*

		NV	NVT	NL	NLT
Mean	PL-Tw	9.80(2.30)	6.97(0.30)	3.39(0.94)	2.78(0.46)
	PL-St	25.9(3.13)	7.00(0.00)	2.85(1.73)	0.67(0.82)
	Nl-Tw	8.89(1.77)	7.00(0.00)	-	-
	PL-mb	25.9(2.94)	7.00(0.00)	4.30(2.36)	1.00(0.95)
	mgcv	30.0(0.00)	7.00(0.00)	0.00(0.00)	0.00(0.00)
	mcmc	7.22(0.57)	6.97(0.18)	1.41(0.97)	1.36(0.96)
$\tau = 0.1$	PL-Tw	8.04(1.17)	6.91(0.29)	3.44(0.89)	2.87(0.37)
	PL-St	29.7(1.01)	7.00(0.00)	0.76(1.53)	0.08(0.31)
	Nl-Tw	7.63(0.94)	6.96(0.20)	-	-
	PL-mb	29.3(1.10)	7.00(0.00)	4.58(2.25)	0.58(0.64)
	mcmc	8.90(1.86)	6.80(0.41)	4.60(1.94)	2.61(0.67)
$\tau = 0.3$	PL-Tw	7.04(1.20)	7.00(0.00)	3.01(0.10)	3.00(0.00)
	PL-St	29.2(1.44)	7.00(0.00)	1.85(1.61)	0.25(0.54)
	Nl-Tw	7.07(0.29)	7.00(0.00)	-	-
	PL-mb	29.0(1.35)	7.00(0.00)	4.92(2.17)	0.65(0.70)
	mcmc	7.74(0.87)	7.00(0.00)	2.70(0.53)	2.67(0.55)
$\tau = 0.5$	PL-Tw	7.00(0.00)	7.00(0.00)	3.00(0.00)	3.00(0.00)
	PL-St	29.2(1.22)	7.00(0.00)	1.85(1.88)	0.28(0.53)
	Nl-Tw	7.05(0.22)	7.00(0.00)	-	-
	PL-mb	29.1(1.35)	7.00(0.00)	4.58(2.30)	0.63(0.71)
	mcmc	8.09(1.60)	7.00(0.00)	2.41(0.68)	2.37(0.67)
$\tau = 0.7$	PL-Tw	7.02(0.14)	7.00(0.00)	3.02(0.14)	3.00(0.00)
	PL-St	29.0(1.61)	7.00(0.00)	1.68(1.67)	0.25(0.54)
	Nl-Tw	7.04(0.20)	7.00(0.00)	-	-
	PL-mb	28.9(1.36)	7.00(0.00)	5.17(2.27)	0.65(0.78)
	mcmc	8.00(1.23)	7.00(0.00)	2.40(0.72)	2.40(0.72)
$\tau = 0.9$	PL-Tw	7.80(1.08)	6.93(0.26)	3.39(0.85)	2.85(0.36)
	PL-St	29.9(0.42)	7.00(0.00)	0.65(0.99)	0.11(0.31)
	Nl-Tw	7.43(0.77)	6.97(0.17)	-	-
	PL-mb	29.0(1.71)	7.00(0.00)	4.96(2.39)	0.55(0.66)
	mcmc	8.71(2.27)	6.64(0.72)	4.23(1.94)	2.50(0.51)

Table A.22: *Average (standard deviation) of variable selection results over 100 replicates by different approaches, for $N = 500$, $p = 15$, and Student's t errors.*

		NV	NVT	NL	NLT
Mean	PL-Tw	8.75(1.58)	6.97(0.22)	2.78(1.02)	2.39(0.80)
	PL-St	14.6(0.97)	7.00(0.00)	0.48(0.92)	0.24(0.59)
	Nl-Tw	8.06(1.03)	6.98(0.14)	-	-
	PL-mb	14.7(0.66)	7.00(0.00)	0.60(0.75)	0.22(0.44)
	mgcv	15.0(0.00)	7.00(0.00)	0.00(0.00)	0.00(0.00)
	mcmc	7.24(0.50)	7.00(0.00)	1.71(1.20)	1.61(1.13)
$\tau = 0.1$	PL-Tw	7.15(0.39)	7.00(0.00)	3.03(0.22)	2.99(0.10)
	PL-St	15.0(0.20)	7.00(0.00)	0.21(0.76)	0.12(0.46)
	Nl-Tw	7.12(0.38)	7.00(0.00)	-	-
	PL-mb	15.0(0.22)	7.00(0.00)	0.76(1.07)	0.25(0.52)
	mcmc	7.71(0.91)	7.00(0.00)	2.49(1.01)	2.20(0.81)
$\tau = 0.3$	PL-Tw	7.00(0.00)	7.00(0.00)	3.00(0.00)	3.00(0.00)
	PL-St	15.0(0.00)	7.00(0.00)	0.06(0.31)	0.02(0.14)
	Nl-Tw	7.03(0.17)	7.00(0.00)	-	-
	PL-mb	15.0(0.22)	7.00(0.00)	0.58(0.81)	0.17(0.38)
	mcmc	7.42(0.82)	7.00(0.00)	2.59(0.73)	2.52(0.73)
$\tau = 0.5$	PL-Tw	7.00(0.00)	7.00(0.00)	3.00(0.00)	3.00(0.00)
	PL-St	15.0(0.10)	7.00(0.00)	0.05(0.22)	0.02(0.14)
	Nl-Tw	7.01(0.10)	7.00(0.00)	-	-
	PL-mb	15.0(0.20)	7.00(0.00)	0.56(0.81)	0.12(0.33)
	mcmc	7.60(1.13)	7.00(0.00)	2.18(1.05)	2.18(1.05)
$\tau = 0.7$	PL-Tw	7.01(0.10)	7.00(0.00)	3.01(0.10)	3.00(0.00)
	PL-St	15.0(0.10)	7.00(0.00)	0.11(0.31)	0.04(0.20)
	Nl-Tw	7.01(0.10)	7.00(0.00)	-	-
	PL-mb	15.0(0.14)	7.00(0.00)	0.48(0.87)	0.14(0.38)
	mcmc	7.50(0.97)	7.00(0.00)	2.36(0.96)	2.30(0.88)
$\tau = 0.9$	PL-Tw	7.05(0.22)	7.00(0.00)	3.00(0.20)	2.98(0.14)
	PL-St	14.9(0.35)	7.00(0.00)	0.27(0.69)	0.06(0.28)
	Nl-Tw	7.15(0.36)	7.00(0.00)	-	-
	PL-mb	14.9(0.33)	7.00(0.00)	0.65(0.95)	0.15(0.36)
	mcmc	7.66(0.99)	6.97(0.18)	2.70(1.02)	2.40(0.86)

Table A.23: *Average (standard deviation) of variable selection results over 100 replicates by different approaches, for $N = 500$, $p = 30$, and Student's t errors.*

		NV	NVT	NL	NLT
Mean	PL-Tw	9.44(2.28)	6.90(0.59)	3.34(0.97)	2.58(0.67)
	PL-St	26.2(3.71)	6.94(0.45)	2.50(1.89)	0.72(0.88)
	Nl-Tw	8.64(2.63)	6.91(0.59)	-	-
	PL-mb	26.5(3.77)	6.96(0.28)	3.86(2.03)	0.77(0.86)
	mgcv	30.0(0.00)	7.00(0.00)	0.00(0.00)	0.00(0.00)
	mcmc	7.30(0.70)	7.00(0.00)	1.34(1.09)	1.17(0.99)
$\tau = 0.1$	PL-Tw	7.21(0.50)	7.00(0.00)	3.11(0.37)	2.99(0.10)
	PL-St	29.3(1.70)	7.00(0.00)	1.35(1.46)	0.23(0.45)
	Nl-Tw	7.15(0.41)	7.00(0.00)	-	-
	PL-mb	29.4(1.41)	7.00(0.00)	4.24(2.33)	0.54(0.77)
	mcmc	8.24(1.07)	7.00(0.00)	3.10(1.24)	2.48(0.68)
$\tau = 0.3$	PL-Tw	7.01(0.10)	7.00(0.00)	3.01(0.10)	3.00(0.00)
	PL-St	29.1(1.53)	7.00(0.00)	1.72(1.78)	0.34(0.57)
	Nl-Tw	7.03(0.17)	7.00(0.00)	-	-
	PL-mb	29.4(1.09)	7.00(0.00)	3.94(2.34)	0.54(0.67)
	mcmc	9.06(3.02)	7.00(0.00)	2.33(0.99)	2.22(0.82)
$\tau = 0.5$	PL-Tw	7.00(0.00)	7.00(0.00)	3.00(0.00)	3.00(0.00)
	PL-St	29.5(1.43)	7.00(0.00)	1.37(1.67)	0.23(0.47)
	Nl-Tw	7.03(0.17)	7.00(0.00)	-	-
	PL-mb	29.3(1.03)	7.00(0.00)	3.66(2.15)	0.49(0.61)
	mcmc	8.01(1.13)	7.00(0.00)	2.44(0.77)	2.40(0.86)
$\tau = 0.7$	PL-Tw	7.02(0.14)	7.00(0.00)	3.01(0.10)	3.00(0.00)
	PL-St	29.1(1.89)	7.00(0.00)	1.40(1.79)	0.25(0.54)
	Nl-Tw	7.04(0.20)	7.00(0.00)	-	-
	PL-mb	29.1(1.42)	7.00(0.00)	4.05(2.33)	0.61(0.82)
	mcmc	8.11(1.20)	7.00(0.00)	2.40(0.86)	2.39(0.81)
$\tau = 0.9$	PL-Tw	7.26(0.60)	6.99(0.10)	3.10(0.44)	2.97(0.17)
	PL-St	29.4(1.32)	7.00(0.00)	1.36(1.77)	0.23(0.55)
	Nl-Tw	7.21(0.52)	7.00(0.00)	-	-
	PL-mb	29.4(1.03)	7.00(0.00)	4.23(2.40)	0.49(0.61)
	mcmc	9.26(1.76)	6.91(0.25)	3.43(1.52)	2.06(0.98)

A.3 Sample R Codes

```

library(MASS)
library(stats)

#####
##                                                    ##
##          Truncated Power Spline Basis Matrices      ##
##          and Second Derivative Penalty Matrices     ##
##                                                    ##
#####

TPS<-function(X,q=3,knots=9,remove=2,a=NULL,b=NULL,c=NULL)

#Input list:

#X(matrix): predictor matrix
#need to be pre-rescaled so that supported in [0,1]

#q(scalar/vector): default=3
#(scalar) degree of TPS, the same for all covariates
#(vector) degrees of TPS, specific for each covariate

#knots(scalar/vector/matrix): default=9
#(scalar) number of equidistant knots
#          the same for all covariates
#(vector of integers) numbers of equidistant knots
#          specific for each covariate
#(vector of values in (0,1)) sequence of internal knots

```



```
#                                the same for all covariates
#(matrix) sequences of internal knots
#                                specific for each covariate

#remove(scalar/vector): default=2
#indicator of whether to remove B_-1 and/or B_0 (see Thesis)
#(=0) do not remove
#(=1) remove B_-1 only
#(=2) remove B_-1 and B_0
#(scalar) the same for all covariates
#(vector) specific for each covariate

#a(vector/matrix): default=NULL
#first set of coefficients to modify TPS basis (see Thesis)
#(=NULL) do as in Thesis
#(vector) user-specified coefficients
#                                the same for all covariates
#(matrix) user-specified coefficients
#                                specific for each covariate

#b(vector/matrix): default=NULL
#second set of coefficients to modify TPS basis (see Thesis)
#(=NULL) do as in Thesis
#(vector) user-specified coefficients
#                                the same for all covariates
#(matrix) user-specified coefficients
#                                specific for each covariate

#c(scalar/vector): default=NULL
```

```
#coefficient to modify B_0 as B_0-c
#if B_0 is not removed (remove = 0 or 1)
#(=NULL) use sample averages
#(scalar) user-specified coefficient
#           the same for all covariates
#(vector) user-specified coefficients
#           specific for each covariate

#End of input list

#Output list:

#B(N*...*p array): truncated power spline basis matrices
#specific for each covariate

#Omega(...*...*p array): second derivative penalty matrices
#specific for each covariate

#q(p vector): degrees of TPS, specific for each covariate

#knots(p*... matrix): sequences of internal knots
#specific for each covariate

#remove(p vector):
#indicators of whether to remove B_{-1} and/or B_0 (see Thesis)
#(=0) do not remove
#(=1) remove B_{-1} only
#(=2) remove B_{-1} and B_0
#specific for each covariate
```

```

#K(p vector): numbers of basic TPS
#specific for each covariate

#a(p*... matrix):
#first set of coefficients to modify TPS basis (see Thesis)
#specific for each covariate

#b(p*... matrix):
#second set of coefficients to modify TPS basis (see Thesis)
#specific for each covariate

#c(p vector):
#coefficients to modify B_0 as B_0-c
#if B_0 is not removed (remove = 0 or 1)
#specific for each covariate

#End of output list

{
### Preparations ###
N<-dim(X)[1] #sample size
p<-dim(X)[2] #number of covariates
if(length(q)==1) q<-rep(q,p)

if(length(knots)==1)
  knots<-seq(0,1,length=(knots+2))[2:(knots+1)]
if(is.null(dim(knots))&&knots[1]<1)
  knots<-matrix(rep(knots,p),nrow=p,byrow=TRUE)

```

```

if(is.null(dim(knots))&& knots[1]>=1)
{
  temp<-knots
  knots<-matrix(0,nrow=p,ncol=max(temp))
  for(j in 1:p)
  {
    knots[j,1:temp[j]]<-
      seq(0,1,length=(temp[j]+2))[2:(temp[j]+1)]
  }
}

if(length(remove)==1) remove<-rep(remove,p)
K<-rep(0,p) #number of basic TPS for each covariate
for(j in 1:p)
{
  K[j]<-q[j]+sum(knots[j,]>0)+1-remove[j]
}

if(is.null(a))
{
  a<-matrix(0,nrow=p,ncol=max(K+remove-2))
  for(j in 1:p)
  {
    for(m in 1:(q[j]-1))
    {
      a[j,m]<--cov(X[,j],X[,j]^(m+1))/var(X[,j])
    }
    for(n in q[j):(K[j]+remove[j]-2))
    {

```

```

        temp<-(X[,j]-knots[j,(n-q[j]+1)])^q[j]*
            (X[,j]>knots[j,(n-q[j]+1)])
        a[j,n]<--cov(X[,j],temp)/var(X[,j])
    }
}

if(is.null(dim(a)))
    a<-matrix(rep(a,p),nrow=p,byrow=TRUE)

if(is.null(b))
{
    b<-matrix(0,nrow=p,ncol=max(K+remove-2))
    for(j in 1:p)
    {
        for(m in 1:(q[j]-1))
        {
            b[j,m]<-(cov(X[,j]^2,X[,j]^(m+1))-
                cov(X[,j],X[,j]^(m+2)))/var(X[,j])
        }
        for(n in q[j):(K[j]+remove[j]-2))
        {
            temp<-(X[,j]-knots[j,(n-q[j]+1)])^q[j]*
                (X[,j]>knots[j,(n-q[j]+1)])
            b[j,n]<-(cov(X[,j]^2,temp)-
                cov(X[,j],X[,j]*temp))/var(X[,j])
        }
    }
}

if(is.null(dim(b)))

```

```

b<-matrix(rep(b,p),nrow=p,byrow=TRUE)

if(is.null(c))
{
  c<-rep(0,p)
  for(j in 1:p) c[j]<-mean(X[,j])
}
if(length(c)==1) c<-rep(c,p)

### B and Omega ###
B<-array(0,dim=c(N,max(K),p))
Omega<-array(0,dim=c(max(K),max(K),p))

for(j in 1:p)
{
  if(remove[j]<1)
  {
    B[,1,j]<-rep(1,N)
  }
  if(remove[j]<2)
  {
    B[(2-remove[j]),j]<-X[,j]-c[j]
  }
  for(m in (3-remove[j]):(q[j]+1-remove[j]))
  {
    B[,m,j]<-X[,j]^(m-1+remove[j])+
      a[j,(m-2+remove[j])]*X[,j]+
      b[j,(m-2+remove[j])]
  }
}

```

```

for(n in (q[j]+2-remove[j]):K[j])
{
  temp<-(X[,j]-knots[j,(n-q[j]-1+remove[j]))]^q[j]*
    (X[,j]>knots[j,(n-q[j]-1+remove[j]))]
  B[,n,j]<-temp+
    a[j,(n-2+remove[j])]*X[,j]+
    b[j,(n-2+remove[j])]
}

for(m in (3-remove[j]):(q[j]+1-remove[j]))
{
  for(n in (3-remove[j]):(q[j]+1-remove[j]))
  {
    f<-function(x)
    {
      x^(m-3+remove[j])*x^(n-3+remove[j])
    }
    Omega[m,n,j]<-(m-2+remove[j])*(n-2+remove[j])*
      (m-1+remove[j])*(n-1+remove[j])*
      integrate(f,0,1)$value
  }
}

for(m in (q[j]+2-remove[j]):K[j])
{
  for(n in (3-remove[j]):(q[j]+1-remove[j]))
  {
    f<-function(x)
    {
      (x-knots[j,(m-q[j]-1+remove[j]))]^(q[j]-2)*

```

```

      (x^(n-3+remove[j]))
    }
    Omega[m,n,j]<-(n-2+remove[j])*(n-1+remove[j])*
      (q[j]-1)*q[j]*
      integrate(f,
        knots[j,(m-q[j]-1+remove[j])],1)$value
    Omega[n,m,j]<-Omega[m,n,j]
  }
}
for(m in (q[j]+2-remove[j]):K[j])
{
  for(n in (q[j]+2-remove[j]):K[j])
  {
    f<-function(x)
    {
      (x-knots[j,(m-q[j]-1+remove[j])])^(q[j]-2)*
      (x-knots[j,(n-q[j]-1+remove[j])])^(q[j]-2)
    }
    Omega[m,n,j]<-(q[j]-1)^2*q[j]^2*
      integrate(f,
        max(knots[j,(m-q[j]-1+remove[j])],
          knots[j,(n-q[j]-1+remove[j])]),1)$value
  }
}
}#end of j in 1:p

return(list(B=B,Omega=Omega,q=q,knots=knots,
  remove=remove,K=K,a=a,b=b,c=c))
}

```



```
#####
##                                                                 ##
##           Boosting for PLAM (Mean Regression)                  ##
##                                                                 ##
#####

PboostM<-function(X,Y,M=NULL,Xv=NULL,Yv=NULL,Xp=NULL,
                  q=3,knots=9,W=rep(1,2*dim(X)[2]))

#Input list:

#X(matrix): predictor matrix
#need to be pre-rescaled so that supported in [0,1]

#Y(vector): response vector

#M(scalar): default=NULL
#desired number of iterations
#(=NULL) validation data required

#Xv(matrix): validation predictor matrix, default=NULL
#need to be pre-rescaled so that supported in [0,1]

#Yv(vector): validation response vector, default=NULL

#Xp(matrix): prediction matrix, default=NULL
```

```
#need to be pre-rescaled so that supported in [0,1]

#q(scalar/vector): default=3
#(scalar) degree of TPS, the same for all covariates
#(vector) degrees of TPS, specific for each covariate

#knots(scalar/vector/matrix): default=9
#(scalar) number of equidistant knots
#           the same for all covariates
#(vector of integers) numbers of equidistant knots
#           specific for each covariate
#(vector of values in (0,1)) sequence of internal knots
#           the same for all covariates
#(matrix) sequences of internal knots
#           specific for each covariate

#W(vector): default=rep(1,2*dim(X)[2]).
#weights for linear (W[1:p]) & nonlinear (W[(p+1):(2p)])
#parts of covariate effects.

#End of input list

#Output list:

#beta0(scalar): estimated intercept

#F(N*(2p) matrix): estimated function values for training data
#F[,1:p] for linear parts & F[, (p+1):(2p)] for nonlinear parts
```

```

#Yf(N vector): fitted response vector

#Fp(Np*(2p) matrix): estimated function values for prediction data
#Fp[,1:p] for linear parts & Fp[(p+1):(2p)] for nonlinear parts
#NULL if Xp is NULL

#Yp(Np vector): predicted response vector
#NULL if Xp is NULL

#vs(... vector):
#which part of which covariate is chosen in every iteration
#1:p for linear parts of covariates 1:p
#(p+1):(2p) for nonlinear parts of covariates 1:p

#te(2*... matrix): training RMSE and ADE in every iteration

#tev(2*... matrix): validation RMSE and ADE in every iteration
#NULL if Xv is NULL

#End of output list

{
### Preparations ###
nu<-0.1 #step size
#max number of iterations
if(!is.null(M)) Mmax<-M
if(is.null(M)) Mmax<-10000
N<-dim(X)[1] #training sample size
if(is.null(M)) Nv<-dim(Xv)[1] #validation sample size

```

```

if(!is.null(Xp)) Np<-dim(Xp)[1] #prediction sample size
p<-dim(X)[2] #number of covariates

tps<-TPS(X,q,knots,remove=2,a=NULL,b=NULL,c=NULL)
B<-tps$B #basic TPS matrices of training data
Omega<-tps$Omega #second derivative penalty matrices
q<-tps$q
knots<-tps$knots
K<-tps$K #number of basic TPS for each covariate
a<-tps$a
b<-tps$b
c<-tps$c
if(is.null(M)) Bv<-TPS(Xv,q=q,knots=knots,remove=2,a=a,b=b,c=c)$B
  #basic TPS matrices of validation data
if(!is.null(Xp)) Bp<-TPS(Xp,q=q,knots=knots,remove=2,a=a,b=b,c=c)$B
  #basic TPS matrices of prediction data
lambda<-rep(0,p) #tunning parameters for penalized splines
for(j in 1:p)
{
  lambda[j]<-uniroot(function(x)
    sum(diag(B[,1:K[j],j])%*%
      solve(t(B[,1:K[j],j])%*%
        B[,1:K[j],j]+x*Omega[1:K[j],1:K[j],j])%*%
        t(B[,1:K[j],j]))) -1,
    lower=10^(-6),upper=10^6)[[1]]
}

### Centralization of X, Xv & Xp ###
for(j in 1:p)

```

```

{
  X[,j]<-X[,j]-c[j]
  if(is.null(M)) Xv[,j]<-Xv[,j]-c[j]
  if(!is.null(Xp)) Xp[,j]<-Xp[,j]-c[j]
}

### Storage of frequently used and unchanged matrices ###
HL<-array(0,dim=c(p,N))
HN<-array(0,dim=c(max(K),N,p))
for(j in 1:p)
{
  HL[j,]<-solve(t(X[,j])%*%X[,j])%*%t(X[,j])
  HN[1:K[j],,j]<-solve(t(B[,1:K[j],j])%*%B[,1:K[j],j]+
    lambda[j]*Omega[1:K[j],1:K[j],j])%*%t(B[,1:K[j],j]))
}

### Initialization ###
beta0<-rep(0,Mmax)
beta<-matrix(0,nrow=p,ncol=Mmax)
theta<-array(0,dim=c(max(K),p,Mmax))
F<-array(0,dim=c(N,2*p,Mmax))
vs<-rep(0,Mmax)
te<-matrix(0,nrow=2,ncol=Mmax)
tev<-NULL
if(is.null(M)) tev<-matrix(0,nrow=2,ncol=Mmax)
eta<-rep(0,N) #current estimation (linear+nonlinear)
U<-2*(Y-eta) #current negative gradients
m<-0

```

```

### Iterating ###
while(m<Mmax)
{
  m<-m+1

### Find the best linear fit ###
  Jlin<-1
  b<-HL[1,]%*%U
  Lin<-(t(U)%*%U-t(U-X[,1]*b)%*%(U-X[,1]*b))*W[1]
  for(j in 2:p)
  {
    bj<-HL[j,]%*%U
    Linj<-(t(U)%*%U-t(U-X[,j]*bj)%*%(U-X[,j]*bj))*W[j]
    if(Linj>Lin)
    {
      Lin<-Linj
      b<-bj
      Jlin<-j
    }
  }

### Find the best nonlinear fit ###
  Jnon<-1
  th<-HN[1:K[1],,1]%*%U
  Non<-(t(U)%*%U-t(U-B[,1:K[1],1]%*%th)%*%
    (U-B[,1:K[1],1]%*%th))*W[p+1]
  for(j in 2:p)
  {
    thj<-HN[1:K[j],,j]%*%U

```

```

Nonj<-(t(U)%*%U-t(U-B[,1:K[j],j]%*%thj)%*%
      (U-B[,1:K[j],j]%*%thj))*W[p+j]
if (Nonj>Non)
{
  Non<-Nonj
  th<-thj
  Jnon<-j
}
}

### Compare linear & nonlinear fit ###
if (Lin>=Non)
{
  beta[Jlin,m]<-beta[Jlin,m]+nu*b
  F[,Jlin,m]<-F[,Jlin,m]+nu*X[,Jlin]*b
  eta<-eta+nu*X[,Jlin]*b
  vs[m]<-Jlin
}

if (Non>Lin)
{
  theta[1:K[Jnon],Jnon,m]<-theta[1:K[Jnon],Jnon,m]+nu*th
  F[,p+Jnon,m]<-F[,p+Jnon,m]+nu*B[,1:K[Jnon],Jnon]%*%th
  eta<-eta+nu*B[,1:K[Jnon],Jnon]%*%th
  vs[m]<-p+Jnon
}

### Update estimated intercept ###
if ((m/10)==floor(m/10))

```

```
{
  b0<-mean(Y-eta)
  beta0[m]<-beta0[m]+b0
  eta<-eta+b0
}

### Update negative gradients ###
U<-2*(Y-eta)

### Calculate training errors ###
te[1,m]<-mean((eta-Y)^2)^0.5
te[2,m]<-mean(abs(eta-Y))

### Calculate validation errors ###
if(is.null(M))
{
  etav<-beta0[m]*rep(1,Nv)+Xv%%beta[,m]
  for(j in 1:p) etav<-etav+Bv[,1:K[j],j]%%theta[1:K[j],j,m]
  tev[1,m]<-mean((etav-Yv)^2)^0.5
  tev[2,m]<-mean(abs(etav-Yv))
}

### Update estimates ###
if(m<Mmax)
{
  beta0[m+1]<-beta0[m]
  beta[,m+1]<-beta[,m]
  theta[,m+1]<-theta[,m]
  F[,m+1]<-F[,m]
```



```

    }

    } #End of iterations

if(!is.null(M)) stop<-Mmax
if(is.null(M)) stop<-which.min(tev[1,])
Yf<-beta0[stop]*rep(1,N)+X%%beta[,stop]
Yp<-NULL
Fp<-NULL
if(!is.null(Xp))
{
    Yp<-beta0[stop]*rep(1,Np)+Xp%%beta[,stop]
    Fp<-matrix(0,nrow=Np,ncol=2*p)
}
for(j in 1:p)
{
    Yf<-Yf+B[,1:K[j],j]%%theta[1:K[j],j,stop]
    if(!is.null(Xp))
    {
        Fp[,j]<-Xp[,j]*beta[j,stop]
        Fp[,p+j]<-Bp[,1:K[j],j]%%theta[1:K[j],j,stop]
        Yp<-Yp+Fp[,p+j]
    }
}
return(list(beta0=beta0[stop],F=F[, ,stop],Yf=Yf,Fp=Fp,
           Yp=Yp,vs=vs[1:stop],te=te,tev=tev))
}

```

References

- Allison, P. D. *Logistic regression using SAS: Theory and application*. SAS Institute (2012).
- Aneiros-Pérez, G. and Vieu, P. “Nonparametric time series prediction: A semi-functional partial linear modeling.” *Journal of Multivariate Analysis*, 99(5):834–857 (2008).
- Arjas, E. “A graphical method for assessing goodness of fit in Cox’s proportional hazards model.” *Journal of the American Statistical Association*, 83(401):204–212 (1988).
- Avriel, M. *Nonlinear programming: Analysis and methods*. Courier Corporation (2003).
- Bartlett, P. L. and Traskin, M. “Adaboost is consistent.” *Journal of Machine Learning Research*, 8:2347–2368 (2007).
- Bender, R., Augustin, T., and Blettner, M. “Generating survival times to simulate Cox proportional hazards models.” *Statistics in Medicine*, 24(11):1713–1723 (2005).

- Benner, A., Zucknick, M., Hielscher, T., Ittrich, C., and Mansmann, U. "High-dimensional Cox models: The choice of penalty as part of the model building process." *Biometrical Journal*, 52(1):50–69 (2010).
- Betensky, R. A., Lindsey, J. C., Ryan, L. M., and Wand, M. P. "A local likelihood proportional hazards model for interval censored data." *Statistics in Medicine*, 21(2):263–275 (2002).
- Binder, H., Allignol, A., Schumacher, M., and Beyersmann, J. "Boosting for high-dimensional time-to-event data with competing risks." *Bioinformatics*, 25(7):890–896 (2009).
- Binder, H. and Schumacher, M. "Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models." *BMC Bioinformatics*, 9(1):14–23 (2008).
- Binder, H. and Tutz, G. "A comparison of methods for the fitting of generalized additive models." *Statistics and Computing*, 18(1):87–99 (2008).
- Bollaerts, K., Eilers, P. H. C., and Aerts, M. "Quantile regression with monotonicity restrictions using P-splines and the L_1 -norm." *Statistical Modelling*, 6(3):189–207 (2006).
- Bontemps, C., Simioni, M., and Surry, Y. "Semiparametric hedonic price models: Assessing the effects of agricultural nonpoint source pollution." *Journal of Applied Econometrics*, 23(6):825–842 (2008).
- Bøvelstad, H. M., Nygard, S., Størvold, H. L., Aldrin, M., Borgan, Ø., Frigessi, A., and Lingjærde, O. C. "Predicting survival from microarray data: A comparative study." *Bioinformatics*, 23(16):2080–2087 (2007).

- Breiman, L. “Arcing the edge.” Technical report, Technical Report 486, Statistics Department, University of California at Berkeley (1997).
- . “Arcing classifier (with discussion and a rejoinder by the author).” *The Annals of Statistics*, 26(3):801–849 (1998).
- . “Prediction games and arcing algorithms.” *Neural Computation*, 11(7):1493–1517 (1999).
- Breslow, N. “Discussion of Professor Cox’s paper.” *Journal of the Royal Statistical Society: Series B*, 34:216–217 (1972).
- . “Covariance analysis of censored survival data.” *Biometrics*, 89–99 (1974).
- Brown, D., Kauermann, G., and Ford, I. “A partial likelihood approach to smooth estimation of dynamic covariate effects using penalised splines.” *Biometrical Journal*, 49(3):441–452 (2007).
- Buchinsky, M. “Changes in the US wage structure 1963-1987: Application of quantile regression.” *Econometrica*, 62(2):405–458 (1994).
- . “Estimating the asymptotic covariance matrix for quantile regression models a Monte Carlo study.” *Journal of Econometrics*, 68(2):303–338 (1995).
- . “Recent advances in quantile regression models: A practical guideline for empirical research.” *Journal of Human Resources*, 33(1):88–126 (1998).
- Buchinsky, M. and Hahn, J. “An alternative estimator for the censored quantile regression model.” *Econometrica*, 653–671 (1998).

- Bühlmann, P. “Boosting for high-dimensional linear models.” *The Annals of Statistics*, 559–583 (2006).
- Bühlmann, P. and Hothorn, T. “Boosting algorithms: Regularization, prediction and model fitting.” *Statistical Science*, 22(4):477–505 (2007).
- . “Twin boosting: Improved feature selection and prediction.” *Statistics and Computing*, 20(2):119–138 (2010).
- Bühlmann, P. and Yu, B. “Boosting with the L_2 loss: Regression and classification.” *Journal of the American Statistical Association*, 98(462):324–339 (2003).
- Buja, A., Hastie, T., and Tibshirani, R. “Linear smoothers and additive models.” *The Annals of Statistics*, 453–510 (1989).
- Cai, J., Fan, J., Jiang, J., and Zhou, H. “Partially linear hazard regression for multivariate survival data.” *Journal of the American Statistical Association*, 102(478):538–551 (2007).
- Cai, T. and Betensky, R. A. “Hazard regression for interval-censored data with penalized spline.” *Biometrics*, 59(3):570–579 (2003).
- Cai, T., Hyndman, R. J., and Wand, M. P. “Mixed model-based hazard estimation.” *Journal of Computational and Graphical Statistics*, 11(4):784–798 (2002).
- Cauchy, A. “Méthode générale pour la résolution des systemes déquations simultanées.” *Comp. Rend. Sci. Paris*, 25(1847):536–538 (1847).
- Champion, M., Cierco-Ayrolles, C., Gadat, S., and Vignes, M. “Sparse regression and support recovery with L_2 -boosting algorithms.” *Journal of Statistical Planning and Inference*, 155:19–41 (2014).

- Chang, M. N. and Yang, G. L. “Strong consistency of a nonparametric estimator of the survival function with doubly censored data.” *The Annals of Statistics*, 1536–1547 (1987).
- Chernozhukov, V., Fernández-Val, I., and Galichon, A. “Quantile and probability curves without crossing.” *Econometrica*, 78(3):1093–1125 (2010).
- Choudhary, P. K. “Semiparametric regression for assessing agreement using tolerance bands.” *Computational Statistics & Data Analysis*, 51(12):6229–6241 (2007).
- Cole, T. J. and Green, P. J. “Smoothing reference centile curves: The LMS method and penalized likelihood.” *Statistics in Medicine*, 11(10):1305–1319 (1992).
- Cox, D. R. “Regression models and life-tables.” *Journal of the Royal Statistical Society: Series B*, 187–220 (1972).
- . “Partial likelihood.” *Biometrika*, 62(2):269–276 (1975).
- Cox, D. R. and Oakes, D. *Analysis of survival data*. CRC Press (1984).
- de Boor, C. *A practical guide to splines*. Springer (1978).
- Dette, H. and Volgushev, S. “Non-crossing non-parametric estimates of quantile curves.” *Journal of the Royal Statistical Society: Series B*, 70(3):609–627 (2008).
- Drucker, H., Cortes, C., Jackel, L. D., LeCun, Y., and Vapnik, V. “Boosting and other ensemble methods.” *Neural Computation*, 6(6):1289–1301 (1994).

- Efron, B. “The efficiency of Cox’s likelihood function for censored data.” *Journal of the American statistical Association*, 72(359):557–565 (1977).
- Efron, B. and Gong, G. “A leisurely look at the bootstrap, the jackknife, and cross-validation.” *The American Statistician*, 37(1):36–48 (1983).
- Eubank, R. L. *Nonparametric regression and spline smoothing*. CRC Press (1999).
- Eubank, R. L., Huang, C., and Wang, S. “Adaptive order selection for spline smoothing.” *Journal of Computational and Graphical Statistics*, 382–397 (2012).
- Fahrmeir, L. and Lang, S. “Bayesian semiparametric regression analysis of multicategorical time-space data.” *Annals of the Institute of Statistical Mathematics*, 53(1):11–30 (2001).
- Fan, J. and Li, R. “Variable selection via nonconcave penalized likelihood and its oracle properties.” *Journal of the American Statistical Association*, 96(456):1348–1360 (2001).
- . “Variable selection for Cox’s proportional hazards model and frailty model.” *Annals of Statistics*, 74–99 (2002).
- Feigl, P. and Zelen, M. “Estimation of exponential survival probabilities with concomitant information.” *Biometrics*, 826–838 (1965).
- Fenske, N., Burns, J., Hothorn, T., and Rehfuess, E. A. “Understanding child stunting in India: A comprehensive analysis of socio-economic, nutritional and environmental determinants using additive quantile regression.” *PloS one*, 8(11):e78692 (2013a).

- Fenske, N., Fahrmeir, L., Hothorn, T., Rzehak, P., and Höhle, M. “Boosting structured additive quantile regression for longitudinal childhood obesity data.” *The International Journal of Biostatistics*, 9(1):1–18 (2013b).
- Fenske, N., Kneib, T., and Hothorn, T. “Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression.” *Journal of the American Statistical Association*, 106(494):494–510 (2011).
- Fine, J. P. and Gray, R. J. “A proportional hazards model for the sub-distribution of a competing risk.” *Journal of the American Statistical Association*, 94(446):496–509 (1999).
- Fisher, L. D. and Lin, D. Y. “Time-dependent covariates in the Cox proportional-hazards regression model.” *Annual Review of Public Health*, 20(1):145–157 (1999).
- Fox, M. and Rubin, H. “Admissibility of quantile estimates of a single location parameter.” *The Annals of Mathematical Statistics*, 1019–1030 (1964).
- Freund, Y. “Boosting a weak learning algorithm by majority.” *Information and Computation*, 121(2):256–285 (1995).
- Freund, Y. and Schapire, R. E. “A decision-theoretic generalization of on-line learning and an application to boosting.” In *Computational Learning Theory*, 23–37 (1995).
- . “Experiments with a new boosting algorithm.” In *ICML*, volume 96, 148–156 (1996).

- . “A decision-theoretic generalization of on-line learning and an application to boosting.” *Journal of Computer and System Sciences*, 55(1):119–139 (1997).
- Friedman, J., Hastie, T., and Tibshirani, R. “Additive logistic regression: A statistical view of boosting.” *The Annals of Statistics*, 28(2):337–407 (2000).
- Friedman, J. H. “Greedy function approximation: A gradient boosting machine.” *The Annals of Statistics*, 29(5):1189–1232 (2001).
- . “Stochastic gradient boosting.” *Computational Statistics & Data Analysis*, 38(4):367–378 (2002).
- Frome, E. L. “The analysis of rates using Poisson regression models.” *Biometrics*, 665–674 (1983).
- Geiß, S., Löffler, D., Körner, B., Engelke, M., Sawal, G., and Bachhausen, P. “Determination of the sum of short chain chlorinated n-alkanes with a chlorine content between 50% and 67% in sediment samples by GC–ECNI-MS and quantification by multiple linear regression.” *Microchemical Journal*, 119:30–39 (2015).
- Gertheiss, J., Hogger, S., Oberhauser, C., and Tutz, G. “Selection of ordinally scaled independent variables with applications to international classification of functioning core sets.” *Journal of the Royal Statistical Society: Series C*, 60(3):377–395 (2011).
- Glasser, M. “Exponential survival with covariance.” *Journal of the American Statistical Association*, 62(318):561–568 (1967).

- Goeman, J. J. “ L_1 penalized estimation in the Cox proportional hazards model.” *Biometrical Journal*, 52(1):70–84 (2010).
- González-Recio, O., Weigel, K. A., Gianola, D., Naya, H., and Rosa, G. J. M. “ L_2 -Boosting algorithm applied to high-dimensional problems in genomic selection.” *Genetics Research*, 92(03):227–237 (2010).
- Gray, R. J. “Some diagnostic methods for Cox regression models through hazard smoothing.” *Biometrics*, 93–102 (1990).
- Groll, A. and Tutz, G. “Regularization for generalized additive mixed models by likelihood-based boosting.” *Methods of Information in Medicine*, 51(2):168–177 (2012).
- Guisan, A., Edwards, T. C., and Hastie, T. J. “Generalized linear and generalized additive models in studies of species distributions: Setting the scene.” *Ecological Modelling*, 157(2):89–100 (2002).
- Hastie, T. J. and Tibshirani, R. J. “Generalized additive models.” *Statistical Science*, 297–310 (1986).
- . “Generalized additive models: Some applications.” *Journal of the American Statistical Association*, 82(398):371–386 (1987a).
- . “Non-parametric logistic and proportional odds regression.” *Applied Statistics*, 260–276 (1987b).
- . *Generalized additive models*. CRC Press (1990).
- He, S., Chen, H., Zhu, Z., Ward, D. G., Cooper, H. J., Viant, M. R., Heath, J. K., and Yao, X. “Robust twin boosting for feature selection from high-dimensional omics data with label noise.” *Information Sciences*, 291:1–18 (2015).

- He, X. “Quantile curves without crossing.” *The American Statistician*, 51(2):186–192 (1997).
- He, X., Fung, W. K., and Zhu, Z. “Robust estimation in generalized partial linear models for clustered data.” *Journal of the American Statistical Association*, 100(472):1176–1184 (2005).
- Hoerl, A. E. and Kennard, R. W. “Ridge regression: Biased estimation for nonorthogonal problems.” *Technometrics*, 12(1):55–67 (1970).
- Hofner, B., Boccuto, L., and Goeker, M. “Controlling false discoveries in high-dimensional situations: Boosting with stability selection.” *Technical Report, arXiv:1411.1285* (2014).
- Hofner, B., Hothorn, T., and Kneib, T. “Variable selection and model choice in structured survival models.” *Computational Statistics*, 28(3):1079–1101 (2013).
- Hofner, B., Hothorn, T., Kneib, T., and Schmid, M. “A framework for unbiased model selection based on boosting.” *Journal of Computational and Graphical Statistics*, 20(4):956–971 (2011).
- Hofner, B., Kneib, T., and Hothorn, T. “A unified framework of constrained regression.” *Statistics and Computing*, 26(1-2):1–14 (2016).
- Holm, A. M., Sivertsen, E. A., Tunheim, S. H., Haug, T., Bjerkeli, V., Yndestad, A., Aukrust, P., and Frøland, S. S. “Gene expression analysis of peripheral T cells in a subgroup of common variable immunodeficiency shows predominance of CCR7(-) effector-memory T cells.” *Clinical & Experimental Immunology*, 138(2):278–289 (2004).

- Holst, U., Hössjer, O., Björklund, C., Ragnarson, P., and Edner, H. “Locally weighted least squares kernel regression and statistical evaluation of LIDAR measurements.” *Environmetrics*, 7(4):401–416 (1996).
- Horowitz, J. L. “A smoothed maximum score estimator for the binary response model.” *Econometrica: Journal of the Econometric Society*, 505–531 (1992).
- Hosmer, J. D. W., Lemeshow, S., and Sturdivant, R. X. *Applied logistic regression*. John Wiley & Sons (2013).
- Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., and Hofner, B. “Model-based boosting 2.0.” *The Journal of Machine Learning Research*, 11:2109–2113 (2010).
- Hothorn, T., Bühlmann, S., P. and Dudoit, Molinaro, A., and van der Laan, M. J. “Survival ensembles.” *Biostatistics*, 7(3):355–373 (2006).
- Hothorn, T., Kneib, T., and Bühlmann, P. “Conditional transformation models.” *Journal of the Royal Statistical Society: Series B*, 76(1):3–27 (2014).
- Hu, Y., Zhao, K., and Lian, H. “Bayesian quantile regression for partially linear additive models.” *Statistics and Computing*, 25(3):651–668 (2015).
- Jenssen, T. K., Kuo, W., Stokke, T., and Hovig, E. “Associations between gene expressions in breast cancer and patient survival.” *Human Genetics*, 111(4-5):411–420 (2002).
- Jiang, W. “Process consistency for AdaBoost.” *Annals of Statistics*, 13–29 (2004).

- Johnson, B. A. and Long, Q. “Survival ensembles by the sum of pairwise differences with application to lung cancer microarray studies.” *The Annals of Applied Statistics*, 5(2A):1081–1101 (2011).
- Kammann, E. E. and Wand, M. P. “Geoadditive models.” *Journal of the Royal Statistical Society: Series C*, 52(1):1–18 (2003).
- Kaplan, E. L. and Meier, P. “Nonparametric estimation from incomplete observations.” *Journal of the American Statistical Association*, 53(282):457–481 (1958).
- Kauermann, G. and Khomski, P. “Additive two-way hazards model with varying coefficients.” *Computational Statistics & Data Analysis*, 51(3):1944–1956 (2006).
- Kauermann, G. and Opsomer, J. D. “Data-driven selection of the spline dimension in penalized spline regression.” *Biometrika*, 98(1):225–230 (2011).
- Kearns, M. “Thoughts on hypothesis boosting.” *Unpublished manuscript* (1988).
- Kearns, M. and Valiant, L. “Cryptographic limitations on learning Boolean formulae and finite automata.” In *Proceedings of the Twenty-first Annual ACM Symposium on Theory of Computing* (1989).
- Khan, U. T., Valeo, C., and He, J. “Non-linear fuzzy-set based uncertainty propagation for improved DO prediction using multiple-linear regression.” *Stochastic Environmental Research and Risk Assessment*, 27(3):599–616 (2013).

- Klein, J. P. and Moeschberger, M. L. *Survival analysis: Techniques for censored and truncated data*. Springer Science & Business Media (2005).
- Kleinbaum, D. and Klein, M. *Logistic regression: A self-learning text*. Springer Science & Business Media (2010).
- Kneib, T. “Beyond mean regression.” *Statistical Modelling*, 13(4):275–303 (2013).
- Kneib, T., Hothorn, T., and Tutz, G. “Variable selection and model choice in geoaddditive regression models.” *Biometrics*, 65(2):626–634 (2009).
- Koenker, R. “Quantile regression for longitudinal data.” *Journal of Multivariate Analysis*, 91(1):74–89 (2004).
- . *Quantile regression*. Cambridge University Press (2005).
- Koenker, R. and Machado, J. A. F. “Goodness of fit and related inference processes for quantile regression.” *Journal of the American Statistical Association*, 94(448):1296–1310 (1999).
- Kohavi, R. “A study of cross-validation and bootstrap for accuracy estimation and model selection.” In *Ijcai*, volume 14, 1137–1145 (1995).
- Kordas, G. “Smoothed binary regression quantiles.” *Journal of Applied Econometrics*, 21(3):387–407 (2006).
- Kriegler, B. and Berk, R. “Boosting the quantile distribution: A cost-sensitive statistical learning procedure.” *Department of Statistics, UCLA, working paper* (2007).

- Lambert, P. and Eilers, P. H. C. “Bayesian proportional hazards model with time-varying regression coefficients: A penalized Poisson regression approach.” *Statistics in Medicine*, 24(24):3977–3989 (2005).
- Lee, B. P. L., Rushlow, W. J., Chakraborty, C., and Lala, P. K. “Differential gene expression in premalignant human trophoblast: Role of IGFBP-5.” *International Journal of Cancer*, 94(5):674–684 (2001).
- Leitenstorfer, F. and Tutz, G. “Generalized monotonic regression based on B-splines with an application to air pollution data.” *Biostatistics*, 8(3):654–673 (2007).
- Leng, C. and Zhang, H. H. “Nonparametric model selection in hazard regression.” *Journal of Nonparametric Statistics*, 18:417–429 (2007).
- Li, H. and Luan, Y. “Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data.” *Bioinformatics*, 21(10):2403–2409 (2005).
- Lian, H., Chen, X., and Yang, J. Y. “Identification of partially linear structure in additive models with an application to gene expression prediction from sequences.” *Biometrics*, 68(2):437–445 (2012).
- Lian, H., Li, J., and Hu, Y. “Shrinkage variable selection and estimation in proportional hazards models with additive structure and high dimensionality.” *Computational Statistics & Data Analysis*, 63:99–112 (2013).
- Liang, H. “Estimation in partially linear models and numerical comparisons.” *Computational Statistics & Data Analysis*, 50(3):675–687 (2006).
- Liang, H., Thurston, S. W., Ruppert, D., Apanasovich, T., and Hauser, R.

- “Additive partial linear models with measurement errors.” *Biometrika*, 95(3):667–678 (2008).
- Lin, D. Y. “On the Breslow estimator.” *Lifetime Data Analysis*, 13(4):471–480 (2007).
- Lin, D. Y. and Wei, L. J. “The robust inference for the Cox proportional hazards model.” *Journal of the American Statistical Association*, 84(408):1074–1078 (1989).
- Lin, D. Y. and Ying, Z. “Semiparametric inference for the accelerated life model with time-dependent covariates.” *Journal of Statistical Planning and Inference*, 44(1):47–63 (1995).
- Lindsey, J. C. and Ryan, L. M. “Methods for interval-censored data.” *Statistics in Medicine*, 17(2):219–238 (1998).
- Liu, D., Lin, X., and Ghosh, D. “Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models.” *Biometrics*, 63(4):1079–1088 (2007).
- Liu, H., Li, J., and Wong, L. “Use of extreme patient samples for outcome prediction from gene expression data.” *Bioinformatics*, 21(16):3377–3384 (2005).
- Lu, W. and Li, L. “Boosting method for nonlinear transformation models with censored survival data.” *Biostatistics*, 9(4):658–667 (2008).
- Lu, X., Chen, G., Singh, R. S., Song, K., and Peter, X. “A class of partially linear single-index survival models.” *Canadian Journal of Statistics*, 34(1):97–112 (2006).

- Ma, S., Huang, J., Xie, Y., and Yi, N. “Identification of breast cancer prognosis markers using integrative sparse boosting.” *Methods of Information in Medicine*, 51(2):152–161 (2012).
- Machado, J. A. F. and Silva, J. M. C. S. “Quantiles for counts.” *Journal of the American Statistical Association*, 100(472):1226–1237 (2005).
- Mallows, C. L. “Some comments on C_p .” *Technometrics*, 15(4):661–675 (1973).
- Manski, C. F. “Maximum score estimation of the stochastic utility model of choice.” *Journal of Econometrics*, 3(3):205–228 (1975).
- . “Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator.” *Journal of Econometrics*, 27(3):313–333 (1985).
- Mason, L., Baxter, J., Bartlett, P., and Frean, M. “Boosting algorithms as gradient descent in function space.” NIPS (1999).
- Mayr, A., Binder, H., Gefeller, O., and Schmid, M. “Extending statistical boosting.” *Methods Inf Med*, 53(6):428–435 (2014).
- Mayr, A., Fenske, N., Hofner, B., Kneib, T., and Schmid, M. “Generalized additive models for location, scale and shape for high dimensional data: A flexible approach based on boosting.” *Journal of the Royal Statistical Society: Series C*, 61(3):403–427 (2012a).
- Mayr, A., Hothorn, T., and Fenske, N. “Prediction intervals for future BMI values of individual children - A non-parametric approach by quantile boosting.” *BMC Medical Research Methodology*, 12(1):1–13 (2012b).

- Mayr, A. and Schmid, M. “Boosting the concordance index for survival data - A unified framework to derive and evaluate biomarker combinations.” *PloS one*, 9(1):e84483 (2014).
- Meinshausen, N. and Bühlmann, P. “Stability selection.” *Journal of the Royal Statistical Society: Series B*, 72(4):417–473 (2010).
- Mittlböck, M. and Heinzl, H. “Measures of explained variation in gamma regression models.” *Communications in Statistics-Simulation and Computation*, 31(1):61–73 (2002).
- Murphy, S. A. “Asymptotic theory for the frailty model.” *The Annals of Statistics*, 182–198 (1995).
- Nelson, W. “Hazard plotting for incomplete failure data.” *Journal of Quality Technology*, 1(1):27–52 (1969).
- Newey, W. K. and Powell, J. L. “Asymmetric least squares estimation and testing.” *Econometrica: Journal of the Econometric Society*, 819–847 (1987).
- Oakes, D. “Survival times: Aspects of partial likelihood.” *International Statistical Review*, 235–252 (1981).
- Peto, R. “Experimental survival curves for interval-censored data.” *Applied Statistics*, 86–91 (1973).
- Ridgeway, G. “The state of boosting.” *Computing Science and Statistics* (1999).
- Rigby, R. A. and Stasinopoulos, D. M. “Generalized additive models for location, scale and shape.” *Journal of the Royal Statistical Society: Series C*, 54(3):507–554 (2005).

- Robinzonov, N. and Hothorn, T. “Boosting for estimating spatially structured additive models.” In *Statistical Modelling and Regression Structures*, 181–196. Springer (2010).
- Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., and Giltane, J. M. “The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma.” *New England Journal of Medicine*, 346(25):1937–1947 (2002).
- Ruppert, D. “Selecting the number of knots for penalized splines.” *Journal of Computational and Graphical Statistics*, 735–757 (2012).
- Ruppert, D. and Carroll, R. J. “Theory & methods: Spatially-adaptive penalties for spline fitting.” *Australian & New Zealand Journal of Statistics*, 42(2):205–223 (2000).
- Ruppert, D., Wand, M. P., and Carroll, R. J. *Semiparametric regression*. Cambridge University Press (2003).
- . “Semiparametric regression during 2003–2007.” *Electronic Journal of Statistics*, 3:1193–1256 (2009).
- Sastry, N. “A nested frailty model for survival data, with an application to the study of child survival in northeast Brazil.” *Journal of the American Statistical Association*, 92(438):426–435 (1997).
- Schapire, R. E. “The strength of weak learnability.” *Machine Learning*, 5(2):197–227 (1990).

- Scheipl, F., Fahrmeir, L., and Kneib, T. “Spike-and-slab priors for function selection in structured additive regression models.” *Journal of the American Statistical Association*, 107(500):1518–1532 (2012).
- Scheipl, F., Kneib, T., and Fahrmeir, L. “Penalized likelihood and bayesian function selection in regression models.” *JASStA Advances in Statistical Analysis*, 97(4):349–385 (2013).
- Schmid, M. and Hothorn, T. “Flexible boosting of accelerated failure time models.” *BMC Bioinformatics*, 9(1):269–281 (2008).
- Schnabel, S. K. and Eilers, P. H. C. “Simultaneous estimation of quantile curves using quantile sheets.” *AStA Advances in Statistical Analysis*, 97(1):77–87 (2013).
- Schoenfeld, D. A. “Sample-size formula for the proportional-hazards regression model.” *Biometrics*, 499–503 (1983).
- Shah, R. D. and Samworth, R. J. “Variable selection with error control: Another look at stability selection.” *Journal of the Royal Statistical Society: Series B*, 75(1):55–80 (2013).
- Shim, J., Hwang, C., and Seok, K. H. “Non-crossing quantile regression via doubly penalized kernel machine.” *Computational Statistics*, 24(1):83–94 (2009).
- Simonoff, J. S. and Tsai, C. “Semiparametric and additive model selection using an improved Akaike information criterion.” *Journal of Computational and Graphical Statistics*, 8(1):22–40 (1999).

- Snyman, J. *Practical mathematical optimization: An introduction to basic optimization theory and classical and new gradient-based algorithms*, volume 97. Springer Science & Business Media (2005).
- Sobotka, F. and Kneib, T. “Geoaddivitive expectile regression.” *Computational Statistics & Data Analysis*, 56(4):755–767 (2012).
- Speckman, P. “Kernel smoothing in partial linear models.” *Journal of the Royal Statistical Society: Series B*, 413–436 (1988).
- Steyerberg, E. W., Eijkemans, M. J. C., and Habbema, J. D. F. “Step-wise selection in small data sets: A simulation study of bias in logistic regression analysis.” *Journal of Clinical Epidemiology*, 52(10):935–942 (1999).
- Stone, M. “Cross-validatory choice and assessment of statistical predictions.” *Journal of the Royal Statistical Society: Series B*, 111–147 (1974).
- . “An asymptotic equivalence of choice of model by cross-validation and Akaike’s criterion.” *Journal of the Royal Statistical Society: Series B*, 44–47 (1977).
- Taieb, S. B., Huser, R., Hyndman, R. J., and Genton, M. G. “Probabilistic time series forecasting with boosted additive models: An application to smart meter data.” Technical report, Monash University, Department of Econometrics and Business Statistics (2015).
- Takeuchi, I., Le, Q. V., Sears, T. D., and Smola, A. J. “Nonparametric quantile estimation.” *The Journal of Machine Learning Research*, 7:1231–1264 (2006).

- Therneau, T. M. and Grambsch, P. M. *Modeling survival data: Extending the Cox model*. Springer Science & Business Media (2000).
- Tibshirani, R. “Regression shrinkage and selection via the lasso.” *Journal of the Royal Statistical Society: Series B*, 58(1):267–288 (1996).
- Tsiatis, A. A. and Davidian, M. “A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error.” *Biometrika*, 88(2):447–458 (2001).
- Tutz, G. and Binder, H. “Flexible modelling of discrete failure time including time-varying smooth effects.” *Statistics in Medicine*, 23(15):2445–2461 (2004).
- . “Generalized additive modeling with implicit variable selection by likelihood-based boosting.” *Biometrics*, 62(4):961–971 (2006).
- Tutz, G. and Gertheiss, J. “Feature extraction in signal regression: A boosting technique for functional data regression.” *Journal of Computational and Graphical Statistics*, 19(1):154–174 (2010).
- Tutz, G. and Reithinger, F. “A boosting approach to flexible semiparametric mixed models.” *Statistics in Medicine*, 26(14):2872–2900 (2007).
- Tutz, G. and Ulbricht, J. “Penalized regression with correlation-based penalty.” *Statistics and Computing*, 19(3):239–253 (2009).
- van der Laan, M. J. and Robins, J. M. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media (2003).
- van Houwelingen, H. C., Bruinsma, T., Hart, A. A. M., van’t Veer, L. J., and Wessels, L. F. A. “Cross-validated Cox regression on microarray gene expression data.” *Statistics in Medicine*, 25(18):3201–3216 (2006).

- Wagner, J. M. and Shimshak, D. G. “Stepwise selection of variables in data envelopment analysis: Procedures and managerial perspectives.” *European Journal of Operational Research*, 180(1):57–67 (2007).
- Wahba, G. *Spline models for observational data*. SIAM (1990).
- Wang, H. J., Zhu, Z., and Zhou, J. “Quantile regression in partially linear varying coefficient models.” *The Annals of Statistics*, 3841–3866 (2009).
- Wang, J., Coombes, K. R., Highsmith, W. E., Keating, M. J., and Abruzzo, L. V. “Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: A meta-analysis of three microarray studies.” *Bioinformatics*, 20(17):3166–3178 (2004).
- Wang, Z. “HingeBoost: ROC-based boost for classification and variable selection.” *The International Journal of Biostatistics*, 7(1):1–30 (2011).
- Wang, Z. and Wang, C. Y. “Buckley-James boosting for survival analysis with high-dimensional biomarker data.” *Statistical Applications in Genetics and Molecular Biology*, 9(1):1–33 (2010).
- Wood, S. *Generalized additive models: An introduction with R*. CRC press (2006).
- Wu, N., Huang, J., Schmalz, B., and Fohrer, N. “Modeling daily chlorophyll a dynamics in a German lowland river using artificial neural networks and multiple linear regression approaches.” *Limnology*, 15(1):47–56 (2014).
- Xia, Y. and Härdle, W. “Semi-parametric estimation of partially linear single-index models.” *Journal of Multivariate Analysis*, 97(5):1162–1184 (2006).

- Xiao, Z. “Quantile cointegrating regression.” *Journal of Econometrics*, 150(2):248–260 (2009).
- Yatchew, A. “An elementary estimator of the partial linear model.” *Economics Letters*, 57(2):135–143 (1997).
- Yee, T. W. “Quantile regression via vector generalized additive models.” *Statistics in Medicine*, 23(14):2295–2315 (2004).
- Yee, T. W. and Mitchell, N. D. “Generalized additive models in plant ecology.” *Journal of Vegetation Science*, 2(5):587–602 (1991).
- Yeh, I. C. “Modeling of strength of high-performance concrete using artificial neural networks.” *Cement and Concrete Research*, 28(12):1797–1808 (1998).
- Zhang, H. H., Cheng, G., and Liu, Y. “Linear or nonlinear? Automatic structure discovery for partially linear models.” *Journal of the American Statistical Association*, 106(495):1099–1112 (2011).
- Zhang, H. H. and Lu, W. “Adaptive LASSO for Cox’s proportional hazards model.” *Biometrika*, 94(3):691–703 (2007).
- Zhang, T. and Yu, B. “Boosting with early stopping: Convergence and consistency.” *Annals of Statistics*, 1538–1579 (2005).
- Zhao, J. “General sparse boosting: Improving feature selection of L_2 boosting by correlation-based penalty family.” *Communications in Statistics-Simulation and Computation*, 44(6):1612–1640 (2015).
- Zhao, W. and Elsayed, E. A. “A general accelerated life model for step-stress testing.” *IIE Transactions*, 37(11):1059–1069 (2005).

- Zheng, S. “Gradient descent algorithms for quantile regression with smooth approximation.” *International Journal of Machine Learning and Cybernetics*, 2(3):191–207 (2011).
- . “QBoost: Predicting quantiles with boosting for regression and binary classification.” *Expert Systems with Applications*, 39(2):1687–1697 (2012).
- Zhou, S. and Shen, X. “Spatially adaptive regression splines and accurate knot selection schemes.” *Journal of the American Statistical Association*, 96(453):247–259 (2001).
- Zou, H. “The adaptive LASSO and its oracle properties.” *Journal of the American Statistical Association*, 101(476):1418–1429 (2006).
- Zou, H. and Hastie, T. “Regularization and variable selection via the elastic net.” *Journal of the Royal Statistical Society: Series B*, 67(2):301–320 (2005).
- Zucker, D. M. and Karr, A. F. “Nonparametric survival analysis with time-dependent covariate effects: A penalized partial likelihood approach.” *The Annals of Statistics*, 329–353 (1990).

List of Publications During Ph.D.

Studies in MAS, SPMS, NTU

1. Tang, X., Li, J., and Lian, H. (2013). Empirical likelihood for partially linear proportional hazards models with growing dimensions. *Journal of Multivariate Analysis*, 121, 22-32.
2. Lian, H., Li, J., and Tang, X. (2014). SCAD-penalized regression in additive partially linear proportional hazards models with an ultra-high-dimensional linear part. *Journal of Multivariate Analysis*, 125, 50-64.
3. Tang, X., and Lian, H. (2015). Mean and quantile boosting for partially linear additive models. *Statistics and Computing*, 1-12.
4. Tang, X., Hong, Z., Hu, Y., and Lian, H. (2015). Gaussian process models for non-parametric functional regression with functional responses. *Communications in Statistics-Theory and Methods*, 44(16), 3428-3445.