Shape-Constrained Estimation in High Dimensions

Min Xu

June 2015
CMU-ML-15-103

Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
John Lafferty, Chair
Aarti Singh
Larry Wasserman
Ming Yuan (UW Madison)

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

This research was funded in parts by the grants NSF IIS1116740, ONR N000141210762, NSF CCF0625879, and AFOSR FA95500910373

The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

In memory of my grandfather.

Abstract

Shape-constrained estimation techniques such as convex regression or log-concave density estimation offer attractive alternatives to traditional nonparametric methods. Shape-constrained estimation often has an easy-to-optimize likelihood, no tuning parameter, and an adaptivity property where the sample complexity adapts to the complexity of the underlying functions. In this dissertation, we posit that shape-constrained estimation has an additional advantage in that they are naturally suited to the high-dimensional regime, where the number of variables is large relative to the number of samples.

In the first part of this dissertation, we study high dimensional convex regression and demonstrate that convex regression surprisingly has the additive faithfulness property, where the additive approximation is guaranteed to capture all relevant variables even if the underlying function is not additive. We propose a practical variable selection procedure for high dimensional convex regression based on this observation. The overall work provides a practical smoothing-free semi-parametric generalization of the Lasso.

We generalize our work on high dimensional convex regression to discrete choice models, in which a consumer chooses between m items $x_1, ..., x_m$ with probability proportional to $\exp f(x_i)$ for a utility function f. We show that additive faithfulness applies also in this setting. We accordingly adapt our method to the estimation of the utility function.

In the last part, we consider the problem of learning the orientation pattern in additive shape-constraint models. Brute force search in this problem requires times exponential in the dimensionality. We propose a relaxation approach, based on trend filtering and motivated by our identifiability analysis, that is computationally efficient and effective.

Acknowledgments

Happy graduate students come from good advisors and the happiest graduate students come from John Lafferty. To me, John made the research inspirational, the work enjoyable, the frustration tolerable, and the journey memorable. I would not have lasted long in the PhD program without John's boundless support and endless encouragement.

Aarti Singh and Larry Wasserman were an integral part of my graduate student experience–it is a rare pleasure to be able to do research and teach courses with mentors friendly and passionate about the pursuit of knowledge and understanding. I am fortunate to have Ming Yuan with his sharp insights on my thesis committee. I also thank Carlos Guestrin, Emily Fox, Shuheng Zhou, Peter Buhlmann, Ryan Tibshirani, Rina Foygel Barber, and Anupam Gupta for having taught me much.

Tie-Yan Liu and Tao Qin gave me a fun and fruitful summer at Microsoft Research Asia in 2012 and Rayid Ghani provided me another great summer at the Data Science for Social Good Program in 2013. I am indebted to them for the memorable experiences.

Though this thesis is mine, the ideas and work within come from many more. I am thankful to my collaborators: the inimitable Sivaraman Balakrishnan and Akshay Krishnamurthy, the wiser older brother Khalid El-Arini, Han Liu, Sabyasachi Chatterjee, Minhua Chen, Yuxue Qi.

Graduate school was a wild ride and I have had the good luck of sharing it with fellows students who made life sometimes easier, sometimes harder, but never dull.

Madalina Fiterau, Leila Wehbe, Ankur Parikh, Aaditya Ramdas were my trench buddies almost from day one. Yang Xu, Jing Xiang, Yisong Yue were great to talk to and learn from. Qirong Ho, Yucheng Low, Joseph Gonzalez, Alona Fyshe, Prashant Reddy, Diyi Yang, Nan Li, Favonia, Julian Shun, Suresh, Yifei Ma, Brendan O'Connor, Rob Hall, James Sharpnack, Mladen Kolar, Han Liu, Haijie Gu from CMU have all colored my life. Likewise with Yuancheng Zhu, Walter Dempsey, Dinah Shender, Qinqing Zheng, Sun Siqi and Meng-wen Zhang from Chicago. Helen Li, as loving as she is lovely.

Lastly, I owe too much to my ever loving, ever caring parents.

Contents

List of Figures

CHAPTER 1

INTRODUCTION

Nonparametric estimation methods, such as kernel regression or random forest, are flexible and powerful because of they impose weak assumptions on the underlying function. Their disadvantages are that they require more time for computate and more samples for estimate. Nonparametric methods are particularly vulnerable to the curse of dimensionality. Their drawbacks are dramatically exacerbated when the data is high-dimensional, i.e. when the dataset has a large number of variables relative to the number of samples.

In parametric regression, stunning recent advances have shown that under a sparsity assumption, in which most variables are assumed to be uninformative, it is tractable to identify the relevant variables and estimate the function as if the data is low-dimensional. Some analogous results have followed for high-dimensional nonparametric regression but there is still a large gap; there currently exist no method for high-dimensional nonparametric regression that is as practical and theoretically justifiable as parametric methods like the Lasso.

This thesis tackles the problem of high-dimensional nonparametric estimation through shape constraints. Shape-constrained estimation has a rich history and extensive research on topics such as convex or monotone regression and log-concave density estimation. Shape-constraints differ from the usual smoothness assumptions in several ways:

1. It is often possible to directly optimize the likelihood.
2. It is often free of tuning parameters, such as the bandwidth in kernel regression.
3. It exhibits adaptivity; the sample complexity can adapt to the complexity of the underlying function to be learned. (Guntuboyina and Sen, 2013a; Cai and Low, 2011)

In this thesis, we posit an additional advantage: that shape constraints are naturally suited toward high-dimensional estimation.

We focus on monotone functions and convex/concave functions in this thesis, but some of the analysis extends to higher orders of shape-constraints.

Shape-constraint assumptions arise naturally from real data. For example, the income of a person is a increasing function of the education quality, the price of a

house is a decreasing function of the neighborhood crime level. Estimation of convex functions arises naturally in several applications. Examples include geometric programming (Boyd and Vandenberghe, 2004), computed tomography (Prince and Willsky, 1990), target reconstruction (Lele et al., 1992), image analysis (Goldenshluger and Zeevi, 2006) and circuit design (Hannah and Dunson, 2012). Other applications include queuing theory (Chen and Yao, 2001) and economics, where it is of interest to estimate concave utility functions (Meyer and Pratt, 1968). Utility functions can be assumed concave because of the phenonmenon of diminishing returns. Beyond cases where the shape-constraint assumption is natural, the shape-constrained estimation can be attractive as a tractable, nonparamametric relaxation of the linear model.

Shape-constrained estimation has a long history. Much of the earlier work on isotonic regression is described by the classic "4B" text (Barlow et al., 1972). Mair et al. (2009) too provides a good history. Research into convex regression began in the 1950s (Hildreth, 1954) for estimation of production and Engel curves. The earlier works were focused on the univariate case and the least square estimator's properties were investigated by Hanson and Pledger (1976), Groeneboom et al. (2001), Mammen (1991), and more.

Recently, there has been increased research activity on shape-constrained estimation. Guntuboyina and Sen (2013b) analyze univariate convex regression and show surprisingly that the risk of the MLE is adaptive to the complexity of the true function. Seijo and Sen (2011) and Lim and Glynn (2012) study maximum likelihood estimation of multivariate convex regression and independently establish its consistency. Cule et al. (2010c) and Kim and Samworth (2014) analyze log-concave density estimation and prove consistency of the MLE; the latter further show that log-concave density estimation has minimax risk lower bounded by $n^{-2/(d+1)}$ for $d \geq 2$, refuting a common notion that the condition of convexity is equivalent, in estimation difficulty, to the condition of having two bounded derivatives. Additive shape-constrained estimation has also been studied; Pya and Wood (2014) propose a penalized B-spline estimator while Chen and Samworth (2014) show the consistency of the MLE.

1.1 Thesis Summary

Briefly, we study high dimensional convex regression in Chapter 2, convex utility estimation for discrete choice model in Chapter 3, and shape-constraint pattern selection for additive models in Chapter 4. Lastly, in Chapter 5, we discuss open questions raised by the work in this thesis.

Chapter 2

In Chapter 2 of the thesis, we give a practical procedure that performs variable selection for high dimensional convex regression. Our main result is a population level analysis showing that an additive projection can faithfully recover the set of relevant variables of a possibly non-additive convex function under weak assumptions on the underlying density; we refer to this phenonmenon as additive faithfulness.

Our estimation procedure is a two stage procedure where we fit an additive convex function in the first stage and fit several decoupled univariate concave functions in the second stage. The second concave fitting stage is un-intuitive and generally necessary as shown in our theory. Our optimization method is a backfitting procedure where each iteration is a quadratic program. It is computationally practical and effective on both simulated and real data.

We also perform a finite sample analysis on our estimation procedure and prove variable screening consistency. Whereas variable selection for general smooth functions is impossible unless $n = \exp(s)$ (as proved by Comminges and Dalalyan (2012)), we show that variable selection for convex regression is consistent even if $n = O(\text{poly}(s))$ where n is the sample size and s is the number of relevant variables.

Chapter 3

Chapter 3 generalizes the work of Chapter 2 to the discrete choice model, which is a more general form of the logistic loss. In discrete choice model, a consumer chooses one of m items $\mathbf{x}_1, ..., \mathbf{x}_m$ to purchase and decides on item $\mathbf{x}_i$ with probability proportional to $\exp f(\mathbf{x}_i)$ where f is a concave utility function.

We show that a form of additive faithfulness also holds in this setting and extend our estimation procedure to the discrete choice model. We verify that our method is effective in a real world dataset from a survey that we designed and conducted.

Chapter 4

Chapter 4 studies the problem of orientation pattern selection for an additive shape-constraint model. More precisely, we fit $\sum_{j=1}^{d} f_j$ where f_j could be either monotone increasing or decreasing.

We show in the $d = 2$ case that the problem is identifiable and that the correct pattern can be recovered by minimizing the L_1 norm of the differences $\sum_{i=1}^{n-1} |f_{ij} - f_{i+1,j}|$. This observation motivates an estimator where we use a L_1 of differences regularization. This estimator can also be interpreted as the convex relaxation of the computationally inefficient combinatorial search.

1.2 Properties of Shape-Constrained Functions

We give some basic properties of monotone and convex/concave functions and then describe simple estimation problems that involve these functions.

Monotonic Functions

A monotonic function from $\mathbb{R}$ to $\mathbb{R}$ is easy to visualize, but the notion of monotoncity can actually be much more general.

Definition 1.2.1. Let C be a partially ordered set. A function $f : C \to \mathbb{R}$ is *monotone increasing* if $f(x) \geq f(y)$ if $x \succeq y$.

If $C = \mathbb{R}^p$, we can use the ordering that $x \geq y$ if $x_j \geq y_j$ for all j. f is monotone under this ordering if and only if, for all $j = 1, ..., p$, for any fixed $\mathbf{x}_{-j}$, $f(x_j, \mathbf{x}_{-j})$ is a monotonic 1-dimensional function of x_j. Another interesting example is if C is a directed acyclic graph.

Classic results in real analysis state that monotone functions have a countable number of discontinuities and that every function of bounded variation can be written as a sum of a monotone increasing and decreasing function.

Given finite samples, the MLE (assuming Gaussian error) of monotone regression is a finite dimensional optimization even though the set of monotone functions is infinite dimensional.

Definition 1.2.2. (LSE for Monotone Functions)
Suppose we have samples $(X_i, y_i)_{i=1,...,n}$ where the X_i's are drawn from some distribution on C. The least square estimator (LSE) is

$$\begin{aligned} &\min_{f_i} \sum_{i=1}^{n} (f_i - y_i)^2 \\ &\text{s.t. } f_i \geq f_j \text{ for all } (i,j) \text{ such that } X_i \succeq X_j \end{aligned}$$

The well known Pool Adjacent Violator algorithm (PAVA), first described by Ayer et al. (1955) can efficiently solve this optimization, in time $O(n \log n)$ for totally ordered X_i's. Dykstra (1981) provides generalizations of PAVA that apply to partially ordered sets.

The estimated function is defined only on X_i in the training set but interpolation can be used to evaluate the estimated function on a general $\mathbf{x}$.

Convex Functions

Convex functions intuitively have a bowl-shaped graph. They have numerous equivalent characterizations. We list three which are useful for us.

Definition 1.2.3. Let $C \subset \mathbb{R}^d$ be a convex set. $f : C \to \mathbb{R}$ is convex if $f(\lambda x + 1(1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$ for all $\lambda \in [0,1]$ and for all $x, y \in C$.

Equivalently, f is convex iff for every $x \in C$, there exists a subgradient $\nabla f(x) \in \mathbb{R}^d$ such that $f(y) \geq f(x) + \nabla f(x)^\mathsf{T}(y-x)$.

If f is twice-differentiable, then f is convex iff the Hessian is positive semidefinite for all x in the interior of C.

The first order characterization for one dimensional convex functions is particularly simple; it says that the derivative must be non-decreasing. This simple observation is useful in reducing the computational complexity of many of our estimation algorithms.

Convex functions are analytically nice because they are continuous on the interior of the support and thus measurable. In fact, a classic result by Aleksandorov (1939) shows that convex functions are almost everywhere twice differentiable.

A useful well-known property of convex functions is that the sum of a convex and a concave function can represent any function with a bounded second derivative (Yuille and Rangarajan, 2003). This is analogous to how any function of bounded variation can be written as the sum of an increasing and a decreasing function.

Proposition 1.2.1. *Let $h : \mathbb{R}^d \to \mathbb{R}$ be a twice-differentiable function with a bounded second derivative. Then $h = f + g$ where f is convex and g is concave.*

Proof. The Hessian of $h(x) + c\sum_{j=1}^d x_j^2$ is Hessian$(h) + cI_d$. Since the second derivative of h is bounded, there exists a positive scalar c such that $f(x) = h(x) + c\sum_{j=1}^d x_j^2$ is convex. Let $g(x) = -c\sum_{j=1}^d x_j^2$ and the claim follows. □

Analogous to the case of monotone regression, the MLE for convex regression is a finite dimensional optimization.

$$\min_{f_i, \beta_i} \frac{1}{n} \sum_{i=1}^n (y_i - f_i)^2$$
$$\text{s.t.} f_{i'} \geq f_i + \beta_i^\mathsf{T}(x_{i'} - x_i)$$

where β_i is the p-dimensional subgradient vector at x_i. This optimization is a Quadratic Program and can be solved efficiently with interior point methods. Again,

the estimated function is defined only on X_i in the training set but we can evaluate the estimated function on a general $\mathbf{x}$ with interpolation.

1.3 Background on High Dimensional Statistics

High dimensional data is, simply put, data with a large number of covariates–often more than the number of samples. Statistical problems become challenging in this regime and many classical methods fail entirely.

Let us first consider a linear model $y = X\beta$ with n samples and p covariates. When $p << n$, the ordinary least square estimate $\widehat{\beta}_{OLS} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}y$ forms a good estimate of y. In the high dimensional regime where p is large, $\widehat{\beta}_{OLS}$ overfits; its training error decreases as one uses more covariates. If $p > n$, $\widehat{\beta}$ cannot even be computed since $X^\mathsf{T}X$ is non-invertible.

For high dimensional data, it is reasonable to assume that most of the covariates are not predictive of the output and thus irrelevant. We formalize this assumption mathematically by assuming that β is *sparse*–it has only s non-zero entries where $s << p$. In the vocabulary of high-dimensional statistics, we say that p is the *ambient dimension*.

If the *relevant variables* were known a priori, then the problem is easy because we can ignore the irrelevant variables and put the problem in the low dimensional regime; the challenge thus is *variable selection*–to identify the set S of relevant variables. One approach is to search over the subsets–possibly using a greedy method for computational efficiency–and score them with a criterion such as Mallows's C_p, Akaike Information Criterion, or Bayesian Information Criterion, to achieve a good balance of both low training error and low model complexity (Hastie et al., 2009).

Another approach shows that the L_1-regularized M-estimator, also known as *lasso*, is effective at producing a sparse estimator $\widehat{\beta}$ whose non-zero entries approximate S. Astonishingly, it is also known that the lasso can consistently estimate the parameters so long as $\frac{s \log p}{n} \to 0$. In other words, the ambient dimension can be exponential in the number of samples.

$$\widehat{\beta}_{lasso} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\mathsf{T}\beta)^2 + \lambda \|\beta\|_1$$

Figure 1.1: The lasso estimator. λ is a tuning parameter that balances training error and sparsity of the output $\widehat{\beta}_{lasso}$.

It is harder to adapt nonparametric models to the high dimensional setting. One relatively easy case is the *additive model*, where the p-dimensional regression function $f(\mathbf{x})$ is assumed to decompose as a sum of p univariate functions $\sum_{j=1}^{p} f_j(x_j)$. In this case, many researchers were able to derive nonparametric analogues of the lasso. Ravikumar et al. (2009) for examples penalizes a sum of L_2 norms of the component functions:

$$\min_{f \text{ smooth}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} f_j(x_{ij}) \right)^2 + \lambda \sum_{j=1}^{p} \sqrt{\frac{1}{n} \sum_{i=1}^{n} f_j(x_{ij})^2}$$

In non-additive nonparametric regression, variable selection is a notoriously difficult problem. Lafferty and Wasserman (2008) develop a greedy procedure for adjusting bandwidths in a local linear regression estimator, and show that the procedure achieves the minimax rate as if the relevant variables were isolated in advance. But the method only provably scales to dimensions p that grow logarithmically in the sample size n, i.e., $p = O(\log n)$. This is in contrast to the high dimensional scaling behavior known to hold for sparsity selection in linear models using ℓ_1 penalization, where n is logarithmic in the dimension p. Bertin and Lecué (2008) develop an optimization-based approach in the nonparametric setting, applying the lasso in a local linear model at each test point. Here again, however, the method only scales as $p = O(\log n)$, the low-dimensional regime. An approximation theory approach to the same problem is presented in DeVore et al. (2011), using techniques based on hierarchical hashing schemes, similar to those used for "junta" problems (Mossel et al., 2004). Here it is shown that the sample complexity scales as $n > \log p$ if one adaptively selects the points on which the high-dimensional function is evaluated.

Comminges and Dalalyan (2012) show that the exponential scaling $n = O(\log p)$ is achievable if the underlying function is assumed to be smooth with respect to a Fourier basis. They also give support for the intrinsic difficulty of variable selection in nonparametric regression, giving lower bounds showing that consistent variable selection is not possible if $n < \log p$ or if $n < \exp s$, where s is the number of relevant variables. Variable selection over kernel classes is studied by Koltchinskii and Yuan (2010).

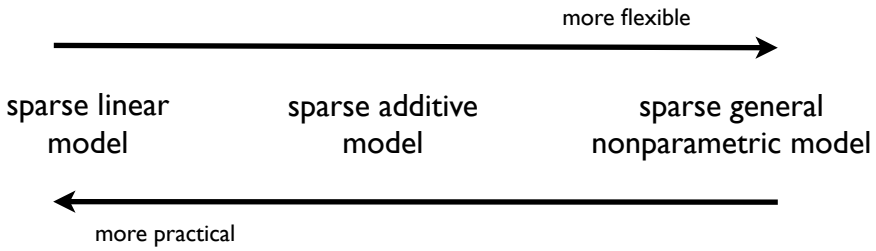


Figure 1.2: A summary of the sparse models.

1.4 Notation

Indexing Convention

Unless otherwise stated, we let i index samples, j, k index features. The variable p, d will be used to represent the dimensionality and n the number of samples.

$\mathbf{x}$	vector	X	random variable (possibly random vector)
$\mathbf{x}_{-k}$	vector $\mathbf{x}$ with k-th coordinate removed	X_S	$S \subset \{1, ..., p\}$, X restricted to variables in S
$X^{(i)}$	i-th sample	$\bar{X}$	sample mean
$\mathbb{E}[\cdot \mid x_k]$	shorthand for $\mathbb{E}[\cdot \mid X_k = x_k]$	$x_{(j)}$	the j-th largest entry of a vector $\mathbf{x}$
L^2	Lebesgue square integrable space	$L^2(P)$	square integrable space w.r.t. distribution P
$\mathbf{1}_n$	all ones vector	$\mathbf{1}_S$	vector 1 in set S, 0 else
$\|f\|_P^2$	$L_2(P)$: $\mathbb{E}f(X)^2$	$\langle f, g \rangle_n$	empirical inner product $\frac{1}{n}\sum_{i=1}^n f(X_i)g(X_i)$
$\|f\|_n^2$	empirical L2: $\frac{1}{n}\sum_{i=1}^n f(X_i)^2$	$\mathcal{C}^1$	set of univariate convex functions
$\mathcal{C}_B^1$	univariate convex function bounded by B	$\mathcal{C}_1^p$	additive function with p univariate convex functions

Figure 1.3: Notations used in the dissertation

CHAPTER 2

HIGH DIMENSIONAL CONVEX REGRESSION

2.1 Introduction

In this chapter we study the problem of variable selection in multivariate convex regression. Assuming that the regression function is convex and sparse, our goal is to identify the relevant variables. We show that it suffices to estimate a sum of one-dimensional convex functions, leading to significant computational and statistical advantages. This is in contrast to general nonparametric regression, where fitting an additive model can result in false negatives. Our approach is based on a two-stage quadratic programming procedure. In the first stage, we fit an convex additive model, imposing a sparsity penalty. In the second stage, we fit a concave function on the residual for each variable. As we show, this non-intuitive second stage is in general necessary. Our first result is that this procedure is faithful in the population setting, meaning that it results in no false negatives, under mild assumptions on the density of the covariates. Our second result is a finite sample statistical analysis of the procedure, where we upper bound the statistical rate of variable screening consistency. An additional contribution is to show how the required quadratic programs can be formulated to be more scalable. We give simulations to illustrate our method, showing that it performs in a manner that is consistent with our analysis.

While nonparametric, the convex regression problem is naturally formulated using finite dimensional convex optimization, with no additional tuning parameters. The convex additive model can be used for convenience, without assuming it to actually hold, for the purpose of variable selection. As we show, our method scales to high dimensions, with a dependence on the intrinsic dimension s that scales polynomially, rather than exponentially as in the general case analyzed in Comminges and Dalalyan (2012).

Related Work

Perhaps more closely related to the present work is the framework studied by Raskutti et al. (2012) for sparse additive models, where sparse regression is considered under an additive assumption, with each component function belonging to an RKHS. An advantage of working over an RKHS is that nonparametric regression with a sparsity-

inducing regularization penalty can be formulated as a finite dimensional convex cone optimization. On the other hand, smoothing parameters for the component Hilbert spaces must be chosen, leading to extra tuning parameters that are difficult to select in practice. There has also been work on estimating sparse additive models over a spline basis, for instance the work of Huang et al. (2010), but these approaches too require the tuning of smoothing parameters.

Chapter Outline

In the following section we give a high-level summary of our technical results, including additive faithfulness, variable selection consistency, and high dimensional scaling. In Section 2.3 we give a detailed account of our method and the conditions under which we can guarantee consistent variable selection. In Section 2.4 we show how the required quadratic programs can be reformulated to be more efficient and scalable. In Section 2.5 we give the details of our finite sample analysis, showing that a sample size growing as $n = O\big(\text{poly}(s) \log p\big)$ is sufficient for variable selection. In Section 2.6 we report the results of simulations that illustrate our methods and theory. The full proofs are given in a technical appendix.

2.2 Overview of Results

In this section we provide a high-level description of our technical results. The full technical details, the precise statement of the results, and their detailed proofs are provided in following sections.

Our main contribution is an analysis of an additive approximation for identifying relevant variables in convex regression. We prove a result that shows when and how the additive approximation can be used without introducing false negatives in the population setting. In addition, we develop algorithms for the efficient implementation of the quadratic programs required by the procedure.

Faithful screening

The starting point for our approach is the observation that least squares nonparametric estimation under convexity constraints is equivalent to a finite dimensional

quadratic program. Specifically, the infinite dimensional optimization

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{n} (Y_i - f(\mathbf{x}_i))^2 \\ \text{subject to} \quad & f : \mathbb{R}^p \to \mathbb{R} \text{ is convex} \end{aligned} \tag{2.2.1}$$

is equivalent to the finite dimensional quadratic program

$$\begin{aligned} \text{minimize}_{f,\beta} \quad & \sum_{i=1}^{n} (Y_i - f_i)^2 \\ \text{subject to} \quad & f_j \geq f_i + \beta_i^T (\mathbf{x}_j - \mathbf{x}_i), \text{ for all } i, j. \end{aligned} \tag{2.2.2}$$

Here f_i is the estimated function value $f(\mathbf{x}_i)$, and the vectors $\beta_i \in \mathbb{R}^d$ represent supporting hyperplanes to the epigraph of f. See Boyd and Vandenberghe (2004), Section 6.5.5. Importantly, this finite dimensional quadratic program does not have tuning parameters for smoothing the function.

This formulation of convex regression is subject to the curse of dimensionality. Moreover, attempting to select variables by regularizing the subgradient vectors β_i with a group sparsity penality is not effective. Intuitively, the reason is that all p components of the subgradient β_i appear in every convexity constraint $f_j \geq f_i + \beta_i^T(\mathbf{x}_j - \mathbf{x}_i)$; small changes to the subgradients may not violate the constraints. Experimentally, we find that regularization with a group sparsity penality will make the subgradients of irrelevant variables small, but may not zero them out completely.

This motivates us to consider an additive approximation. As we show, this leads to an effective variable selection procedure. The shape constraints play an essential role. For general regression, using an additive approximation for variable selection may make errors. In particular, the nonlinearities in the regression function may result in an additive component being wrongly zeroed out. We show that this cannot happen for convex regression under appropriate conditions.

We say that a differentiable function f depends on variable x_k if $\partial_{x_k} f \neq 0$ with probability greater than zero. An additive approximation is given by

$$\{f_k^*\}, \mu^* := \underset{f_1,\ldots,f_p,\mu}{\arg\min} \Big\{ \mathbb{E}\Big(f(X) - \mu - \sum_{k=1}^{p} f_k(X_k) \Big)^2 \, : \, \mathbb{E} f_k(X_k) = 0 \Big\}. \tag{2.2.3}$$

We say that f is *additively faithful* in case $f_k^* = 0$ implies that f does not depend on coordinate k. Additive faithfulness is a desirable property since it implies that an additive approximation may allow us to screen out irrelevant variables.

Our first result shows that convex multivariate functions are additively faithful under the following assumption on the distribution of the data.

Definition 2.2.1. Let $p(\mathbf{x})$ be a density supported on $[0,1]^p$. Then p satisfies the *boundary flatness condition* if for all j, and for all $\mathbf{x}_{-j}$,

$$\frac{\partial p(\mathbf{x}_{-j} \,|\, x_j)}{\partial x_j} = \frac{\partial^2 p(\mathbf{x}_{-j} \,|\, x_j)}{\partial x_j^2} = 0 \quad \text{at } x_j = 0 \text{ and } x_j = 1.$$

As discussed in Section 2.3, this is a relatively weak condition. Our first result is that this condition suffices in the population setting of convex regression.

Theorem. (Theorem 2.3.1) Let $p(\mathbf{x})$ be a positive density supported on $C = [0,1]^p$ that satisfies the boundary flatness property. If f is convex with a bounded second derivative on an open set around C, then f is additively faithful under p.

Intuitively, an additive approximation zeroes out variable k when, fixing x_k, every "slice" of f integrates to zero. We prove this result by showing that "slices" of convex functions that integrate to zero cannot be "glued together" while still maintaining convexity.

While this shows that convex functions are additively faithful, it is difficult to estimate the optimal additive functions. The difficulty is that f_k^* need not be a convex function, as we show through a counterexample in Section 2.3. It may be possible to estimate f_k^* with smoothing parameters, but, for the purpose of variable screening, it is sufficient in fact to approximate f_k^* by a *convex* additive model.

Our next result states that a convex additive fit, combined with a series of univariate concave fits, is faithful. We abuse notation in the next theorem and let the notation f_k^* represent convex additive components.

Theorem. (Theorem 2.3.3) Suppose $p(\mathbf{x})$ is a positive density on $C = [0,1]^p$ that satisfies the boundary flatness condition. Suppose that f is convex and continuously twice-differentiable on an open set around C. and that $\partial_{x_k} f$, $\partial_{x_k} p(\mathbf{x}_{-k} \,|\, x_k)$, and $\partial_{x_k}^2 p(\mathbf{x}_{-k} \,|\, x_k)$ are all continuous as functions on C. Define

$$\{f_k^*\}_{k=1}^p, \mu^* = \arg\min_{\{f_k\},\mu} \Big\{ \mathbb{E}\Big(f(X) - \mu - \sum_{k=1}^s f_k(X_k) \Big)^2 \,:\, f_k \in \mathcal{C}^1,\, \mathbb{E} f_k(X_k) = 0 \Big\} \tag{2.2.4}$$

where $\mathcal{C}^1$ is the set of univariate convex functions, and, with respective to f_k^*'s from above, define

$$g_k^* = \arg\min_{g_k} \Big\{ \mathbb{E}\Big(f(X) - \mu^* - \sum_{k' \neq k} f_{k'}^*(X_{k'}) - g_k(X_k) \Big)^2 \,:\, g_k \in \text{-}\mathcal{C}^1, \mathbb{E} g_k(X_k) = 0 \Big\}, \tag{2.2.5}$$

with -$\mathcal{C}^1$ denoting the set of univariate concave functions. Then $f_k^* = 0$ and $g_k^* = 0$ implies that f does not depend on x_k, i.e., $\partial_{x_k} f(\mathbf{x}) = 0$ with probability one.

This result naturally suggests a two-stage screening procedure for variable selection. In the first stage we fit a sparse convex additive model $\{\widehat{f}_k\}$. In the second stage we fit a concave function $\widehat{g}_k$ to the residual for each variable having a zero convex component $\widehat{f}_k$. If both $\widehat{f}_k = 0$ and $\widehat{g}_k = 0$, we can safely discard variable x_k. As a shorthand, we refer to this two-stage procedure as AC/DC. In the AC stage we fit an additive convex model. In the DC stage we fit decoupled concave functions on the residuals. The decoupled nature of the DC stage allows all of the fits to be carried out in parallel. The entire process involves no smoothing parameters. Our next result concerns the required optimizations, and their finite sample statistical performance.

Optimization

Given samples (y_i, X_i), AC/DC becomes the following optimization.

$$\{\widehat{f}_k\}_{k=1}^p = \arg\min_{\{f_k \in \mathcal{C}^1\}} \frac{1}{n}\sum_{i=1}^n \left(y_i - \bar{y} - \sum_{k=1}^p f_k(X_{ik})\right)^2 + \lambda \sum_{k=1}^p \|f_k\|_\infty$$

$$\forall k,\ \widehat{g}_k = \arg\min_{g_k \in \mathcal{C}^1} \frac{1}{n}\sum_{i=1}^n \left(y_i - \bar{y} - \sum_{k' \neq k} \widehat{f}_{k'}(X_{ik'}) - g_k(X_{ik})\right)^2 + \lambda \|g_k\|_\infty$$

where $\bar{y}$ is the empirical mean of y. Our estimate of the relevant variables is $\widehat{S} = \{k : \|\widehat{f}_k\| > 0 \text{ or } \|\widehat{g}_k\| > 0\}$.

We present the optimization algorithms in Section 2.4. The convex constraints for the additive functions, analogous to the multivariate constraints (2.2.2), are that each component $f_k(\cdot)$ can be represented by its supporting hyperplanes, i.e.,

$$f_{ki'} \geq f_{ki} + \beta_{ki}(x_{ki'} - x_{ki}) \quad \text{for all } i, i' \tag{2.2.6}$$

where $f_{ki} \coloneqq f_k(x_{ki})$ and β_{ki} is the subgradient at point x_{ki}. While this apparently requires $O(n^2 p)$ equations to impose the supporting hyperplane constraints, in fact, only $O(np)$ constraints suffice. This is because univariate convex functions are characterized by the condition that the subgradient, which is a scalar, must increase monotonically. This observation leads to a reduced quadratic program with $O(np)$ variables and $O(np)$ constraints.

Directly applying a QP solver to this optimization is still computationally expensive for relatively large n and p. We thus develop a block coordinate descent method, where in each step we solve a sparse quadratic program involving $O(n)$ variables and $O(n)$ constraints. This is efficiently solved using optimization packages such as MOSEK. The details of these optimizations are given in Section 2.4.

Finite sample analysis

In Section 2.5 we analyze the finite sample variable selection consistency of AC/DC, without assuming that the true regression function f_0 is additive. Our analysis first establishes a sufficient deterministic condition for variable selection consistency, and then considers a stochastic setting. Our proof technique decomposes the KKT conditions for the optimization in a manner that is similar to the now standard *primal-dual witness* method (Wainwright, 2009).

We prove separate results that allow us to analyze false negative rates and false positive rates. To control false positives, we analyze scaling conditions on the regularization parameter λ_n for group sparsity needed to zero out irrelevant variables $k \in S^c$, where $S \subset \{1, \ldots, p\}$ is the set of variables selected by the AC/DC algorithm in the population setting. To control false negatives, we analyze the restricted regression where the variables in S^c are zeroed out, following the primal-dual strategy.

Each of our theorems uses a subset of the following assumptions:

A1: X_S, X_{S^c} are independent.

A2: f_0 is convex with a bounded second derivative. $\mathbb{E} f_0(X) = 0$.

A3: $\|f_0\|_\infty \leq sB$ and $\|f_k^*\|_\infty \leq B$ for all k.

A4: The noise is mean-zero sub-Gaussian with scale σ, independent of X.

A5: The density $p(\mathbf{x})$ is bounded away from $0/\infty$ and satisfies the boundary flatness condition.

In Assumption A3, $f^* = \sum_k f_k^*$ denotes the optimal additive projection of f_0 in the population setting.

Our analysis involves parameters α_+ and α_-, which are measures of the signal strength of the weakest variable:

$$\alpha_+ = \inf_{f \in \mathcal{C}^p \,:\, \mathrm{supp}(f) \subsetneq \mathrm{supp}(f^*)} \Big\{ \mathbb{E}\big(f_0(X) - f(X)\big)^2 - \mathbb{E}\big(f_0(X) - f^*(X)\big)^2 \Big\}$$

$$\alpha_- = \min_{k \in S \,:\, g_k^* \neq 0} \Big\{ \mathbb{E}\big(f_0(X) - f^*(X)\big)^2 - \mathbb{E}\big(f_0(X) - f^*(X) - g_k^*(X_k)\big)^2 \Big\}.$$

Intuitively, if α_+ is small, then it is easier to make a false omission in the additive convex stage of the procedure. If α_- is small, then it is easier to make a false omission in the decoupled concave stage of the procedure.

We make strong assumptions on the covariates in A1 in order to make very weak assumptions on the true regression function f_0 in A2; in particular, we do not assume that f_0 is additive. Relaxing this condition is an important direction for

future work. We also include an extra boundedness constraint to use new bracketing number results (Kim and Samworth, 2014).

Our main result is the following. Suppose assumptions A1-A5 hold. Let $\{\widehat{f}_i\}$ be any AC solution and let $\{\widehat{g}_k\}$ be any DC solution, both estimated with regularization parameter λ scaling as $\lambda = \Theta\Big(s\widetilde{\sigma}\sqrt{\frac{1}{n}\log^2 np}\Big)$. Suppose in addition that

$$\alpha_f/\widetilde{\sigma} \geq cB^2\sqrt{\frac{s^5}{n^{4/5}}\log^2 np} \tag{2.2.7}$$

$$\alpha_g^2/\widetilde{\sigma} \geq cB^4\sqrt{\frac{s^5}{n^{4/5}}\log^2 2np}. \tag{2.2.8}$$

where $\widetilde{\sigma} \equiv \max(\sigma, B)$ and c is a constant dependent only on b, c_1.

Then, for sufficiently large n, with probability at least $1 - \frac{1}{n}$:

$$\widehat{f}_k \neq 0 \text{ or } \widehat{g}_k \neq 0 \text{ for all } k \in S$$
$$\widehat{f}_k = 0 \text{ and } \widehat{g}_k = 0 \text{ for all } k \notin S.$$

This shows that variable selection consistency is achievable under exponential scaling of the ambient dimension, $p = O(\exp(cn))$ for some $0 < c < 1$, as for linear models. The cost of nonparametric estimation is reflected in the scaling with respect to $s = |S|$, which can grow only as $o(n^{4/25})$.

We remark that Comminges and Dalalyan (2012) show that, even with the product distribution, under traditional smoothness constraints, variable selection is achievable only if $n > O(e^s)$. Here we demonstrate that convexity yields the scaling $n = O(\text{poly}(s))$.

2.3 Population Level Analysis: Additive Faithfulness

For a general regression function, an additive approximation may result in a relevant variable being incorrectly marked as irrelevant. Such mistakes are inherent to the approximation and may persist even in the population setting. In this section we give examples of this phenomenon, and then show how the convexity assumption changes the behavior of the additive approximation. We work with $C = [0, 1]^p$ as the support of the distribution in this section but all of our results apply to general hypercubes. We begin with a lemma that characterizes the components of the additive approximation under mild conditions.

Lemma 2.3.1. *Let P be a distribution on $C = [0,1]^p$ with a positive density function $p(\mathbf{x})$. Let $f : C \to \mathbb{R}$ be in $L^2(P)$. Let*

$$f_1^*, ..., f_p^*, \mu^* := \\ argmin \left\{ \mathbb{E}\Big(f(X) - \mu - \sum_{k=1}^{p} f_k(X_k) \Big)^2 \,:\, f_k \in L^2(P),\, \mathbb{E} f_k(X_k) = 0,\; k = 1, \ldots, p \right\}.$$

With $\mu^* = \mathbb{E} f(X)$,

$$f_k^*(x_k) = \mathbb{E}\Big[f(X) - \sum_{k' \neq k} f_{k'}^*(X_{k'}) \,|\, x_k \Big] - \mathbb{E} f(X), \tag{2.3.1}$$

and this solution is unique.

Lemma 2.3.1 follows from the stationarity conditions of the optimal solution. This result is known, and criterion (2.3.1) is used in the backfitting algorithm for fitting additive models. We include a proof as our results build on it.

Proof. Let $f_1^*, ..., f_p^*, \mu^*$ be the minimizers as defined; they exist since the set of mean zero additive functions is a closed subspace of $L^2(P)$. We first show that the optimal μ is $\mu^* = \mathbb{E} f(X)$ for any $f_1, ..., f_k$ such that $\mathbb{E} f_k(X_k) = 0$. This follows from the stationarity condition, which states that $\mu^* = \mathbb{E}[f(X) - \sum_k f_k(X_k)] = \mathbb{E}[f(X)]$. Uniqueness is apparent because the second derivative is strictly larger than zero and strong convexity is guaranteed.

We now turn our attention toward the f_k^*s. It must be that f_k^* minimizes

$$\min_{f_k} \mathbb{E}\big(f(X) - \mu^* - \sum_{k' \neq k} f_{k'}^*(X_{k'}) - f_k(X_k) \big)^2 \tag{2.3.2}$$

subject to $\mathbb{E} f_k(X_k) = 0$. Fixing x_k, we will show that the value

$$\mathbb{E}[f(X) - \sum_{k' \neq k} f_{k'}^*(X_{k'}) \,|\, x_k] - \mu^* \tag{2.3.3}$$

uniquely minimizes

$$\min_{f_k(x_k)} \int_{\mathbf{x}_{-k}} p(\mathbf{x}) \Big(f(\mathbf{x}) - \sum_{k' \neq k} f_{k'}^*(x_{k'}) - f_k(x_k) - \mu^* \Big)^2 d\mathbf{x}_{-k}. \tag{2.3.4}$$

The first-order optimality condition gives us

$$\int_{\mathbf{x}_{-k}} p(\mathbf{x}) f_k(x_k) d\mathbf{x}_{-k} = \int_{\mathbf{x}_{-k}} p(\mathbf{x})(f(\mathbf{x}) - \sum_{k' \neq k} f^*_{k'}(x_{k'}) - \mu^*) d\mathbf{x}_{-k} \tag{2.3.5}$$

$$p(x_k) f_k(x_k) = \int_{\mathbf{x}_{-k}} p(x_k) p(\mathbf{x}_{-k} \,|\, x_k)(f(\mathbf{x}) - \sum_{k' \neq k} f^*_{k'}(x_{k'}) - \mu^*) d\mathbf{x}_{-k} \tag{2.3.6}$$

$$f_k(x_k) = \int_{\mathbf{x}_{-k}} p(\mathbf{x}_{-k} \,|\, x_k)(f(\mathbf{x}) - \sum_{k' \neq k} f^*_{k'}(x_{k'}) - \mu^*) d\mathbf{x}_{-k} \tag{2.3.7}$$

To prove uniqueness, suppose $\widetilde{f} = \sum_{j=1}^p \widetilde{f}_j$ is another additive function that achieves the same square error. Let $\nu \in [0,1]$, consider $\mathbb{E}\left(f(X) - \mu^* - (f^* + \nu(\widetilde{f} - f^*))\right)^2$ as a function of ν. The objective is strongly convex if $\mathbb{E}(\widetilde{f} - f^*)^2$, and so $\mathbb{E}(\widetilde{f} - f^*)^2 = 0$ by the assumption that f^* and $\widetilde{f}$ are both optimal solutions. By Lemma 2.7.3, we conclude that $\mathbb{E}(f^*_j - \widetilde{f}_j)^2 = 0$ as well and thus, $f^*_j = \widetilde{f}_j$ almost everywhere.

We note that $\mathbb{E}[f(X) - \sum_{k' \neq k} f^*_{k'}(X_{k'}) | x_k] - \mathbb{E}f(X)$ has mean zero as a function of x_k, which shows that f^*_k's are feasible.

□

In the case that the distribution in Lemma 2.3.1 is a product distribution, the additive components take on a simple form.

Corollary 2.3.1. *Let $p(\mathbf{x})$ be a positive density on $C = [0,1]^p$. Let $\mu^*, f^*_k(x_k)$ be defined as in Lemma 2.3.1. Then $\mu^* = \mathbb{E}f(X)$ and $f^*_k(x_k) = \mathbb{E}[f(X) \,|\, x_k] - \mathbb{E}f(X)$ and this solution is unique.*

In particular, under the uniform distribution, $f^*_k(x_k) = \int f(x_k, \mathbf{x}_{-k}) d\mathbf{x}_{-k} - \int f(\mathbf{x}) d\mathbf{x}$.

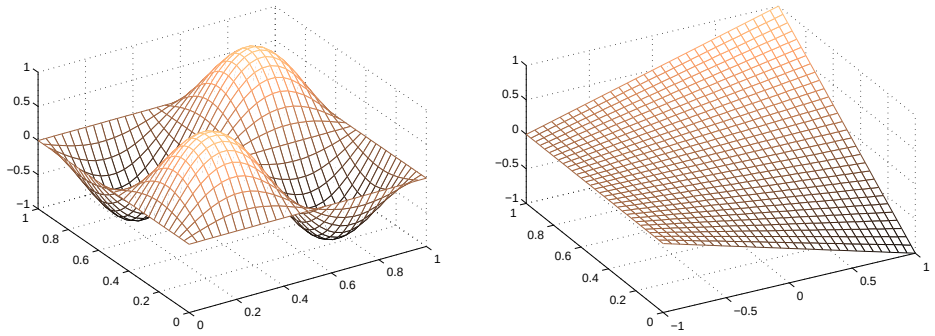
Example 2.3.1. Using Corollary 2.3.1, we give two examples of *additive unfaithfulness* under the uniform distribution—where relevant variables are erroneously marked as irrelevant under an additive approximation. First, consider the following function:

$$f(x_1, x_2) = \sin(2\pi x_1) \sin(2\pi x_2) \quad \text{(egg carton)} \tag{2.3.8}$$

defined for $(x_1, x_2) \in [0,1]^2$. Then $\int_{x_2} f(x_1, x_2) dx_2 = 0$ and $\int_{x_1} f(x_1, x_2) dx_1 = 0$ for each x_1 and x_2. An additive approximation would set $f_1 = 0$ and $f_2 = 0$. Next, consider the function

$$f(x_1, x_2) = x_1 x_2 \quad \text{(tilting slope)} \tag{2.3.9}$$

defined for $x_1 \in [-1, 1]$, $x_2 \in [0, 1]$. In this case $\int_{x_1} f(x_1, x_2)dx_1 = 0$ for each x_2; therefore, we expect $f_2 = 0$ under the additive approximation. This function, for every fixed x_2, is a zero-intercept linear function of x_1 with slope x_2.



(a) egg carton (b) tilting slope

Figure 2.1: Two additively unfaithful functions. Relevant variables are zeroed out under an additive approximation because every "slice" of the function integrates to zero.

In order to exploit additive models in variable selection, it is important to understand when the additive approximation accurately captures all of the relevant variables. We call this property *additive faithfulness*. We first formalize the concept that a multivariate function f *does not depend on* a coordinate x_k.

Definition 2.3.1. Let $C = [0, 1]^p$ and let $f : C \to \mathbb{R}$. We say that f *does not depend on coordinate* k if for all $\mathbf{x}_{-k}$, $f(x_k, \mathbf{x}_{-k})$ is a constant as a function of x_k. If f is differentiable, then f does not depend on k if $\partial_{x_k} f(x_k, \mathbf{x}_{-k})$ is 0 for all $\mathbf{x}_{-k}$.

In addition, suppose we have a distribution P over C and the additive approximation

$$f_k^*, \mu^* := \underset{f_1,\ldots,f_p,\mu}{\arg\min} \Big\{ \mathbb{E}\Big[\Big(f(X) - \sum_{k=1}^{p} f_k(X_k) - \mu \Big)^2 \Big] \ : \ \mathbb{E} f_k(X_k) = 0 \Big\}. \tag{2.3.10}$$

We say that f is *additively faithful* under P if $f_k^* = 0$ implies that f does not depend on coordinate k.

Additive faithfulness is an attractive property because it implies that, in the population setting, the additive approximation yields a consistent variable screening.

Additive Faithfulness of Convex Functions

We now show that under a general class of distributions which we characterize below, convex multivariate functions are additively faithful. To simplify presentation, we restrict our attention to densities bounded away from $0/\infty$, that is, $0 < \inf p(\mathbf{x}) \leq \sup p(\mathbf{x}) < \infty$.

Definition 2.3.2. Let $p(\mathbf{x})$ be a density supported on $[0,1]^p$. We say that $p(\mathbf{x})$ satisfies the *boundary flatness condition* if for all j, for all $\mathbf{x}_{-j}$, and for all $x_j \in [0,\epsilon) \cup (1-\epsilon, 1]$ for some arbitrarily small $\epsilon > 0$, $p(\mathbf{x}_{-j} \,|\, x_j)$ is twice differentiable in x_j, that $p(\mathbf{x}_{-j} \,|\, x_j)$, $\frac{\partial p(\mathbf{x}_{-j} \,|\, x_j)}{\partial x_j}$, $\frac{\partial^2 p(\mathbf{x}_{-j} \,|\, x_j)}{\partial^2 x_j}$ are bounded, and that

$$\frac{\partial p(\mathbf{x}_{-j} \,|\, x_j)}{\partial x_j} = \frac{\partial^2 p(\mathbf{x}_{-j} \,|\, x_j)}{\partial x_j^2} = 0 \quad \text{at } x_j = 0, x_j = 1 \tag{2.3.11}$$

The boundary flatness condition intuitively states that two conditional densities $p(\mathbf{x}_{-j} \,|\, x_j)$ and $p(\mathbf{x}_{-j} \,|\, x_j')$ are similar when x_j and x_j' are both close to the same boundary point. It is thus much more general than product densities. Boundary flatness is a weak condition because it affects only an ϵ-small region around the boundary; $p(\mathbf{x}_{-j} \,|\, x_j)$ can take arbitrary shapes away from the boundary. Boundary flatness also allows arbitrary correlation structure between the variables (provided $p(\mathbf{x}) > 0$). In Section 2.3, we give a detailed discussion of the boundary flatness condition and show examples of boundary flat densities; in particular, we show that any density supported on a compact set can be approximated arbitrarily well by boundary flat densities.

The following theorem is the main result of this section.

Theorem 2.3.1. *Let $p(\mathbf{x})$ be a density supported on $C = [0,1]^p$ and bounded away from $0/\infty$ that satisfies the boundary flatness property.*

Suppose f is a convex with a bounded second derivative on an open set containing C, then f is additively faithful under $p(\mathbf{x})$.

We let the domain of f be slightly larger than C for a technical reason–it is so we can say in the proof that the Hessian of f is positive semidefinite even at the boundary of C.

We pause to give some intuition before we present the full proof. Suppose that the underlying density is a product density first. We know from Lemma 2.3.1 that the additive approximation zeroes out k when, fixing x_k, every "slice" of f integrates to zero, but "slices" of convex functions that integrate to zero cannot be "glued together" while still maintaining convexity. Since the behavior of the whole convex

function is constrained by its behavior at the boundary, the same result holds even if the underlying density is not a product density but merely resembles a product density at the boundary, which is exactly the notion formalized by the boundary flatness condition.

Proof. Fixing k and using the result of Lemma 2.3.1, we need only show that for all x_k, $\mathbb{E}[f(X) - \sum_{k'} f_{k'}(X_{k'}) \,|\, x_k] - \mathbb{E}f(X) = 0$ implies that f does not depend on coordinate k, i.e., $\partial_{x_k} f(\mathbf{x}) = 0$ for all $\mathbf{x}$.

Let us use the shorthand notation that $r(\mathbf{x}_{-k}) = \sum_{k' \neq k} f_{k'}(x_{k'})$ and assume without loss of generality that $\mu^* = E[f(X)] = 0$. We then assume that for all x_k,

$$\mathbb{E}[f(X) - r(X_{-k}) \,|\, x_k] \equiv \int_{\mathbf{x}_{-k}} p(\mathbf{x}_{-k} \,|\, x_k)\big(f(\mathbf{x}) - r(\mathbf{x}_{-k})\big) = 0. \tag{2.3.12}$$

We let $p'(\mathbf{x}_{-k} \,|\, x_k)$ denote $\frac{\partial p(\mathbf{x}_{-k} \,|\, x_k)}{\partial x_k}$ and $p''(\mathbf{x}_{-k} \,|\, x_k)$ denote $\frac{\partial^2 p(\mathbf{x}_{-k} \,|\, x_k)}{\partial x_k^2}$ and likewise for $f'(x_k, \mathbf{x}_{-k})$ and $f''(x_k, \mathbf{x}_{-k})$.

We differentiate with respect to x_k at $x_k = 0, 1$ under the integral. The detail necessary to verify the validity of this operation is technical and given in Section 2.7 of the supplementary material.

$$\int_{\mathbf{x}_{-k}} p'(\mathbf{x}_{-k} \,|\, x_k)\big(f(\mathbf{x}) - r(\mathbf{x}_{-k})\big) + p(\mathbf{x}_{-k} \,|\, x_k) f'(x_k, \mathbf{x}_{-k}) d\mathbf{x}_{-k} = 0 \tag{2.3.13}$$

$$\int_{\mathbf{x}_{-k}} p''(\mathbf{x}_{-k} \,|\, x_k)\big(f(\mathbf{x}) - r(\mathbf{x}_{-k})\big) + 2p'(\mathbf{x}_{-k} \,|\, x_k) f'(x_k, \mathbf{x}_{-k}) + p(\mathbf{x}_{-k} \,|\, x_k) f''(x_k, \mathbf{x}_{-k}) d\mathbf{x}_{-k} = 0. \tag{2.3.14}$$

By the boundary flatness condition, we have that $p''(\mathbf{x}_{-k} \,|\, x_k)$ and $p'(\mathbf{x}_{-k} \,|\, x_k)$ are zero at $x_k = x_k^0 \equiv 0$. The integral equations then reduce to the following:

$$\int_{\mathbf{x}_{-k}} p(\mathbf{x}_{-k} \,|\, x_k^0) f'(x_k^0, \mathbf{x}_{-k}) d\mathbf{x}_{-k} = 0 \tag{2.3.15}$$

$$\int_{\mathbf{x}_{-k}} p(\mathbf{x}_{-k} \,|\, x_k^0) f''(x_k^0, \mathbf{x}_{-k}) d\mathbf{x}_{-k} = 0. \tag{2.3.16}$$

Because f is convex, $f(x_k, \mathbf{x}_{-k})$ must be a convex function of x_k for all $\mathbf{x}_{-k}$. Therefore, for all $\mathbf{x}_{-k}$, $f''(x_k^0, \mathbf{x}_{-k}) \geq 0$. Since $p(\mathbf{x}_{-k} \,|\, x_k^0) > 0$ by the assumption that $p(\mathbf{x})$ is a positive density, we have that $\forall \mathbf{x}_{-k}, f''(x_k^0, \mathbf{x}_{-k}) = 0$ necessarily.

The Hessian of f at $(x_k^0, \mathbf{x}_{-k})$ then has a zero at the k-th main diagonal entry. A positive semidefinite matrix with a zero on the k-th main diagonal entry must have

only zeros on the k-th row and column; see proposition 7.1.10 of Horn and Johnson (1990). Thus, at all $\mathbf{x}_{-k}$, the gradient of $f'(x_k^0, \mathbf{x}_{-k})$ with respect to $\mathbf{x}_{-k}$ must be zero. Therefore, $f'(x_k^0, \mathbf{x}_{-k})$ must be constant for all $\mathbf{x}_{-k}$. By equation 2.3.15, we conclude that $f'(x_k^0, \mathbf{x}_{-k}) = 0$ for all $\mathbf{x}_{-k}$. We can use the same reasoning for the case where $x_k = x_k^1$ and deduce that $f'(x_k^1, \mathbf{x}_{-k}) = 0$ for all $\mathbf{x}_{-k}$.

Because $f(x_k, \mathbf{x}_{-k})$ as a function of x_k is convex, it must be that, for all $x_k \in (0, 1)$ and for all $\mathbf{x}_{-k}$,

$$0 = f'(x_k^0, \mathbf{x}_{-k}) \leq f'(x_k, \mathbf{x}_{-k}) \leq f'(x_k^1, \mathbf{x}_{-k}) = 0 \tag{2.3.17}$$

Therefore f does not depend on x_k.

□

Theorem 2.3.1 plays an important role in our finite sample analysis, where we show that the additive approximation is variable screening consistent, even when the true function is not additive.

Remark 2.3.1. We assume twice differentiability in Theorems 2.3.1 to simplify the proof. We expect, however, that this smoothness condition is not necessary—every convex function can be approximated arbitrarily well by a smooth convex function.

Remark 2.3.2. In Theorem 2.3.1, we do not assume a parametric form for the additive components; the additive approximations may not be faithful if we take a parametric form. For example, suppose we approximate a mean-zero convex function $f(X)$ by a linear form $X\beta$. The optimal linear function in the population setting is $\beta^* = \Sigma^{-1}\text{Cov}(X, f(X))$ where Σ is the covariance matrix. Suppose the X's are independent, follow a symmetric distribution, have unit variance, and suppose $f(\mathbf{x}) = x_1^2 - \mathbb{E}[X_1^2]$, then $\beta_1^* = \mathbb{E}[X_1 f(X)] = \mathbb{E}[X_1^3 - X_1\mathbb{E}[X_1^2]] = 0$.

Boundary Flatness Examples

In this section, we give more examples of boundary flat densities (see Definition 2.3.2) and discuss extending the notion of boundary flatness to densities with a more general support. We first start with an sufficient condition on the *joint density* that ensures boundary flatness.

Example 2.3.2. Boundary flatness is satisfied if the joint density becomes flat at the boundary. To be precise, let $p(\mathbf{x})$ be a joint density bounded away from $0/\infty$ with a bounded second derivative.

Suppose also, for all j,

$$\partial_{x_j} p(x_j, \mathbf{x}_{-j}) = \partial^2_{x_j} p(x_j, \mathbf{x}_{-j}) = 0 \quad \text{at } x_j = 0, 1,.$$

It is then straightforward to show boundary flatness. One can first verify that the derivatives of the marginal density $p(x_j)$ vanishes at $x_j = 0, 1$ and then apply the quotient rule on $\frac{p(x_j, \mathbf{x}_{-j}}{p(x_j)}$ to show that $\partial_{x_j} p(\mathbf{x}_{-j} \,|\, x_j) = \partial^2_{x_j} p(\mathbf{x}_{-j} \,|\, x_j) = 0$ at $x_j = 0, 1$ as well.

The next example shows that any bounded density over a hypercube can be approximated arbitrarily well by boundary flat densities.

Example 2.3.3. Suppose $p_\epsilon(\mathbf{x})$ is a bounded density over $[\epsilon, 1-\epsilon]^p$ for some $0 < \epsilon < 1/2$. Let $q(\mathbf{x})$ be an arbitrary boundary flat density over $[0,1]^p$ (one can take the uniform density for instance). Define a mixture $p_{\lambda,\epsilon}(\mathbf{x}) = \lambda q(\mathbf{x}) + (1-\lambda) p_\epsilon(\mathbf{x})$ where $0 < \lambda \leq 1$, then $p_{\lambda,\epsilon}(\mathbf{x})$ is boundary flat over $[0,1]^p$.

Now, let $p(\mathbf{x})$ be a bounded density over $[0,1]^p$. Let $p_\epsilon(\mathbf{x})$ be the density formed from truncating $p(\mathbf{x})$ in $[\epsilon, 1-\epsilon]^p$. The corresponding mixture $p_{\lambda,\epsilon}(\mathbf{x})$ then approximates $p(\mathbf{x})$ when λ and ϵ are both small.

Since $p_{\lambda,\epsilon}(\mathbf{x})$ remains boundary flat for arbitrarily small ϵ and λ, $p(\mathbf{x})$ can be approximated arbitrarily well (in L_1 for example) by boundary flat densities.

In our discussion so far, we have restricted ourselves to densities supported and positive on the hypercube $[0,1]^p$ to minimize extraneous technical details. It may be possible to extend the analysis to densities whose support is a convex and compact set so long as the marginal density $p(x_j) > 0$ for all x_j in the support. A rigorous analysis of this however is beyond the scope of this paper.

It may also possible to extend similar result to densities with an unbounded support, by using a limit condition $\lim_{|x_k| \to \infty} \frac{\partial p(\mathbf{x}_{-k} \,|\, x_k)}{\partial x_k} = 0$. Such a limit condition however is not obeyed by a correlated multivariate Gaussian distribution. The next example shows that certain convex functions are not additively faithful under certain multivariate Gaussian distributions.

Example 2.3.4. Consider a two dimensional quadratic function $f(\mathbf{x}) = \mathbf{x}^\mathsf{T} H \mathbf{x} + c$ with zero mean where $H = \begin{pmatrix} H_{11} & H_{12} \\ H_{12} & H_{22} \end{pmatrix}$ is positive definite and a Gaussian distribution $X \sim N(0, \Sigma)$ where $\Sigma = \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix}$. As we show in Section 2.8 of the Appendix,

the additive approximation has the following closed form.

$$f_1^*(x_1) = \left(\frac{T_1 - T_2\alpha^2}{1-\alpha^4}\right) x_1^2 + c_1$$
$$f_2^*(x_2) = \left(\frac{T_2 - T_1\alpha^2}{1-\alpha^4}\right) x_2^2 + c_2$$

Where $T_1 = H_{11} + 2H_{12}\alpha + H_{22}\alpha^2$, $T_2 = H_{22} + 2H_{12}\alpha + H_{11}\alpha^2$, c_1, c_2 are constants such that f_1^* and f_2^* both have mean zero. Let $H = \begin{pmatrix} 1.6 & 2 \\ 2 & 5 \end{pmatrix}$, then it is easy to check that if $\alpha = -\frac{1}{2}$, then $f_1^* = 0$ and additive faithfulness is violated, if $\alpha > \frac{1}{2}$, then f_1^* is a concave function. We take the setting where $\alpha = -0.5$, compute the optimal additive functions via numerical simulation, and show the results in Figure 2.2(a)–f_1^* is zero as expected.

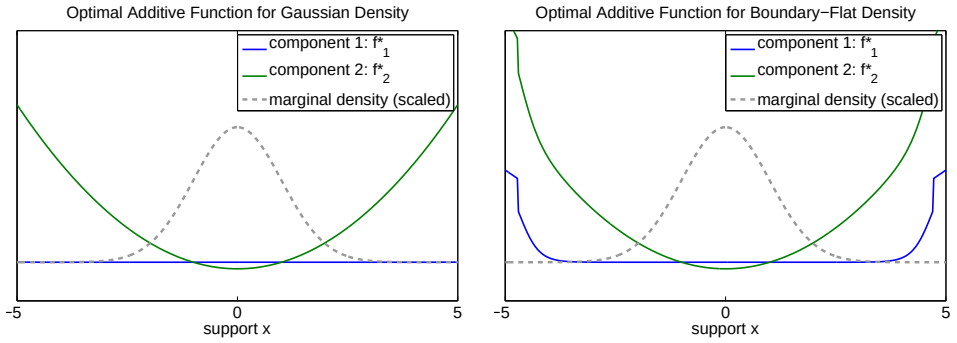
Although the Gaussian distribution does not satisfy the boundary flatness condition, it is possible to approximate the Gaussian distribution arbitrarily well with distributions that do satisfy the boundary flatness conditions. We use the similar idea as that of Example 2.3.3.

Example 2.3.5. Let Σ be as in Example 2.3.4 with $\alpha = -0.5$ so that $f_1^* = 0$. Consider a mixture $\lambda U[-(b+\epsilon), b+\epsilon]^2 + (1-\lambda)N_b(0,\Sigma)$ where $N_b(0,\Sigma)$ is the density of a *truncated* bivariate Gaussian bounded in $[-b,b]^2$ and $U[-(b+\epsilon), b+\epsilon]^2$ is the uniform distribution over a square. The uniform distribution is supported over a slightly larger square to satisfy the boundary flatness conditions.

When b is large, ϵ is small, and λ is small, the mixture closely approximates the Gaussian distribution but is still additively faithful for convex functions. Figure 2.2(b) shows the optimal additive components under the mixture distribution, computed by numerical integration with $b = 5, \epsilon = 0.3, \lambda = 0.0001$. True to our theory, f_1^*, which is zero under the Gaussian distribution, is nonzero under the mixture approximation to the Gaussian distribution. We note that the magnitude $\mathbb{E}f_1^*(X_1)^2$, although non-zero, is very small, consistent with the fact that the mixture distribution closely approximates the Gaussian distribution.

Converse to Faithfulness

It is difficult to find natural conditions under which the opposite direction of additive faithfulness holds—conditions implying that if f does not depend on coordinate k, then f_k^* will be zero in the additive approximation. Suppose, for example, that f is



(a) Gaussian distribution (b) Mixture approximation

Figure 2.2: Optimal additive projection of the quadratic function described in Example 2.3.4 under both the Gaussian distribution described in Example 2.3.4 and under the approximately Gaussian mixture distribution described in Example 2.3.5. For the mixture approximation, we used $b = 5, \epsilon = 0.3, \lambda = 0.0001$ where the parameters are defined in Example 2.3.5. This example shows the effect and the importance of the boundary flatness conditions.

only a function of X_1, X_2, and that (X_1, X_2, X_3) follows a degenerate 3-dimensional distribution where $X_3 = f(X_1, X_2) - f^*(X_1) - f_2^*(X_2)$. In this case X_3 exactly captures the additive approximation error. The best additive approximation of f would have a component $f_3^*(x_3) = x_3$ even though f does not depend on x_3.

The simplest case under which the converse holds as well is the product density. In this case, if f does not depend on X_k, then $\mathbb{E}[f(X) - r(X_{-k}) \,|\, X_k] = 0$ for any function $r(X_{-k})$. In this section, we will generalize the product density into another condition for which we can guarantee the converse.

Theorem 2.3.2. *Let f be a function such that $\mathbb{E}f(X) = 0$. Let S_0 be the set of relevant variables of f. Let $k \notin S_0$ and suppose there exists $k' \in S_0$ such that X_k is independent of $X_{S_0 - \{k'\}}$ conditional on $X_{k'}$. Then, $f_k^* = 0$.*

Proof. Let $r(\mathbf{x}_{S_0}) = \sum_{k \in S_0} \widetilde{f}_k(x_k)$ be the additive projection of f restricted to only the relevant variables S_0. We will show that it is also in fact, the additive projection of f without the variable restriction.

Suppose $k \notin S_0$. Let $k' \in S_0$ be the variable such that $X_k \perp X_{S_0 - \{k'\}} \,|\, X_{k'}$.

Then,

$$\begin{aligned}
&\mathbb{E}\left[f(X) - \sum_{j\in S_0} \widetilde{f}(X_j) \,|\, X_k\right] \\
&= \mathbb{E}_{X_{k'}}\left[\mathbb{E}\left[f(X) - \sum_{j\in S_0} \widetilde{f}(X_j) \,|\, X_{k'}, X_k\right] \,\Big|\, X_k\right] \\
&= \mathbb{E}_{X_{k'}}\left[\mathbb{E}\left[f(X) - \sum_{j\in S_0} \widetilde{f}(X_j) \,|\, X_{k'}\right] \,\Big|\, X_k\right] \\
&= 0
\end{aligned}$$

The third equality follows because $f(X) - \sum_{j\in S_0} \widetilde{f}_j(X_j)$ is a function of X_{S_0} only and thus is independent of X_k when conditioned on $X_{k'}$. For the fourth equality, observe that $\sum_{j\in S_0} \widetilde{f}_j$ is the additive projection of f restricted on S_0 and thus, $\mathbb{E}[f(X) - \sum_{j\in S_0} \widetilde{f}_j(X_j) \,|\, X_{k'}] = 0$ by Lemma 2.3.1.

The theorem follows since this analysis holds for every $k \in S_0$, □

There is an easier way to interprete the condition in Theorem 2.3.2 using the language of graphical models. Let $\mathcal{G}$ be the conditional independence graph of X, that is, for every j, $X_j \perp X_k \,|\, X_{N(j)}$ for every k not in the Graph neighborhood $N(j)$ of node X_j. The condition in Theorem 2.3.2 is equivalent to saying that for every $k \notin S_0$, there exists only one path in $\mathcal{G}$ that connects node X_k to the set of nodes X_{S_0}. See figure 2.3 for a visual example.

Convex Additive Models

Although convex functions are additively faithful—under appropriate conditions—it is difficult to estimate the optimal additive functions f_k^*s as defined in equation (2.3.10). The reason is that f_k^* need not be a convex function, as example 2.3.4 and example 2.3.5 show. It may be possible to estimate f_k^* via smoothing, but we prefer an approach that is free of smoothing parameters. Since the true regression function f is convex, we approximate the additive model with a *convex* additive model. We abuse notation and, for the rest of the paper, use the notation f_k^* to represent convex additive fits:

$$\{f_k^*\}_{k=1}^p = \arg\min \left\{\mathbb{E}\Big(f(X) - \sum_{k=1}^p f_k(X_k)\Big)^2 \,:\, f_k \in \mathcal{C}^1,\, \mathbb{E}f_k(X_k) = 0\right\} \qquad (2.3.18)$$

where $\mathcal{C}^1$ is the set of univariate convex functions.

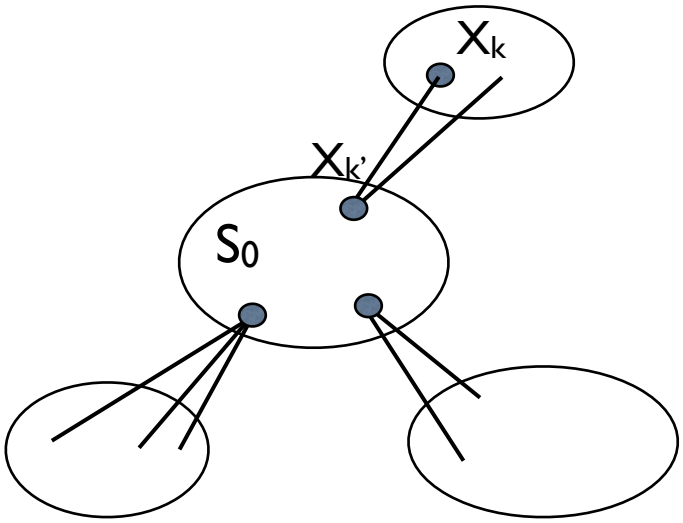


Figure 2.3: A conditional independence graph that satisfies the condition in Theorem 2.3.2

If $p(\mathbf{x})$ is a product density, then $\mathbb{E}[f(X) \mid x_k]$ is convex in x_k and the additive projection is simultaneously the convex additive projection. Thus, in this case, additive faithfulness trivially holds for the convex additive projection. For a general boundary flat density $p(\mathbf{x})$ however, the additive projection need not be convex and we thus cannot say anything about additive faithfulness of the convex additive projection.

Luckily, we can restore faithfulness by coupling the f_k^*'s with a set of univariate concave fits on the *residual* $f - f^*$:

$$g_k^* = \arg\min \Big\{ \mathbb{E}\Big(f(X) - \sum_{k' \neq k} f_{k'}^*(X_{k'}) - g_k(X_k) \Big)^2 \; : \; g_k \in \text{-}\mathcal{C}^1, \mathbb{E} g_k(X_k) = 0 \Big\}. \quad (2.3.19)$$

Theorem 2.3.3. *Suppose $p(\mathbf{x})$ is a density on $C = [0,1]^p$ bounded away from $0/\infty$ that satisfies the boundary flatness condition. Suppose that f is convex with a bounded*

second derivative on an open set around C. Let f_k^* and g_k^* be as defined in equations (2.3.18) *and* (2.3.19)*, then the f_k^*'s and the g_k^*'s are unique. Furthermore, $f_k^* = 0$ and $g_k^* = 0$ implies that $\partial_{x_k} f(\mathbf{x}) = 0$, that is, f does not depend on x_k.*

Before we can prove the theorem, we need a lemma that generalizes Theorem 2.3.1.

Lemma 2.3.2. *Suppose $p(\mathbf{x})$ is a density on $C = [0,1]^p$ bounded away from $0/\infty$ satisfying the boundary flatness condition. Let $f(\mathbf{x})$ be a convex function with a bounded second derivative on an open set around C. Let $\phi(\mathbf{x}_{-k})$ be a bounded function that does not depend on x_k. Then, we have that the unconstrained univariate function*

$$h_k^* = \arg\min_{f_k} \mathbb{E}\Big[\big(f(X) - \phi(X_{-k}) - h_k(X_k)\big)^2\Big] \tag{2.3.20}$$

is given by $h_k^*(x_k) = \mathbb{E}\big[f(X) - \phi(X_{-k}) \,|\, x_k\big]$, and $h_k^* = 0$ implies that $\partial_{x_k} f(\mathbf{x}) = 0$.

Proof. In the proof of Theorem 2.3.1, the only property of $r(\mathbf{x}_{-k})$ we used was the fact that $\partial_{x_k} r(\mathbf{x}_{-k}) = 0$. Therefore, the proof here is identical to that of Theorem 2.3.1 except that we replace $r(\mathbf{x}_{-k})$ with $\phi(\mathbf{x}_{-k})$. □

Proof of theorem 2.3.3. Fix k. Let f_k^* and g_k^* be defined as in equation 2.3.18 and equation 2.3.19. Let $\phi(\mathbf{x}_{-k}) \equiv \sum_{k' \neq k} f_{k'}^*(x_{k'})$. Each $f_{k'}^*$ is convex and thus continuous on $(0,1)$. $f_{k'}^*(x_{k'})$ is defined at $x_{k'} = 0, 1$; thus, $f_{k'}^*$ must be bounded and $\phi(\mathbf{x}_{-k})$ is bounded.

We have that

$$f_k^* = \arg\min_{f_k} \Big\{\mathbb{E}\big(f(X) - \sum_{k' \neq k} f_{k'}^*(X_{k'}) - f_k\big)^2 \,:\, f_k \in \mathcal{C}^1,\, \mathbb{E} f_k(X_k) = 0\Big\} \tag{2.3.21}$$

$$g_k^* = \arg\min_{g_k} \Big\{\mathbb{E}\big(f(X) - \sum_{k' \neq k} f_{k'}^*(X_{k'}) - g_k\big)^2 \,:\, g_k \in \text{-}\mathcal{C}^1,\, \mathbb{E} g_k(X_k) = 0\Big\} \tag{2.3.22}$$

Let us suppose that $f_k^* = g_k^* = 0$. It must be then that

$$\arg\min_{c \in \mathbb{R}} \mathbb{E}\left(f(X) - \phi(X_{-k}) - c(X_k^2 - m_k^2)\right)^2 = 0$$

where $m_k^2 \equiv \mathbb{E} X_k^2$; this is because $c(x_k^2 - m_k^2)$ is either convex or concave in x_k and it is centered, i.e. $\mathbb{E}[X_k^2 - m_k^2] = 0$. Since the optimum has a closed form $c^* = \frac{\mathbb{E}\left[(f(X) - \phi(X_{-k}))(X_k^2 - m_k^2)\right]}{\mathbb{E} X_k^2 - m_k^2}$, we deduce that

$$\mathbb{E}\big[(f(X) - \phi(X_{-k}))(X_k^2 - m_k^2)\big] = \mathbb{E}[(f(X) - \phi(X_{-k}))X_k^2] = \mathbb{E}[\mathbb{E}[f(X) - \phi(X_{-k}) \,|\, X_k] X_k^2] = 0$$

We denote $h_k^*(x_k) = \mathbb{E}[f(X) - \phi(X_{-k}) \,|\, x_k]$. $f(\mathbf{x})$ and $\phi(\mathbf{x}_{-k})$ are both bounded and so $h_k^*(x_k)$ is bounded as well. Therefore, h_k^* is square integrable and there exists a fourier series $s_n(x_k)$ convergent to h_k^* in L_2. Since $p(\mathbf{x})$ is bounded,

$$\lim_{n\to\infty} \mathbb{E}\left(s_n(X_k) - h_k^*(X_k)\right)^2 \to 0$$

as well.

If we can show that $\mathbb{E}h_k^*(X_k)^2 = 0$, we would apply Lemma 2.3.2 and finish the proof. So let us suppose for sake of contradiction that $\mathbb{E}h_k^*(X_k)^2 > 0$.

Let $0 < \epsilon < 1$ be fixed and let n be large enough such that $\mathbb{E}(s_n(X_k) - h_k^*(X_k))^2 \leq \epsilon \mathbb{E}h_k^*(X_k)^2$.

Since $s_n(x_k)$ is twice-differentiable and has a second derivative bounded away from $-\infty$, there exist some positive scalar α such that $s_n(x_k) + \alpha(x_k^2 - m_k^2)$ has a non-negative second derivative and is thus convex.

Because we assumed $f^* = g^* = 0$, it must be that

$$\operatorname*{arg\,min}_{c\in\mathbb{R}} \mathbb{E}\Big(f(X) - \phi(X_{-k}) - c\big(s_n(X_k) + \alpha(X_k^2 - m_k^2)\big)\Big)^2 = 0$$

This is because $c\big(s_n(x_k) + \alpha(x_k^2 - m_k^2)\big)$ is convex for $c \geq 0$ and concave for $c \leq 0$ and it is a centered function.

Again, $c^* = \frac{\mathbb{E}[(f(X) - \phi(X_{-k}))\left(s_n(X_k) + \alpha(X_k^2 - m_k^2)\right)]}{\mathbb{E}\left(s_n(X_k) + \alpha(X_k^2 - m_k^2)\right)^2} = 0$, so

$$\begin{aligned}
\mathbb{E}[(f(X) - \phi(X_{-k}))\big(s_n(X_k) + \alpha(X_k^2 - m_k^2)\big)] &= \mathbb{E}[(f(X) - \phi(X_{-k}))s_n(X_k)] \\
&= \mathbb{E}\Big[\mathbb{E}[f(X) - \phi(X_{-k}) \,|\, X_k]s_n(X_k)\Big] \\
&= \mathbb{E}h_k^*(X_k)s_n(X_k) = 0
\end{aligned}$$

where the first equality follows because $\mathbb{E}[(f(X) - \phi(X_{-k}))(X_k^2 - m_k^2)] = 0$.

We have chosen s_n such that $\mathbb{E}(h_k^*(X_k) - s_n(X_k))^2 \leq \epsilon \mathbb{E}h_k^*(X_k)^2$ for some $\epsilon < 1$. But, $\mathbb{E}(h_k^*(X_k) - s_n(X_k))^2 = \mathbb{E}h_k^*(X_k)^2 - 2\mathbb{E}h_k^*(X_k)s_n(X_k) + \mathbb{E}s_n(X_k)^2 \geq \mathbb{E}h_k^*(X_k)^2$. This is a contradiction and therefore, $\mathbb{E}h_k^*(X_k)^2 = 0$.

Now we use Lemma 2.3.2 with $\phi(\mathbf{x}_{-k}) = f(\mathbf{x}) - \sum_{k'\neq k} f_{k'}^*(x_{k'})$ and conclude that $f_k^* = 0$ and $g_k^* = 0$ together imply that f does not depend on x_k.

Now we turn to uniqueness. Suppose for sake of contradiction that f^* and $\widetilde{f}$ are optimal solutions to (2.3.18) and $\mathbb{E}(\widetilde{f} - f^*)^2 > 0$. $f^* + \lambda(\widetilde{f} - f^*)$ for any $\lambda \in [0, 1]$ must then also be an optimal solution by convexity of the objective and constraint. However, the second derivative of the objective $\mathbb{E}(f - f^* - \lambda(\widetilde{f} - f^*))^2$ with respect to λ is $2\mathbb{E}(\widetilde{f} - f^*)^2 > 0$. The objective is thus strongly convex and $\mathbb{E}(f^* - \widetilde{f})^2 = 0$. We now apply Lemma 2.7.3 by letting $\phi_k = f_k^* - \widetilde{f}_k$. We conclude that $\mathbb{E}(f_k^* - \widetilde{f}_k)^2 = 0$ for all k. The uniqueness of g^* is proved similarly. □

Estimation Procedure

Theorem 2.3.3 naturally suggests a two-stage screening procedure for variable selection in the population setting. In the first stage, we fit a convex additive model.

$$f_1^*, ..., f_p^* = \underset{f_1, ..., f_p \in \mathcal{C}_0^1, \mu}{\operatorname{argmin}} \mathbb{E}\Big(f(X) - \mu - \sum_{k=1}^{p} f_k(X_k)\Big)^2 \tag{2.3.23}$$

where we denote $\mathcal{C}_0^1$ ($\text{-}\mathcal{C}_0^1$) as the set of one-dimensional convex (resp. concave) functions with population mean zero. In the second stage, for every variable marked as irrelevant in the first stage, we fit a univariate *concave* function separately on the residual for that variable. For each k such that $f_k^* = 0$:

$$g_k^* = \underset{g_k \in \text{-}\mathcal{C}_0^1}{\operatorname{argmin}} \mathbb{E}\Big(f(X) - \mu^* - \sum_{k'} f_{k'}^*(X_{k'}) - g_k(X_k)\Big)^2 \tag{2.3.24}$$

We screen out S^C, any variable k that is zero after the second stage, and output S.

$$S^c = \big\{k : f_k^* = 0 \text{ and } g_k^* = 0\big\}. \tag{2.3.25}$$

We refer to this procedure as AC/DC (additive convex/decoupled concave). Theorem 2.3.3 guarantees that the true set of relevant variables S_0 must be a subset of S.

It is straightforward to construct a finite sample variable screening procedure, which we describe in Figure 2.4. We use an ℓ_∞/ℓ_1 penalty in equation (2.3.26) and an ℓ_∞ penalty in equation (2.3.24) to encourage sparsity. Other penalties can also produce sparse estimates, such as a penalty on the derivative of each of the component functions. The $\|\cdot\|_\infty$ norm is convenient for both theoretical analysis and implementation.

After selecting the variable set $\widehat{S}$, one can refit a low-dimensional non-additive convex function to build the best predictive model. If refitting is undesirable for

AC/DC Algorithm for Variable Selection in Convex Regression

Input: $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$, regularization parameter λ.

AC Stage: Estimate a sparse additive convex model:

$$\widehat{f}_1, ..., \widehat{f}_p, \widehat{\mu} = \underset{f_1,...,f_p \in \mathcal{C}_0^1}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \Big(y_i - \mu - \sum_{k=1}^{p} f_k(x_{ik}) \Big)^2 + \lambda \sum_{k=1}^{p} \|f_k\|_\infty \quad (2.3.26)$$

DC Stage: Estimate concave functions for each k such that $\|\widehat{f}_k\|_\infty = 0$:

$$\widehat{g}_k = \underset{g_k \in \text{-}\mathcal{C}_0^1}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \Big(y_i - \widehat{\mu} - \sum_{k'} \widehat{f}_{k'}(x_{ik'}) - g_k(x_{ik}) \Big)^2 + \lambda \|g_k\|_\infty \quad (2.3.27)$$

Output: Component functions $\{\widehat{f}_k\}$ and relevant variables $\widehat{S}$ where

$$\widehat{S}^c = \big\{ k : \|\widehat{f}_k\| = 0 \text{ and } \|\widehat{g}_k\| = 0 \big\}. \quad (2.3.28)$$

Figure 2.4: The AC/DC algorithm for variable selection in convex regression. The AC stage fits a sparse additive convex regression model, using a quadratic program that imposes an group sparsity penalty for each component function. The DC stage fits decoupled concave functions on the residuals, for each component that is zeroed out in the AC stage.

whatever reason, the AC/DC outputs can also be used for prediction. Given a new sample $\mathbf{x}$, we let $\widehat{y} = \sum_k \widehat{f}_k(\mathbf{x}_k) + \sum_k \widehat{g}_k(\mathbf{x}_k)$. Note that $\widehat{g}_k = 0$ for k such that $\widehat{f}_k \neq 0$ in AC/DC. The next section describes how to compute this function evaluation.

The optimization in (2.3.26) appears to be infinite dimensional, but it is equivalent to a finite dimensional quadratic program. In the following section, we give the details of this optimization, and show how it can be reformulated to be more computationally efficient.

2.4 Optimization

We now describe in detail the optimization algorithm for the additive convex regression stage. The second decoupled concave regression stage follows a very similar procedure.

Let $bdsx_i \in \mathbb{R}^p$ be the covariate, let y_i be the response and let ϵ_i be the mean zero noise. The regression function $f(\cdot)$ we estimate is the sum of univariate functions $f_k(\cdot)$ in each variable dimension and a scalar offset μ. We impose additional constraints that each function $f_k(\cdot)$ is convex, which can be represented by its supporting hyperplanes, i.e.,

$$f_{i'k} \geq f_{ik} + \beta_{ik}(x_{i'k} - x_{ik}) \quad \text{for all } i, i' = 1, \ldots, n, \tag{2.4.1}$$

where $f_{ik} \coloneqq f_k(x_{ik})$ is the function value and β_{ik} is a subgradient at point x_{ik}. This ostensibly requires $O(n^2 p)$ constraints to impose the supporting hyperplane constraints. In fact, only $O(np)$ constraints suffice, since univariate convex functions are characterized by the condition that the subgradient, which is a scalar, must increase monotonically. This observation leads to the optimization

$$\begin{aligned}
\min_{\{f_k, \beta_k\}, \mu} \quad & \frac{1}{2n} \sum_{i=1}^{n} \Big(y_i - \mu - \sum_{k=1}^{p} f_{ik} \Big)^2 + \lambda \sum_{k=1}^{p} \|f_k\|_\infty \\
\text{subject to} \quad & \text{for all } k = 1, \ldots, p\text{:} \\
& f_{\pi_k(i+1)k} = f_{\pi_k(i)k} + \beta_{\pi_k(i)k}(x_{\pi_k(i+1)k} - x_{\pi_k(i)k}), \text{ for } i = 1, \ldots, n-1 \\
& \sum_{i=1}^{n} f_{ik} = 0, \\
& \beta_{\pi_k(i+1)k} \geq \beta_{\pi_k(i)k} \text{ for } i = 1, \ldots, n-2.
\end{aligned} \tag{2.4.2}$$

Here f_k denotes the vector $f_k = (f_{1k}, f_{2_k} \ldots, f_{nk})^T \in \mathbb{R}^n$ and $\{\pi_k(1), \pi_k(2), \ldots, \pi_k(n)\}$ are the indices in the sorted ordering of the values of coordinate k:

$$x_{\pi_k(1)k} \leq x_{\pi_k(2)k} \leq \cdots \leq x_{\pi_k(n)k}. \tag{2.4.3}$$

We can solve for μ explicitly as $\mu = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}$. This follows from the KKT conditions and the constraints $\sum_i f_{ik} = 0$.

The sparse convex additive model optimization in (2.4.2) is a quadratic program with $O(np)$ variables and $O(np)$ constraints. Directly applying a QP solver for f and β is computationally expensive for relatively large n and p. However, notice that variables in different feature dimensions are only coupled in the squared error

term $(y_i - \mu - \sum_{k=1}^{p} f_{ik})^2$. Hence, we can apply the block coordinate descent method, where in each step we solve the following QP subproblem for $\{f_k, \beta_k\}$ with the other variables fixed. In matrix notation, the optimization is

$$\begin{aligned} \min_{f_k, \beta_k, \gamma_k} \quad & \frac{1}{2n}\|r_k - f_k\|_2^2 + \lambda\gamma_k \\ \text{such that} \quad & P_k f_k = \operatorname{diag}(P_k \boldsymbol{x}_k)\beta_k \\ & D_k \beta_k \leq 0 \\ & -\gamma_k \mathbf{1}_n \leq f_k \leq \gamma_k \mathbf{1}_n \\ & \mathbf{1}_n^{\mathsf{T}} f_k = 0 \end{aligned} \tag{2.4.4}$$

where $\beta_k \in \mathbb{R}^{n-1}$ is the vector $\beta_k = (\beta_{1k}, \ldots, \beta_{(n-1)k})^T$, and $r_k \in \mathbb{R}^n$ is the residual vector $r_k = (y_i - \widehat{\mu} - \sum_{k' \neq k} f_{ik'})^T$. In addition, $P_k \in \mathbb{R}^{(n-1)\times n}$ is a permutation matrix where the i-th row is all zeros except for the value -1 in position $\pi_k(i)$ and the value 1 in position $\pi_k(i+1)$, and $D_k \in \mathbb{R}^{(n-2)\times(n-1)}$ is another permutation matrix where the i-th row is all zeros except for a value 1 in position $\pi_k(i)$ and a value -1 in position $\pi_k(i+1)$. We denote by $\operatorname{diag}(v)$ the diagonal matrix with diagonal entries v. The extra variable γ_k is introduced to impose the regularization penalty involving the ℓ_∞ norm.

This QP subproblem involves $O(n)$ variables, $O(n)$ constraints and a sparse structure, which can be solved efficiently using optimization packages. In our experiments we use MOSEK (www.mosek.com). We cycle through all covariates k from 1 to p multiple times until convergence. Empirically, we observe that the algorithm converges in only a few cycles. We also implemented an ADMM solver for (2.4.2) (Boyd et al., 2011), but found that it is not as efficient as this blockwise QP solver.

After optimization, the function estimate for an input vector $\boldsymbol{x}$ is, according to (2.4.1),

$$\widehat{f}(\boldsymbol{x}) = \sum_{k=1}^{p} \widehat{f}_k(x_k) + \widehat{\mu} = \sum_{k=1}^{p} \max_i \Big\{ \widehat{f}_{ik} + \widehat{\beta}_{ik}(x_k - x_{ik}) \Big\} + \widehat{\mu}. \tag{2.4.5}$$

The univariate concave function estimation required in the DC stage is a straightforward modification of optimization (2.4.4). It is only necessary to modify the linear inequality constraints so that the subgradients are non-increasing: $\beta_{\pi_k(i+1)k} \leq \beta_{\pi_k(i)k}$.

Alternative Formulation

Optimization (2.4.2) can be reformulated in terms of the second derivatives. The alternative formulation replaces the order constraints $\beta_{\pi_k(i+1)k} \geq \beta_{\pi_k(i)k}$ with positivity constraints, which simplifies the analysis.

Define $d_{\pi_k(i)k}$ as the second derivative: $d_{\pi_k(1)k} = \beta_{\pi_k(1)k}$, and $d_{\pi_k(i)k} = \beta_{\pi_k(i)k} - \beta_{\pi_k(i-1)k}$ for $i > 1$. The convexity constraint is equivalent to the constraint that $d_{\pi_k(i)k} \geq 0$ for all $i > 1$.

It is easy to verify that $\beta_{\pi_k(i)k} = \sum_{j \leq i} d_{\pi_k(j)k}$ and

$$\begin{aligned}
f_k(x_{\pi_k(i)k}) =& f_k(x_{\pi_k(i-1)k}) + \beta_{\pi_k(i-1)k}(x_{\pi_k(i)k} - x_{\pi_k(i-1)k}) \\
=& f_k(x_{\pi_k(1)k}) + \sum_{j<i} \beta_{\pi_k(j)k}(x_{\pi_k(j+1)k} - x_{\pi_k(j)k}) \\
=& f_k(x_{\pi_k(1)k}) + \sum_{j<i} \sum_{j' \leq j} d_{\pi_k(j')k}(x_{\pi_k(j+1)k} - x_{\pi_k(j)k}) \\
=& f_k(x_{\pi_k(1)k}) + \sum_{j'<i} d_{\pi_k(j')k} \sum_{i>j\geq j'} (x_{\pi_k(j+1)k} - x_{\pi_k(j)k}) \\
=& f_k(x_{\pi_k(1)k}) + \sum_{j'<i} d_{\pi_k(j')k}(x_{\pi_k(i)k} - x_{\pi_k(j')k}).
\end{aligned}$$

We can write this more compactly in matrix notation as

$$\begin{bmatrix} f_k(x_{1k}) \\ f_k(x_{2k}) \\ \vdots \\ f_k(x_{nk}) \end{bmatrix} = \begin{bmatrix} (x_{1k} - x_{\pi_k(1)k})_+ & \cdots & (x_{1k} - x_{\pi_k(n-1)k})_+ \\ \cdots & & \\ (x_{nk} - x_{\pi_k(1)k})_+ & \cdots & (x_{nk} - x_{\pi_k(n-1)k})_+ \end{bmatrix} \begin{bmatrix} d_{\pi_k(1)k} \\ \cdots \\ d_{\pi_k(n-1)k} \end{bmatrix} + \mu_k$$

$$\equiv \Delta_k d_k + \mu_k \tag{2.4.6}$$

where Δ_k is a $n \times n-1$ matrix such that $\Delta_k(i,j) = (x_{ik} - x_{\pi_k(j)k})_+$, $d_k = (d_{\pi_k(1)k}, \ldots, d_{\pi_k(n-1)k})$, and $\mu_k = f_k(x_{\pi_k(1)k})\mathbf{1}_n$. Because f_k has to be centered, $\mu_k = -\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\Delta_k d_k$, and therefore

$$\Delta_k d_k + \mu_k = \Delta_k d_k - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\Delta_k d_k = \bar{\Delta}_k d_k \tag{2.4.7}$$

where $\bar{\Delta}_k \equiv \Delta_k - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\Delta_k$ is Δ_k with the mean of each column subtracted.

The above derivations prove the following proposition, which states that (2.4.2) has an alternative formulation.

Proposition 2.4.1. *Let $\{\widehat{f}_k, \widehat{\beta}_k\}_{k=1,\ldots,p}$ be an optimal solution to* (2.4.2) *and suppose $\bar{Y} = 0$. Define vectors $\widehat{d}_k \in \mathbb{R}^{n-1}$ such that $\widehat{d}_{\pi_k(1)k} = \widehat{\beta}_{\pi_k(1)k}$ and $\widehat{d}_{\pi_k(i)k} = \widehat{\beta}_{\pi_k(i)k} - \widehat{\beta}_{\pi_k(i-1)k}$ for $i > 1$. Then $\widehat{f}_k = \bar{\Delta}_k\widehat{d}_k$ and $\widehat{d}_k$ is an optimal solution to the following optimization:*

$$\begin{aligned}
\min_{\{d_k \in \mathbb{R}^{n-1}\}_{k=1,\ldots p}} \quad & \frac{1}{2n}\left\|Y - \sum_{k=1}^p \bar{\Delta}_k d_k\right\|_2^2 + \lambda_n \sum_{k=1}^p \|\bar{\Delta}_k d_k\|_\infty \\
\textit{such that} \quad & d_{\pi_k(2)k}, \ldots, d_{\pi_k(n-1)k} \geq 0 \qquad \textit{(convexity)}.
\end{aligned} \tag{2.4.8}$$

Likewise, suppose $\{\widehat{d}_k\}_{k=1,\dots p}$ *is a solution to* (2.4.8)*, define* $\widehat{\beta}_{\pi_k(i)k} = \sum_{j\le i} \widehat{d}_{\pi_k(j)k}$ *and* $\widehat{f}_k = \bar{\Delta}_k \widehat{d}_k$*. Then* $\{\widehat{f}_k, \widehat{\beta}_k\}_{k=1,\dots,p}$ *is an optimal solution to* (2.4.2)*.* $\bar{\Delta}$ *is the* n *by* $n-1$ *matrix defined by* (2.4.7)*.*

The decoupled concave postprocessing stage optimization is again similar. Specifically, suppose $\widehat{d}_k$ is the output of optimization (2.4.8), and define the residual vector

$$\widehat{r} = Y - \sum_{k=1}^{p} \bar{\Delta}_k \widehat{d}_k. \tag{2.4.9}$$

Then for all k such that $\widehat{d}_k = 0$, the DC stage optimization is formulated as

$$\begin{aligned} \min_{c_k} \ & \frac{1}{2n}\left\|\widehat{r} - \Delta_k c_k\right\|_2^2 + \lambda_n \|\Delta_k c_k\|_\infty \\ \text{such that} \ & c_{\pi_k(2)k}, \dots, c_{\pi_k(n-1)k} \le 0 \qquad \text{(concavity)}. \end{aligned} \tag{2.4.10}$$

We can use either the off-centered Δ_k matrix or the centered $\bar{\Delta}_k$ matrix because the concave estimations are decoupled and hence are not subject to non-identifiability under additive constants.

2.5 Analysis of Variable Screening Consistency

Our goal is to show that variable screening consistency. That is, as $n, p \to \infty$, $\mathbb{P}(\widehat{S} = S)$ approaches 0 where $\widehat{S}$ is the set of variables outputted by AC/DC in the finite sample setting (Figure 2.4) and S is the set of variables outputted in the population setting (2.3.25).

We divide our analysis into two parts. We first establish a sufficient deterministic condition for consistency of the sparsity pattern screening procedure. We then consider the stochastic setting and argue that the deterministic conditions hold with high probability. Note that in all of our results and analysis, we let c, C represent absolute constants; the actual values of c, C may change from line to line. We derived two equivalent optimizations for AC/DC: (2.4.2) outputs $\widehat{f}_k, \widehat{g}_k$ and (2.4.8) outputs the second derivatives $\widehat{d}_k$. Their equivalence is established in Proposition 2.4.1 and we use both $\widehat{d}_k$ and $\widehat{f}_k$ in our analysis. We will also assume in this section that the true regression function f_0 has mean-zero and therefore, we will omit the intercept term $\widehat{\mu}$ from our estimation procedure.

In our analysis, we assume that an upper bound B to $\|\widehat{f}_k\|_\infty$ is imposed in the optimization procedure, where B is chosen to also upper bound $\|f_k^*\|_\infty$ (same B as in Assumption A3 in Section 2.5). This B-boundedness constraint is added so that

we may use the convex function bracketing results from Kim and Samworth (2014) to establish uniform convergence between the empirical risk and the population risk. We emphasize that this constraint is not needed in practice and we do not use it for any of our simulations.

Deterministic Setting

We analyze Optimization 2.4.8 and construct an additive convex solution $\{\widehat{d}_k\}_{k=1,\ldots,p}$ that is zero for $k \in S^c$, where S is the set of relevant variables, and show that it satisfies the KKT conditions for optimality of optimization (2.4.8). We define $\widehat{d}_k$ for $k \in S$ to be a solution to the restricted regression (defined below). We also show that $\widehat{c}_k = 0$ satisfies the optimality condition of optimization (2.4.10) for all $k \in S^c$.

Definition 2.5.1. We define the *restricted regression* problem

$$\min_{d_k} \frac{1}{n}\left\|Y - \sum_{k\in S} \bar{\Delta}_k d_k\right\|_2^2 + \lambda_n \sum_{k\in S} \|\bar{\Delta}_k d_k\|_\infty \quad \text{such that } d_{\pi_k(2)k}, \ldots, d_{\pi_k(n-1)k} \geq 0$$

where we restrict the indices k in optimization (2.4.8) to lie in some set S which contains the true relevant variables.

Theorem 2.5.1 (Deterministic setting)**.** *Let $\{\widehat{d}_k\}_{k\in S}$ be a minimizer of the restricted regression as defined above. Let $\widehat{r} \coloneqq Y - \sum_{k\in S} \bar{\Delta}_k \widehat{d}_k$ be the restricted regression residual.*

Let $\pi_k(i)$ be a reordering of X_k in ascending order so that $X_{\pi_k(n)k}$ is the largest entry. Let $\mathbf{1}_{\pi_k(i:n)}$ be 1 on the coordinates $\pi_k(i), \pi_k(i+1), ..., \pi_k(n)$ and 0 elsewhere. Define $\mathsf{range}_k = X_{\pi_k(n)k} - X_{\pi_k(1)k}$.

Suppose for all $k \in S^c$, for all $i = 1, \ldots, n$, $\lambda_n > \mathsf{range}_k |\frac{32}{n} \widehat{r}^\mathsf{T} \mathbf{1}_{\pi_k(i:n)}|$. Suppose also that for all $k \in S^c$, $\max_{i=1,\ldots,n-1} \frac{X_{\pi_k(i+1)k} - X_{\pi_k(i)k}}{\mathsf{range}_k} \leq \frac{1}{16}$, and $\mathsf{range}_k \geq 1$.

Then the following two statements hold.

1. *Let $\widehat{d}_k = 0$ for $k \in S^c$. Then $\{\widehat{d}_k\}_{k=1,\ldots,p}$ is an optimal solution to optimization (2.4.8). Furthermore, any solution to the optimization program (2.4.8) must be zero on S^c.*

2. *For all $k \in S^c$, the solution $\widehat{c}_k$ to optimization (2.4.10) must be zero and unique.*

Theorem 2.5.1 states that the estimator produces no false positive so long as λ_n upper bounds the partial sums of the residual $\widehat{r}$ and that the maximum gap between ordered values of X_k is small.

This result holds regardless of whether or not we impose the boundedness conditions in optimization (2.4.8) and (2.4.10). The full proof of Theorem 2.5.1 is in Section 2.7 of the Appendix. We allow S in Theorem 2.5.1 to be any set containing the relevant variables; in Lasso analysis, S is taken to be the set of relevant variables; we will take S to be the set of variables chosen by the additive convex and decoupled concave procedure in the population setting, which is guaranteed to contain the relevant variables because of additive faithfulness.

Theorem 2.5.1 allows us to separately analyze the false negative rates and false positive rates. To control false positives, Theorem 2.5.2 verifies that the conditions in Theorem 2.5.1 hold in a stochastic setting. To control false negatives, Theorem 2.5.3 analyzes the restricted regression with only $|S|$ variables.

The proof of Theorem 2.5.1 analyses the KKT conditions of optimization (2.4.8). This parallels the now standard *primal-dual witness* technique (Wainwright, 2009). The conditions in Theorem 2.5.1 is our analogue to the *mutual incoherence* condition. Our conditions are much more strict however because the estimation is nonparametric—even the low dimensional restricted regression has $s(n-1)$ variables.

The details of the proof are given in Section 2.7 of the Appendix.

Probabilistic Setting

In the probabilistic setting we treat the covariates as random. We adopt the following standard setup:

1. The data $X^{(1)}, \ldots, X^{(n)} \sim P$ are iid from a distribution P with a density $p(\mathbf{x})$ that is supported on $\mathcal{X} = [-1, 1]^p$.

2. The response is $Y = f_0(X) + W$ where W is independent, zero-mean noise; thus $Y^{(i)} = f_0(X^{(i)}) + W^{(i)}$.

3. The regression function f_0 satisfies $f_0(X) = f_0(X_{S_0})$, where $S_0 = \{1, \ldots, s_0\}$ is the set of relevant variables.

Let $\mathcal{C}^1$ denote the set of univariate convex functions supported on $[-1, 1]$, and let $\mathcal{C}_1^p$ denote the set of convex additive functions $\mathcal{C}_1^p \equiv \{f : f = \sum_{k=1}^p f_k,\ f_k \in \mathcal{C}^1\}$. Let $f^*(\mathbf{x}) = \sum_{k=1}^p f_k^*(x_k)$ be the population risk minimizer in $\mathcal{C}_1^p$,

$$f^* = \arg\min_{f \in \mathcal{C}_1^p} \mathbb{E}\big(f_0(X) - f(X)\big)^2. \tag{2.5.1}$$

f^* is the unique minimizer by Theorem 2.3.3. Similarly, we define -$\mathcal{C}^1$ as the set of univariate concave functions supported on $[-1, 1]$ and define

$$g_k^* = \arg\min_{g_k \in \text{-}\mathcal{C}^1} \mathbb{E}\big(f_0(X) - f^*(X) - g_k(X_k)\big)^2. \tag{2.5.2}$$

The $\widehat{g}_k$'s are unique minimizers as well. We let $S = \{k = 1, \ldots, p : f_k^* \neq 0 \text{ or } g_k^* \neq 0\}$ and let $s = |S|$. By additive faithfulness (Theorem 2.3.3), it must be that $S_0 \subset S$ and thus $s \geq s_0$. In some cases, such as when $X_{S_0}, X_{S_0^c}$ are independent, we have $S = S_0$. Each of our theorems will use a subset of the following assumptions:

A1: X_S, X_{S^c} are independent.

A2: f_0 is convex with a bounded second derivative on an open set around $[-1, 1]^p$. $\mathbb{E}f_0(X) = 0$.

A3: $\|f_0\|_\infty \leq sB$ and $\|f_k^*\|_\infty \leq B$ for all k.

A4: W is mean-zero sub-Gaussian, independent of X, with scale σ; i.e., for all $t \in \mathbb{R}$, $\mathbb{E}e^{t\epsilon} \leq e^{\sigma^2 t^2/2}$.

A5: The density $p(\mathbf{x})$ satisfies the boundary flatness condition (Definition 2.3.2), and $0 < c_l \leq \inf p(\mathbf{x}) \leq \sup p(\mathbf{x}) \leq c_u < \infty$ for two constants c_l, c_u.

By assumption A1, f_k^* must be zero for $k \notin S$. We define α_+, α_- as a measure of the signal strength of the weakest variable:

$$\begin{aligned} \alpha_+ &= \min_{f \in \mathcal{C}_1^p \,:\, \text{supp}(f) \subsetneq \text{supp}(f^*)} \Big\{\mathbb{E}\big(f_0(X) - f(X)\big)^2 - \mathbb{E}\big(f_0(X) - f^*(X)\big)^2\Big\} \\ \alpha_- &= \min_{k \in S \,:\, g_k^* \neq 0} \Big\{\mathbb{E}\big(f_0(X) - f^*(X)\big)^2 - \mathbb{E}\big(f_0(X) - f^*(X) - g_k^*(X_k)\big)^2\Big\} \end{aligned} \tag{2.5.3}$$

α_+ is a lower bound on the excess risk incurred by any additive convex function whose support is strictly smaller than f^*. α_+ is achieved by some $f \neq f^*$ because the set $\{f \in \mathcal{C}_1^p : \text{supp}(f) \subsetneq \text{supp}(f^*)\}$ is a finite union of closed convex sets. $\alpha_+ > 0$ since f^* is the unique risk minimizer. Likewise, α_- lower bounds the excess risk of any decoupled concave fit of the residual $f_0 - f^*$ that is strictly more sparse than the optimal decoupled concave fit $\{\widehat{g}_k^*\}$; $\alpha_- > 0$ by the uniqueness of $\{g_k^*\}$ as well. These quantities play the role of the absolute value of the smallest nonzero coefficient in the true linear model in lasso theory. Intuitively, if α_+ is small, then it is easier to make a false omission in the additive convex stage of the procedure. If α_- is small, then it is easier to make a false omission in the decoupled concave stage of the procedure. If X is independent, then α_+ can be simplified to $\min_{k \,:\, f_k^* \neq 0} \mathbb{E}f_k^*(X)^2$ and α_- becomes unnecessary.

Remark 2.5.1. We make strong assumptions on the covariates in A1 in order to make weak assumptions on the true regression function f_0 in A2. In particular, we do not assume that f_0 is additive. An important direction for future work is to weaken assumption A1. Our simulation experiments indicate that the procedure can be effective even when the relevant and irrelevant variables are correlated.

Theorem 2.5.2 (Controlling false positives)**.** *Suppose assumptions A1-A5 hold. Define* $\widetilde{\sigma} \equiv \max(\sigma, B)$ *and define* $\mathsf{range}_k = X_{\pi_k(n)k} - X_{\pi_k(1)k}$. *Suppose* $p \leq O\big(\exp(cn)\big)$ *and* $n \geq C$ *for some positive constants* C *and* $0 < c < \frac{c_l}{32}$. *Suppose also*

$$\lambda_n \geq 768 s \widetilde{\sigma} \sqrt{\frac{\log^2 np}{n}}. \tag{2.5.4}$$

Then with probability at least $1 - \frac{24}{n}$, *for all* $k \in S^c$, *for all* $i = 1, \ldots, n$,

$$\lambda_n \geq \mathsf{range}_k \left| \frac{32}{n} \widehat{r}^{\mathsf{T}} \mathbf{1}_{(i':n)_k} \right|, \tag{2.5.5}$$

$\max_{i'} \frac{X_{\pi_k(i'+1)k} - X_{\pi_k(i')k}}{\mathsf{range}_k} \leq \frac{1}{16}$, $\mathsf{range}_k \geq 1$, *and both the AC solution* $\widehat{f}_k$ *from optimization* (2.4.8) *and the DC solution* $\widehat{g}_k$ *from optimization* (2.4.10) *are zero.*

The proof of Theorem 2.5.2 exploits independence of $\widehat{r}$ and X_k under assumption A1; when $\widehat{r}$ and X_k are independent, $\widehat{r}^{\mathsf{T}} \mathbf{1}_{(i':n)}$ is the sum of $n - i' + 1$ random coordinates of $\widehat{r}$. We can then use concentration of measure results for sampling without replacement to argue that $|\frac{1}{n} \widehat{r}^{\mathsf{T}} \mathbf{1}_{(i':n)}|$ is small with high probability. The result of Theorem 2.5.1 is then used. The full proof of Theorem 2.5.2 is in Section 2.7 of the Appendix.

Theorem 2.5.3 (Controlling false negatives)**.** *Suppose assumptions A1-A5 hold. Let* $\widehat{f}$ *be any AC solution to the restricted regression with* B*-boundedness constraint, and let* $\widehat{g}_k$ *be any DC solution to the restricted regression with* B*-boundedness constraint. Let* $\widetilde{\sigma}$ *denote* $\max(\sigma, B)$. *Suppose*

$$\lambda_n \leq 768 s \widetilde{\sigma} \sqrt{\frac{\log^2 np}{n}} \tag{2.5.6}$$

and that n *is sufficiently large so that, for some constant* $c' > 1$,

$$\frac{n^{4/5}}{\log np} \geq c' B^4 \widetilde{\sigma}^2 s^5. \tag{2.5.7}$$

Assume that the signal-to-noise ratio satisfies

$$\frac{\alpha_+}{\widetilde{\sigma}} \geq cB^2\sqrt{\frac{s^5 c_u^{1/2}}{n^{4/5}}\log^2 np} \tag{2.5.8}$$

$$\frac{\alpha_-^2}{\widetilde{\sigma}} \geq cB^2\sqrt{\frac{s^5 c_u^{1/2}}{n^{4/5}}\log^2 np} \tag{2.5.9}$$

where c is a constant. Then with probability at least $1 - \frac{C}{n}$ for some constant C, $\widehat{f}_k \neq 0$ or $\widehat{g}_k \neq 0$ for all $k \in S$.

This is a finite sample version of Theorem 2.3.1. We need stronger assumptions in Theorem 2.5.3 to use our additive faithfulness result, Theorem 2.3.1. The full proof of Theorem 2.5.3 is in Section 2.7 of the appendix.

Combining Theorems 2.5.2 and 2.5.3 we obtain the following result.

Corollary 2.5.1. *Suppose the assumptions of Theorem 2.5.2 and Theorem 2.5.3 hold. Then with probability at least $1 - \frac{C}{n}$*

$$\widehat{f}_k \neq 0 \text{ or } \widehat{g}_k \neq 0 \text{ for all } k \in S \tag{2.5.10}$$

$$\widehat{f}_k = 0 \text{ and } \widehat{g}_k = 0 \text{ for all } k \notin S \tag{2.5.11}$$

for some constant C.

The above corollary implies that consistent variable selection is achievable with an exponential scaling of the ambient dimension scaling, $p = O(\exp(cn))$ for some $0 < c < 1$, just as in parametric models. The cost of nonparametric modeling through shape constraints is reflected in the scaling with respect to the number of relevant variables, which can scale as $s = o(n^{4/25})$.

Remark 2.5.2. Comminges and Dalalyan (2012) have shown that under tradtional smoothness constraints, even with a product distribution, variable selection is achievable only if $n > O(e^{s_0})$. It is interesting to observe that because of additive faithfulness, the convexity assumption enables a much better scaling of $n = O(\text{poly}(s_0))$, demonstrating that geometric constraints can be quite different from the previously studied smoothness conditions.

2.6 Experiments

We perform both synthetic and real data experiments.

Simulations

We first illustrate our methods using a simulation of the model

$$Y_i = f_o(x_{iS}) + \epsilon_i \quad (i = 1, 2, \ldots, n).$$

Here x_i denotes data sample i drawn from some distribution P. f_0 is the true regression function. x_{iS} is a subset of x_i with dimension $|S| = s$, where S represents the set of relevant variables, and ϵ_i is additive noise drawn from $\mathcal{N}(0, \sigma)$. For all simulations, we set σ such that the signal-to-noise ratio(SNR, $\frac{\text{std}(Y)}{\sigma}$) is 5. Also, for all simulations except the sixth, we choose the set of relevant variables S uniformly at random among all variables $\{1, ..., p\}$.

We study both the independent case where $P = N(0, I_p)$ and the correlated case where P is a correlated Gaussian Copula modified slightly to satisfy the boundary flatness condition. We measure the probability of exact selection in the independent case and the probability of screening in the correlated case. We also study both cases where the regression function is parametric (quadratic) and cases where the regression function is nonparametric (softmax of linear forms). In all our experiments, we mark a variable as selected if either the AC estimate $\|\widehat{f}_j\|_\infty$ or the DC estimate $\|\widehat{g}_k\|_\infty$ is larger than 10^{-6}. We set $\lambda = 0.5\sqrt{\frac{\log^2 np}{n}}$ for all the simulations.

For the first three simulations, we will use the quadratic form as our true regression function.

$$f_0(x_{iS}) = x_{iS}^{\mathsf{T}} Q x_{iS}$$

The matrix Q is a symmetric positive definite matrix of dimension $s \times s$. Note that if Q is diagonal, then the true function is convex additive; otherwise the true function is convex but not additive.

In the **first simulation** (Figure 2.6a), we vary the ambient dimension p. We set Q as one on the diagonal and $1/2$ on the off-diagonal with 0.5 probability, set $s = 5$, and $p = 64, 128, 256$ and 512. We draw $X \sim N(0, I_p)$. For each (n, p) combination, we generate 100 independent trials. In Figure 2.6(a), we plot the probability of exact support recovery. We observe that the algorithm performs consistent variable selection even if the dimensionality is large. To give the reader a sense of the running speed, for a data set with $n = 1000$ and $p = 512$, the code runs in roughly two minutes on a machine with 2.3 GHz Intel Core i5 CPU and 4 GB memory.

In the **second simulation** (Figure 2.6b,c), we vary the sparsity of the Q matrix, that is, we vary the extent to which the true function is non-additive. We generate four Q matrices plotted in Figure 2.6(c), where the diagonal elements are all one and the off-diagonal elements are $\frac{1}{2}$ with probability α ($\alpha = 0, 0.2, 0.5, 1$ for the four cases). We show the 4 Q matrices we used in Figure 2.6(c). We fix $p = 128$, $s = 5$,

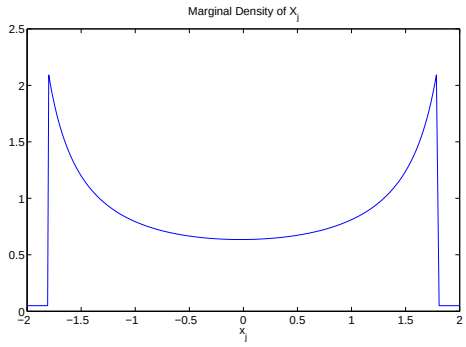


Figure 2.5: Marginal density of the Gaussian Copula and Uniform Mixture

and $X \sim N(0, I_p)$. We again run the AC/DC optimization on 100 independent trials and plot the probability of exact recovery in Figure 2.6(b). The results demonstrate that AC/DC performs consistent variable selection even if the true function is not additive (but still convex).

In the third, fourth, and fifth simulation, we use a correlated design. We generate X from a non-Gaussian boundary flat distribution with covariance Σ. The distribution we used is a mixture of a uniform distribution and a Gaussian Copula.

$$X \sim \gamma U([-2,2]^p) + (1-\gamma)\text{Copula}(0, \Sigma, F)$$

The Gaussian Copula is a way to customize the marginal distributions of a Gaussian random variable while maintaining the same covariance. Gaussian Copula results when one applies a monotone transformation $F^{-1}\Phi$ onto each of the variables of a Gaussian random vector where Φ is the normal CDF and F is the CDF of the new marginal distribution. In all our experiments, we set $\gamma = 0.05$ and set the marginal CDF F so that marginal density of the Copula is bimodal and supported on $[-1.8, 1.8]$. The resulting marginal density of the mixture is shown in Figure 2.5. Notice that boundary flatness holds because the distribution is uniform in the boundary area $[-2,2]^p \backslash [-1.8, 1.8]^p$.

In the **third simulation** (Figure 2.6d,e), we use the non-Gaussian distribution described above and set the covariance $\Sigma_{ij} = \nu^{|i-j|}$ for $\nu = 0, 0.2, 0.5, 0.9$. We use the non-additive Q, same as in the first experiment, with $\alpha = 0.5$ and fix $p = 128, s = 5$. We measure success not through exact recovery but through faithful recovery. We say that a trial is a successful if (1) all relevant variables were recovered and (2) fewer than 20 variables were marked as relevant overall (true sparsisty $s = 5$). We use the same λ as before. The probabilities of success are computed from 40 independent trials and plotted against various values of ν in Figure 2.6(d). Additionally, for $\nu = 0.5$, we show

the number of selected variables versus the sample size as a box-and-whisker plot in Figure 2.6(e). As seen, for small to moderate correlations, AC/DC can successfully recover the relevant variables with only a small number of false positives.

In the fourth and fifth simulation, we use a softmax function as the ground truth

$$f_0(x_{iS}) = \log\left(\sum_{k=1}^{K} \exp(\beta_k^{\mathsf{T}} x_{iS})\right) - \mu \tag{2.6.1}$$

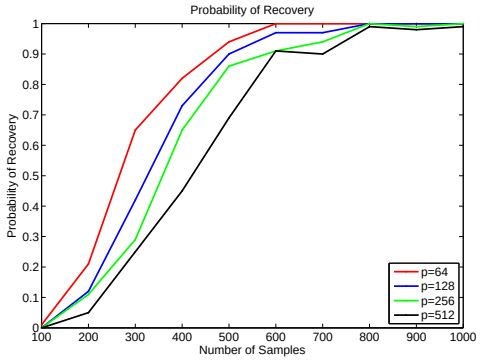
We generate random unit vectors as $\{\beta_k \in \mathbb{R}^s\}_{k=1,\ldots,K}$ and choose μ so that f_0 has mean-zero. We set $K = 7$ for all the experiments.

For the **fourth simulation** (Figure 2.6f,g), we let f_0 be the softmax function and let X be drawn from the boundary flat mixture distribution described earlier with the Toeplitz covariance $\Sigma_{ij} = \nu^{|i-j|}$ for $\nu = 0.5$. We set $s = 5$ and vary $p = 128, 256, 512$. We use the same faithful recovery criteria as the third simulation and plot the probability of faithful recovery against the number of samples in Figure 2.6(f). The probabilities are computed over 30 independent trials. Also, for $p = 256$, we show the number of selected variables versus the sample size as a box-and-whisker plot in Figure 2.6(g). The softmax function is more challenging to estimate than the quadratic function; the softmax function requires about $n > 1500$ to achieve the same success probability as the quadratic function with $n = 1000$. Regardless, we see that increasing the ambient dimension p does not signficantly affect the recovery probability.

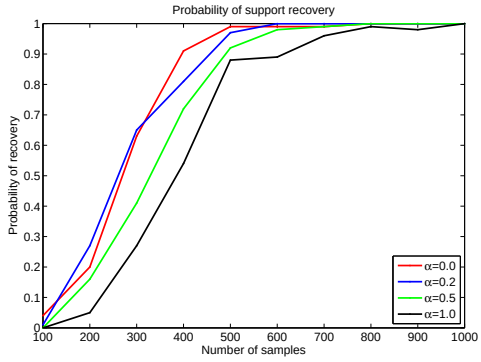
For the **fifth simulation** (Figure 2.7), we compare the variables selected via the AC stage and the variables selected via the DC stage. We use the softmax regression function and X drawn from the boundary flat mixture distribution with a Toeplitz covariance and correlation level $\nu = -0.7$. We set $s = 5, n = 500, p = 128$. We perform 30 independent trials and plot the frequency of variable selection in Figure 2.7. The true variables are X_j for $j = 5, 6, 7, 8, 9, 10$. We plot the frequency of selection among only the first 20 variables, that is, X_j for $j = 1, ..., 20$. We do not plot selection frequencies for variables 21 to 128 because they are almost never selected by either AC or DC. As can be seen, the DC stage is slightly helpful in recovering the true variables but its effect is not significant. We thus believe that the DC stage, though important in theory, is not as important in practice; it may be omitted without significant detriment to the overall result.

Boston housing data

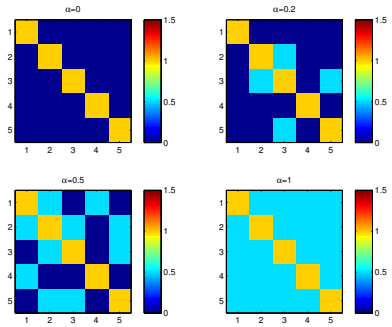
We next use the Boston housing data rather than simulated data. This data set contains 13 covariates, 506 samples and one response variable indicating housing



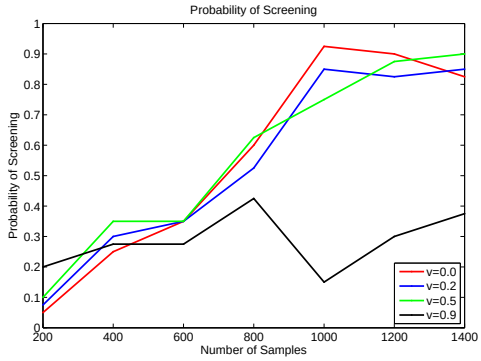
(a) quadratic f_0, independent X, varying p



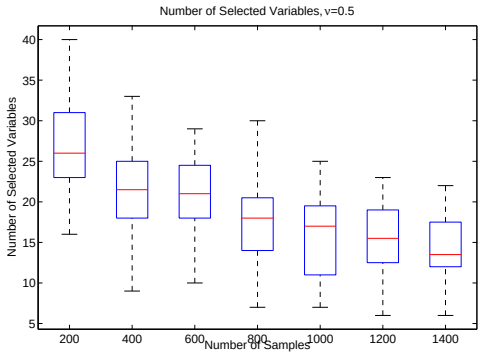
(b) quadratic f_0, independent X, varying Q



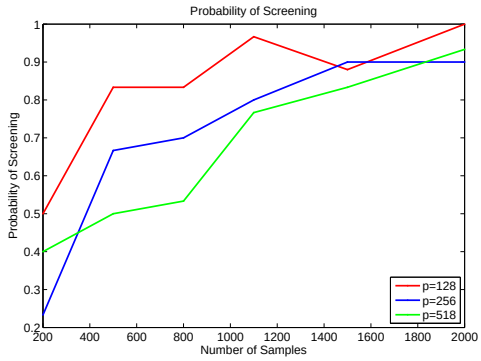
(c) four Q matrices used in (b)



(d) quadratic f_0, correlated X, varying corr ν



(e) recovered support size for $\nu = 0.5$ in (d)



(f) softmax f_0, correlated X, varying p

Figure 2.6: Support recovery results.

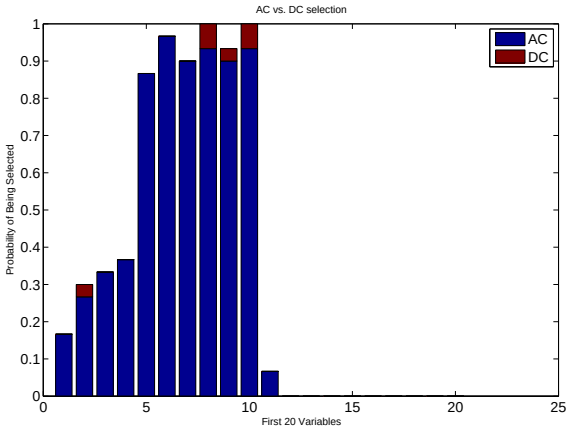


Figure 2.7: Frequency of variable selection among the first 20 variables (X_j for $j = 1, ..., 20$) in the AC stage vs. in the DC stage. The true variables are [5,6,7,8,9,10].

values in suburbs of Boston. The data and detailed description can be found on the UCI Machine Learning Repository website[1].

We first use all $n = 506$ samples (with standardization) in the AC/DC algorithm, using a set of candidate regularization parameters $\{\lambda^{(t)}\}$ ranging from $\lambda^{(1)} = 0$ (no regularization) to 2. For each $\lambda^{(t)}$ we obtain a function value matrix $f^{(t)}$ with $p = 13$ columns and $n = 506$ rows. The non-zero columns in this matrix indicate the variables selected using $\lambda^{(t)}$.

In Figure 2.8(a), we plot on the y-axis the norm $\|f_j^{(t)}\|_\infty$ of every column j against the regularization strength $\lambda^{(t)}$. Instead of plotting the value of $\lambda^{(t)}$ on the x-axis however, we plot the total norm at $\lambda^{(t)}$ normalized against the total norm at $\lambda^{(1)}$: $\frac{\sum_j \|f_j^{(t)}\|_\infty}{\sum_j \|f^{(1)}\|_\infty}$. Thus, as x-axis moves from 0 to 1, the regularization goes from strong to weak. For comparison, we plot the LASSO/LARS result in a similar way in Figure 2.8(b). From the figures we observe that the first three variables selected by AC/DC and LASSO are the same: `LSTAT`, `RM` and `PTRATIO`, consistent with previous findings (Ravikumar et al., 2007). The fourth variable selected by AC/DC is `INDUS` (with $\lambda^{(t)} = 0.7$). We then refit AC/DC with only these four variables without regularization, and plot the estimated additive functions in Figure 2.8(d). When refitting, we constrain a component to be convex if it is non-zero in the AC stage and concave if it is non-zero in the DC stage. As can be seen, these functions contain clear nonlinear effects which cannot be captured by LASSO. The shapes of these

[1] `http://archive.ics.uci.edu/ml/datasets/Housing`

functions, including the concave shape of the PTRATIO function, are in agreement with those obtained by SpAM (Ravikumar et al., 2007).

Next, in order to quantitatively study the predictive performance, we run 3 times 5-fold cross validation, following the same procedure described above—training, variable selection and refitting. A plot of the mean and standard deviation of the predictive mean squared error (MSE) is shown in Figure 2.8(c). Since for AC/DC the same regularization level $\lambda^{(t)}$ may lead to a slightly different number of selected variables in different folds and runs, the values on the x-axis for AC/DC are not necessarily integers. The figure clearly shows that AC/DC has a lower predictive MSE than LASSO. We also compared the performance of AC/DC with that of Additive Forward Regression (AFR) presented in Liu and Chen (2009), and found that they are similar. The main advantages of AC/DC compared with AFR and SpAM are that there are no smoothing parameters required, and the optimization is formulated as a convex program, guaranteeing a global optimum.

2.7 Supplement: Proofs of Technical Results

Proof of the Deterministic Condition for Sparsistency

We restate Theorem 2.5.1 first for convenience. The following holds regardless of whether we impose the B-boundedness condition (see discussion at the beginning of Section 2.5 for definition of the B-boundedness condition).

Theorem 2.7.1. *Let $\{\widehat{d}_k\}_{k\in S}$ be a minimizer of the restricted regression, that is, the solution to optimization (2.4.8) where we restrict $k \in S$. Let $\widehat{r} \coloneqq Y - \sum_{k\in S} \bar{\Delta}_k \widehat{d}_k$ be the restricted regression residual.*

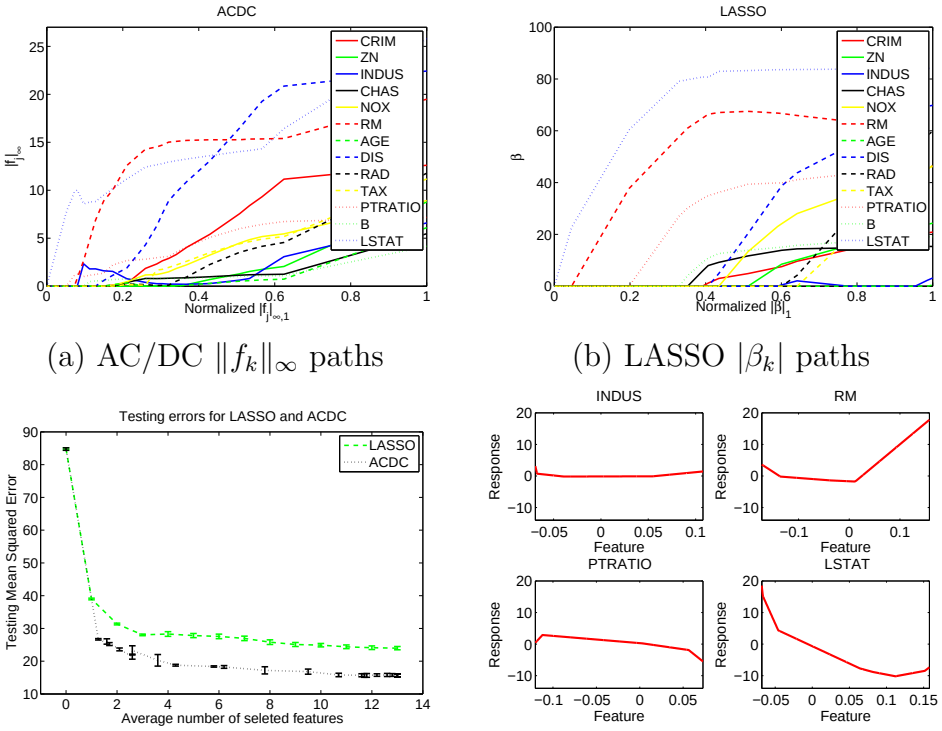
Let $\pi_k(i)$ be an reordering of X_k in ascending order so that $X_{\pi_k(n)k}$ is the largest entry. Let $\mathbf{1}_{\pi_k(i:n)}$ be 1 on the coordinates $\pi_k(i), \pi_k(i+1), ..., \pi_k(n)$ and 0 elsewhere. Define $\mathsf{range}_k = X_{\pi_k(n)k} - X_{\pi_k(1)k}$.

Suppose for all $k \in S^c$ and for all $i = 1, ..., n$, $\lambda_n \geq \mathsf{range}_k |\frac{32}{n} \widehat{r}^{\mathsf{T}} \mathbf{1}_{\pi_k(i:n)}|$, and $\max_{i=1,...,n-1} \frac{X_{\pi_k(i+1)k} - X_{\pi_k(i)k}}{\mathsf{range}_k} \geq \frac{1}{16}$, and $\mathsf{range}_k \geq 1$.

Then the following are true:

1. *Let $\widehat{d}_k = 0$ for $k \in S^c$, then $\{\widehat{d}_k\}_{k=1,...,p}$ is an optimal solution to optimization 2.4.8. Furthermore, any solution to the optimization program 2.4.8 must be zero on S^c.*

2. *For all $k \in S^c$, the solution to optimization 2.4.10 must be zero and unique.*



(a) AC/DC $\|f_k\|_\infty$ paths (b) LASSO $|\beta_k|$ paths

(c) predictive MSE (d) estimated functions from AC/DC

Figure 2.8: Results on Boston housing data, showing regularization paths, MSE and fitted functions.

Proof. We will omit the B-boundedness constraint in our proof here. It is easy to verify that the result of the theorem still holds with the constraint added in.

We begin by considering the first item in the conclusion of the theorem. We will show that with $\{\widehat{d}_k\}_{k=1,..,p}$ as constructed, we can set the dual variables to satisfy the complementary slackness and stationary conditions: $\nabla_{d_k}\mathcal{L}(\widehat{d}) = 0$ for all k.

The Lagrangian is

$$\mathcal{L}(\{d_k\},\nu) = \frac{1}{2n}\Big\|Y - \sum_{k=1}^{p}\bar{\Delta}_k d_k\Big\|_2^2 + \lambda\sum_{k=1}^{p}\|\bar{\Delta}_k d_k\|_\infty - \sum_{k=1}^{p}\sum_{i=2}^{n-1}\nu_{\pi_k(i)k}d_{\pi_k(i)k} \qquad (2.7.1)$$

with the constraint that $\nu_{\pi_k(i)k} \geq 0$ for all k, i.

Because $\{\widehat{d}_k\}_{k\in S}$ is by definition the optimal solution of the restricted regression, it is a consequence that stationarity holds for $k \in S$, that is, $\partial_{\{d_k\}_{k\in S}}\mathcal{L}(d) = 0$, and that the dual variables ν_k for $k \in S$ satisfy complementary slackness.

We now verify that stationarity holds also for $k \in S^c$. We fix one dimension $k \in S^c$ and let $\widehat{r} = Y - \sum_{k' \in S} \bar{\Delta}_{k'} \widehat{d}_{k'}$.

To ease notational burden, let us reorder the samples $\{X_{ik}\}_{i=1,...,n}$ in ascending order so that the i-th sample is the i-th smallest sample. We will from here on write X_{ik} to denote $X_{\pi_k(i)k}$, d_{ik} to denote $d_{\pi_k(i)k}$, etc.

The Lagrangian form of the optimization, in terms of just d_k, is

$$\mathcal{L}(d_k, \nu_k) = \frac{1}{2n} \Big\| Y - \sum_{k' \in S} \bar{\Delta}_{k'} d_{k'} - \bar{\Delta}_k d_k \Big\|_2^2 + \lambda \|\bar{\Delta}_k d_k\|_\infty - \sum_{i=2}^{n-1} \nu_{ik} d_{ik}$$

with the constraint that $\nu_{ik} \geq 0$ for $i = 2, ..., n-1$.

The derivative of the Lagrangian is

$$\partial_{d_k} \mathcal{L}(d_k) = -\frac{1}{n} \bar{\Delta}_k^\mathsf{T} (Y - \sum_{k' \in S} \bar{\Delta}_{k'} d_{k'} - \bar{\Delta}_k d_k) + \lambda \bar{\Delta}_k^\mathsf{T} \mathbf{u} - \begin{pmatrix} 0 \\ \nu_k \end{pmatrix}$$

where $\mathbf{u}$ is the subgradient of $\|\bar{\Delta}_k d_k\|_\infty$. If $\bar{\Delta}_k d_k = 0$, then $\mathbf{u}$ can be any vector whose L_1 norm is less than or equal to 1. $\nu_k \geq 0$ is a vector of Lagrangian multipliers. ν_{k1} does not exist because d_{k1} is not constrained to be non-negative.

We now substitute in $d_{k'} = \widehat{d}_{k'}$ for $k' \in S$, $d_k = 0$ for $k \in S^c$, and $r = \widehat{r}$ and show that the $\mathbf{u}, \nu_k$ dual variables can be set in a way to ensure that stationarity:

$$\partial_{d_k} \mathcal{L}(\widehat{d}_k) = -\frac{1}{n} \bar{\Delta}_k^\mathsf{T} \widehat{r} + \lambda \bar{\Delta}_k^\mathsf{T} \mathbf{u} - \begin{pmatrix} 0 \\ \nu_k \end{pmatrix} = 0$$

where $\|\mathbf{u}\|_1 \leq 1$ and $\nu_k \geq 0$. It is clear that to show stationarity, we only need to show that $[-\frac{1}{n} \bar{\Delta}_k^\mathsf{T} \widehat{r} + \lambda \bar{\Delta}_k^\mathsf{T} \mathbf{u}]_j = 0$ for $j = 1$ and ≥ 0 for $j = 2, ..., n-1$.

Define i^* as the largest index such that $\frac{X_{nk} - X_{i^*k}}{X_{nk} - X_{1k}} \geq 1/2$. We will construct $\mathbf{u} = (a - a', 0, ..., -a, 0, ..., a')$ where a, a' are positive scalars, where $-a$ lies at the i^*-th coordinate, and where the coordinates of $\mathbf{u}$ correspond to the new sample ordering.

We define

$$\begin{aligned}
\kappa &= \frac{1}{\lambda n (X_{nk} - X_{1k})} \left[\Delta_k^\mathsf{T} \widehat{r}\right]_1 \\
a' &= \frac{X_{nk} - X_{1k}}{X_{nk} - X_{i^*k}} \kappa + \frac{X_{i^*k} - X_{1k}}{X_{nk} - X_{i^*k}} \frac{1}{8} \\
a &= \frac{X_{nk} - X_{1k}}{X_{nk} - X_{i^*k}} \kappa + \frac{X_{nk} - X_{1k}}{X_{nk} - X_{i^*k}} \frac{1}{8}
\end{aligned}$$

and we verify two facts: first that the KKT stationarity is satisfied and second, that $\|\mathbf{u}\|_1 < 1$ with high probability. Our claim is proved immediately by combining these two facts.

Because $\widehat{r}$ and $\mathbf{u}$ are both centered vectors, $\bar{\Delta}_k^\mathsf{T}\widehat{r} = \Delta_k^\mathsf{T}\widehat{r}$ and likewise for $\mathbf{u}$. Therefore, we need only show that for $j = 1$, $\lambda\big[\Delta_k^\mathsf{T}\mathbf{u}\big]_j = \big[\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}\big]_j$ and that for $j = 2, ..., n-1$, $\lambda\big[\Delta_k^\mathsf{T}\mathbf{u}\big]_j \geq \big[\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}\big]_j$.

With our explicitly defined form of $\mathbf{u}$, we can characterize $\Delta_k^\mathsf{T}\mathbf{u}$. Note that under sample reordering, the j-th column of Δ_k is $(0, ..., x_{(j+1)k} - x_{jk},\ x_{(j+2)k} - x_{jk},\ ...,\ x_{nk} - x_{jk})$ where the first j-th entries are all 0.

$$\begin{aligned}
\Delta_k^\mathsf{T}\mathbf{u}]_j &= \sum_{i>j}(X_{ik} - X_{jk})\mathbf{u}_i \\
&= \mathbf{u}_{i^*}(X_{i^*k} - X_{jk})\delta_{i^*\geq j} + \mathbf{u}_n(X_{nk} - X_{jk}) \\
&= -a(X_{i^*k} - X_{jk})\delta_{i^*\geq j} + a'(X_{nk} - X_{jk})
\end{aligned}$$

Simple algebra shows then that

$$\big[\Delta_k^\mathsf{T}\mathbf{u}\big]_j = \begin{cases} (-a + a')(X_{i^*k} - X_{jk}) + a'(X_{nk} - X_{i^*k}) & \text{if } j \leq i^* \\ a'(X_{nk} - X_{jk}) & \text{if } j \geq i^* \end{cases} \tag{2.7.2}$$

It is straightforward to check that $\big[\lambda\Delta_k^\mathsf{T}\mathbf{u}\big]_1 = \lambda(X_{nk} - X_{1k})\kappa = \frac{1}{n}\big[\Delta_k^\mathsf{T}\widehat{r}\big]_1$.

To check that $\lambda[\Delta_k^\mathsf{T}\mathbf{u}]_j \geq [\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}]_j$ for $j > 1$, we first characterize $[\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}]_j$:

$$\begin{aligned}
[\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}]_j &= \frac{1}{n}\sum_{i>j}(X_{ik} - X_{jk})\widehat{r}_i \\
&= \frac{1}{n}\sum_{i>j}\sum_{j<i'\leq i}\mathsf{gap}_{i'}\widehat{r}_i \\
&= \frac{1}{n}\sum_{i'>j}\mathsf{gap}_{i'}\sum_{i\geq i'}\widehat{r}_i \\
&= \frac{1}{n}\sum_{i'>j}\mathsf{gap}_{i'}\mathbf{1}_{i':n}^\mathsf{T}\widehat{r}
\end{aligned}$$

where we denote $\mathsf{gap}_{i'} = X_{i'k} - X_{(i'-1)k}$.

We pause for a second here to give a summary of our proof strategy. We leverage two critical observations: first, any two adjacent coordinates in the vector $\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}$ cannot differ by too much. Second, we defined a, a' such that $-a + a' = -\frac{1}{8}$ so

that $\lambda\Delta_k^\mathsf{T}\mathbf{u}$ is a sequence that strictly increases in the first half (for coordinates in $\{1, ..., i^*\}$) and strictly decreases in the second half.

We know $\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}$ and $\lambda\Delta_k^\mathsf{T}\mathbf{u}$ are equal in the first coordinate. We will show that the second sequence increases faster than the first sequence, which will imply that the second sequence is larger than the first in the first half of the coordinates. We will then work similarly but backwards for the second half.

Following our strategy, we first compare $[\lambda\Delta_k^\mathsf{T}\mathbf{u}]_j$ and $[\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}]_j$ for $j = 1, ..., i^* - 1$.

For $j = 1, ..., i^* - 1$, we have that

$$\begin{aligned}\lambda[\Delta_k^\mathsf{T}\mathbf{u}]_{j+1} - \lambda[\Delta_k^\mathsf{T}\mathbf{u}]_j &= \lambda(a - a')\mathsf{gap}_{j+1} \\ &\geq -\mathsf{gap}_{j+1}\frac{1}{n}\mathbf{1}_{(j+1):n}^\mathsf{T}\widehat{r} \\ &= [\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}]_{j+1} - [\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}]_j\end{aligned}$$

The inequality follows because $a - a' = \frac{1}{8}$ and thus $\lambda(a - a') \geq \left|\frac{1}{n}\mathbf{1}_{(j+1):n}^\mathsf{T}\widehat{r}\right|$. Therefore, for all $j = 1, ..., i^*$:

$$[\lambda\Delta_k^\mathsf{T}\mathbf{u}]_j \geq [\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}]_j$$

For $j \geq i^*$, we start our comparison from $j = n - 1$. First, we claim that $a' > \frac{1}{32}$. To prove this claim, note that

$$|\kappa| = \left|\frac{1}{\lambda n}\sum_{i'>1}\mathsf{gap}_{i'}\mathbf{1}_{i':n}^\mathsf{T}\widehat{r}\right| \leq \frac{1}{(X_{kn} - X_{k1})^2}\frac{1}{32}\sum_{i'>1}\mathsf{gap}_{i'} = \frac{1}{32} \tag{2.7.3}$$

because $\sum_{i'>1}\mathsf{gap}_{i'} = X_{nk} - X_{1k}$ by definition and $X_{nk} - X_{1k} \geq 1$ by assumption of the theorem. We note also that

$$\frac{X_{nk} - X_{i^*k}}{X_{nk} - X_{1k}} = \frac{X_{nk} - X_{(i^*+1)k} + X_{(i^*+1)k} - X_{i^*k}}{X_{nk} - X_{1k}} \leq \frac{1}{2} + \frac{1}{16}$$

where the inequality follows because we had assumed that $\frac{X_{(i+1)k} - X_{ik}}{X_{nk} - X_{1k}} \leq \frac{1}{16}$ for all $i = 1, ..., n - 1$.

So, we have

$$\begin{aligned} a' &= \frac{X_{nk} - X_{1k}}{X_{nk} - X_{i^*k}}\kappa + \frac{X_{i^*k} - X_{1k}}{X_{nk} - X_{i^*k}}\frac{1}{8} \\ &= \frac{X_{nk} - X_{1k}}{X_{nk} - X_{i^*k}}\left(\kappa + \frac{X_{i^*k} - X_{1k}}{X_{nk} - X_{1k}}\frac{1}{8}\right) \\ &\geq \frac{X_{nk} - X_{1k}}{X_{nk} - X_{i^*k}}\left(-\frac{1}{32} + (\frac{1}{2} - \frac{1}{16})\frac{1}{8}\right) \\ &\geq \frac{1}{1/2 + 1/16}\left(-\frac{1}{32} + (\frac{1}{2} - \frac{1}{16})\frac{1}{8}\right) \\ &\geq \frac{1}{32} \end{aligned}$$

In the first inequality of the above derivation, we used the fact that $\frac{X_{i^*k} - X_{1k}}{X_{nk} - X_{1k}} \leq \frac{1}{2} - \frac{1}{16}$. In the second inequality, we used the fact that the quantity inside the parentheses is positive and $\frac{X_{nk} - X_{1k}}{X_{nk} - X_{i^*k}} \geq \frac{1}{1/2+1/16}$.

Now consider $j = n - 1$.

$$[\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}]_{n-1} = \frac{1}{n}\mathsf{gap}_n\widehat{r}_n \leq \mathsf{gap}_n\frac{\lambda}{32} \leq \lambda\mathsf{gap}_n a' = \lambda[\Delta_k^\mathsf{T}\mathbf{u}]_{n-1}$$

For $j = i^*, ..., n - 2$, we have that

$$\begin{aligned} \lambda[\Delta_k^\mathsf{T}\mathbf{u}]_j - \lambda[\Delta_k^\mathsf{T}\mathbf{u}]_{j+1} &= \lambda a'\mathsf{gap}_{j+1} \\ &\geq \mathsf{gap}_{j+1}\frac{1}{n}\mathbf{1}_{(j+1):n}^\mathsf{T}\widehat{r} \\ &\geq [\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}]_j - [\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}]_{j+1} \end{aligned}$$

Therefore, for $j = i^*, ..., n - 2$,

$$\lambda[\Delta_k^\mathsf{T}\mathbf{u}]_j \geq \frac{1}{n}[\Delta_k^\mathsf{T}\widehat{r}]_j$$

We conclude then that $\lambda[\Delta_k^\mathsf{T}\mathbf{u}]_j \geq [\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}]_j$ for all $j = 2, ..., n - 1$.

We have thus verified that the stationarity equations hold and now will bound $\|\mathbf{u}\|_1$.

$$\|\mathbf{u}\|_1 = |a - a'| + a + a' \leq \frac{1}{8} + 2a \leq \frac{1}{8} + 4|\kappa| + \frac{1}{2} \leq \frac{1}{8} + \frac{1}{8} + \frac{1}{2} < 1$$

In the third inequality, we used the fact that $|\kappa| \leq \frac{1}{32}$.

We have thus proven that there exists one solution $\{\widehat{d}_k\}_{k=1,\ldots,p}$ such that $\widehat{d}_k = 0$ for all $k \in S^c$. Furthermore, we have shown that the subgradient variables $\mathbf{u}_k$ of the solution $\{\widehat{d}_k\}$ can be chosen such that $\|\mathbf{u}_k\|_1 < 1$ for all $k \in S^c$.

We now prove that if $\{\widehat{d}'_k\}_{k=1,\ldots,p}$ is another solution, then it must be that $\widehat{d}'_k = 0$ for all $k \in S^c$ as well. We first claim that $\sum_{k=1}^p \bar{\Delta}_k \widehat{d}_k = \sum_{k=1}^p \bar{\Delta}_k \widehat{d}'_k$. If this were not true, then a convex combination of $\widehat{d}_k, \widehat{d}'_k$ would achieve a strictly lower objective on the quadratic term. More precisely, let $\zeta \in [0,1]$. If $\sum_{k=1}^p \bar{\Delta}_k \widehat{d}'_k \neq \sum_{k=1}^p \bar{\Delta}_k \widehat{d}_k$, then $\|Y - \sum_{k=1}^p \bar{\Delta}_k\big(\widehat{d}_k + \zeta(\widehat{d}'_k - \widehat{d}_k)\big)\|_2^2$ is strongly convex as a function of ζ. Thus, it cannot be that $\widehat{d}_k$ and $\widehat{d}'_k$ both achieve optimal objective, and we have reached a contradiction.

Now, we look at the stationarity condition for both $\{\widehat{d}_k\}$ and $\{\widehat{d}'_k\}$. Let $\mathbf{u}_k \in \partial\|\bar{\Delta}_k \widehat{d}_k\|_\infty$ and let $\mathbf{u}'_k \in \partial\|\bar{\Delta}_k \widehat{d}'_k\|_\infty$ be the two sets of subgradients. Let $\{\nu_{ik}\}$ and $\{\nu'_{ik}\}$ be the two sets of positivity dual variables, for $k = 1, .., p$ and $i = 1, ...n-1$. Note that since there is no positivity constraint on d_{1k}, we let $\nu_{1k} = 0$ always.

Let us define $\bar{\Delta}$, a $n \times p(n-1)$ matrix, to denote the column-wise concatenation of $\{\bar{\Delta}_k\}_k$ and $\widehat{d}$, a $p(n-1)$ dimensional vector, to denote the concatenation of $\{\widehat{d}_k\}_k$. With this notation, we can express $\sum_{k=1}^p \bar{\Delta}_k \widehat{d}_k = \bar{\Delta}\widehat{d}$.

Since both solutions $(\widehat{d}, \mathbf{u}, \nu)$ and $(\widehat{d}', \mathbf{u}', \nu')$ must satisfy the stationarity condition, we have that

$$\bar{\Delta}^\mathsf{T}(Y - \bar{\Delta}\widehat{d}) + \lambda \begin{pmatrix} \bar{\Delta}_1^\mathsf{T}\mathbf{u}_1 \\ ... \\ \bar{\Delta}_p^\mathsf{T}\mathbf{u}_p \end{pmatrix} - \nu = \bar{\Delta}^\mathsf{T}(Y - \bar{\Delta}\widehat{d}') + \lambda \begin{pmatrix} \bar{\Delta}_1^\mathsf{T}\mathbf{u}'_1 \\ ... \\ \bar{\Delta}_p^\mathsf{T}\mathbf{u}'_p \end{pmatrix} - \nu' = 0.$$

Multiplying both sides of the above equation by $\widehat{d}'$,

$$\widehat{d}'^\mathsf{T}\bar{\Delta}^\mathsf{T}(Y - \bar{\Delta}\widehat{d}) + \lambda \sum_{k=1}^p \widehat{d}_k'^\mathsf{T}\bar{\Delta}_k^\mathsf{T}\mathbf{u}_k - \widehat{d}'^\mathsf{T}\nu = \widehat{d}'^\mathsf{T}\bar{\Delta}^\mathsf{T}(Y - \bar{\Delta}\widehat{d}') + \lambda \sum_{k=1}^p \widehat{d}_k'^\mathsf{T}\bar{\Delta}_k^\mathsf{T}\mathbf{u}'_k - \widehat{d}'^\mathsf{T}\nu'.$$

Since $\bar{\Delta}\widehat{d}' = \bar{\Delta}\widehat{d}$, $\widehat{d}'^\mathsf{T}\nu' = 0$ (complementary slackness), and $\widehat{d}_k'^\mathsf{T}\bar{\Delta}_k^\mathsf{T}\mathbf{u}'_k = \|\widehat{f}'_k\|_\infty$ (where $\widehat{f}'_k = \bar{\Delta}_k\widehat{d}'_k$), we have that

$$\lambda \sum_{k=1}^p \widehat{d}_k'^\mathsf{T}\bar{\Delta}_k^\mathsf{T}\mathbf{u}_k - \widehat{d}'^\mathsf{T}\nu = \lambda \sum_{k=1}^p \|\widehat{f}'_k\|_\infty.$$

On one hand, $\widehat{d}'$ is a feasible solution so $\widehat{d}'^\mathsf{T}\nu \geq 0$ and so

$$\sum_{k=1}^p \widehat{d}_k'^\mathsf{T}\bar{\Delta}_k^\mathsf{T}\mathbf{u}_k \geq \sum_{k=1}^p \|\widehat{f}'_k\|_\infty.$$

On the other hand, by Hölder's inequality,

$$\sum_{k=1}^{p} \widehat{d}_k'^{\mathsf{T}} \bar{\Delta}_k^{\mathsf{T}} \mathbf{u}_k \leq \sum_{k=1}^{p} \|\widehat{f}_k'\|_\infty \|\mathbf{u}_k\|_1.$$

Since $\mathbf{u}_k$ can be chosen so that $\|\mathbf{u}_k\|_1 < 1$ for all $k \in S^c$, we would get a contradiction if $\|\widehat{f}_k'\|_\infty > 0$ for some $k \in S^c$. We thus conclude that $\widehat{d}'$ must follow the same sparsity pattern.

The second item in the theorem concerning optimization 2.4.10 is proven in exactly the same way. The Lagrangian of optimization 2.4.10 is

$$\mathcal{L}_{\text{cave}}(c_k, \nu_k) = \frac{1}{2n}\left\|\widehat{r} - \bar{\Delta}_k c_k\right\|_2^2 + \lambda \|\bar{\Delta}_k c_k\|_\infty + \sum_{k=1}^{p}\sum_{i=2}^{n-1} \nu_{ik} c_{ik}.$$

with $\nu_{ik} \geq 0$. The same reasoning applies to show that $\widehat{c}_k = 0$ for all $k \in S^c$ satisfies KKT conditions sufficient for optimality. □

Proof of False Positive Control

We note that in the following analysis the symbols c, C represent absolute constants. We will often abuse notation and "absorb" new absolute constants into c, C; the actual value of c, C could thus vary from line to line. We first restate the theorem for convenience.

Theorem 2.7.2. *Suppose assumptions A1-A5 hold. Define* $\tilde{\sigma} \equiv \max(\sigma, B)$. *Suppose that* $p \leq O\big(\exp(cn)\big)$ *and* $n \geq C$ *for some constants* $0 < c < 1$ *and* C. *Define* $\mathsf{range}_k = X_{\pi_k(n)k} - X_{\pi_k(1)k}$.

If $\lambda_n \geq 2(12 \cdot 32) s\tilde{\sigma}\sqrt{\frac{1}{n}\log^2 np}$ *then, with probability at least* $1 - \frac{24}{n}$, *for all* $k \in S^c$, *and for all* $i' = 1, ..., n$

$$\lambda_n > \mathsf{range}_k \left| \frac{32}{n} \widehat{r}^{\mathsf{T}} \mathbf{1}_{\pi_k(i':n)} \right|$$

and $\max_{i'} \frac{X_{\pi_k(i'+1)k} - X_{\pi_k(i')k}}{\mathsf{range}_k} \leq \frac{1}{16}$ *and* $\mathsf{range}_k \geq 1$.

Therefore, for all $k \in S^c$, *both the AC solution* $\widehat{f}_k$ *from optimization 2.4.8, and the DC solution* $\widehat{g}_k$ *from optimization 2.4.10 are zero.*

Proof. The key is to note that $\widehat{r}$ and $\Delta_{k,j}$ are independent for all $k \in S^c, j = 1, ..., n$ because $\widehat{r}$ is only dependent on X_S.

Fix j and i. Then $\widehat{r}^{\mathsf{T}}\mathbf{1}_{\pi_k(i':n)}$ is the sum of $n - i' + 1$ random coordinates of $\widehat{r}$. We will use Serfling's theorem on the concentration of measure of sampling without replacement (Corollary 2.7.2). We must first bound $\|\widehat{r}\|_\infty$ and $\frac{1}{n}\sum_{i=1}^n \widehat{r}_i$ before we can use Serfling's results however.

Step 1: *Bounding* $\|\widehat{r}\|_\infty$. We have $\widehat{r}_i = f_0(x_i) + w_i - \widehat{f}(x_i)$ where $\widehat{f}(x_i) = \sum_{k\in S}\bar{\Delta}_k\widehat{d}_k$ is the convex additive function outputted by the restricted regression. Note that both $f_0(x_i)$ and $\widehat{f}(x_i)$ are bounded by $2sB$. Because w_i is sub-Gaussian, $|w_i| \le \sigma\sqrt{2\log\frac{2}{\delta}}$ with probability at least $1-\delta$. By union bound across $i = 1, ..., n$, we have that $\|w\|_\infty \le \sigma\sqrt{2\log\frac{2n}{\delta}}$ with probability at least $1-\delta$.

Putting these observations together,

$$\begin{aligned}\|\widehat{r}\|_\infty &\le 2sB + \sigma\sqrt{2\log\frac{2n}{\delta}})\\ &\le 4s\widetilde{\sigma}\sqrt{\log\frac{2n}{\delta}} \end{aligned} \tag{2.7.4}$$

with probability at least $1-\delta$, where we have defined $\widetilde{\sigma} = \max(\sigma, B)$, and assumed that $\sqrt{\log\frac{2n}{\delta}} \ge 1$. We will eventually take $\delta = O(1/n)$ so this assumption holds under the condition in the theorem which state that $n \ge C$ for some large constant C.

Step 2: *Bounding* $|\frac{1}{n}\widehat{r}^{\mathsf{T}}\mathbf{1}|$. We have that

$$\begin{aligned}\frac{1}{n}\widehat{r}^{\mathsf{T}}\mathbf{1} &= \frac{1}{n}\sum_{i=1}^n f_0(x_i) + w_i - \widehat{f}(x_i)\\ &= \frac{1}{n}\sum_{i=1}^n f_0(x_i) + w_i \quad (\widehat{f} \text{ is centered}).\end{aligned}$$

Since $|f_0(x_i)| \le sB$, the first term $|\frac{1}{n}\sum_{i=1}^n f_0(x_i)|$ is at most $sB\sqrt{\frac{2}{n}\log\frac{2}{\delta}}$ with probability at most $1-\delta$ by Hoeffding's inequality. Since w_i is sub-Gaussian, the second term $|\frac{1}{n}\sum_{i=1}^n w_i|$ is at most $\sigma\sqrt{\frac{2}{n}\log\frac{2}{\delta}}$ with probability at most $1-\delta$. Taking a union bound, we have that

$$\begin{aligned}|\frac{1}{n}\widehat{r}^{\mathsf{T}}\mathbf{1}| &\le sB\sqrt{\frac{2}{n}\log\frac{4}{\delta}} + \sigma\sqrt{\frac{2}{n}\log\frac{4}{\delta}}\\ &\le 4s\widetilde{\sigma}\sqrt{\frac{1}{n}\log\frac{4}{\delta}}\end{aligned} \tag{2.7.5}$$

with probability at least $1-\delta$.

Step 3: *Apply Serfling's theorem.* For any $k \in S^c$, Serfling's theorem states that with probability at least $1-\delta$

$$\left|\frac{1}{n}\widehat{r}^{\mathsf{T}}\mathbf{1}_{\pi_k(i':n)}\right| \leq 2\|\widehat{r}\|_\infty\sqrt{\frac{1}{n}\log\frac{2}{\delta}} + \left|\frac{1}{n}\widehat{r}^{\mathsf{T}}\mathbf{1}\right|$$

We need Serfling's theorem to hold for all $k = 1, ..., p$ and $i' = 1, ..., n$. We also need the events that $\|\widehat{r}\|_\infty$ and $|\frac{1}{n}\widehat{r}^{\mathsf{T}}\mathbf{1}|$ are small to hold. Using a union bound, with probability at least $1-\delta$, for all k, i',

$$\begin{aligned}\left|\frac{1}{n}\widehat{r}^{\mathsf{T}}\mathbf{1}_{\pi_k(i':n)}\right| &\leq 2\|\widehat{r}\|_\infty\sqrt{\frac{1}{n}\log\frac{6np}{\delta}} + \left|\frac{1}{n}\widehat{r}^{\mathsf{T}}\mathbf{1}\right| \\ &\leq 8s\widetilde{\sigma}\sqrt{\log\frac{6n}{\delta}}\sqrt{\frac{1}{n}\log\frac{6np}{\delta}} + 4s\widetilde{\sigma}\sqrt{\frac{1}{n}\log\frac{12}{\delta}} \\ &\leq 12s\widetilde{\sigma}\sqrt{\frac{1}{n}\log^2\frac{12np}{\delta}}\end{aligned}$$

In the second inequality, we used equation (2.7.4) and equation (2.7.5) from steps 1 and 2 respectively. Setting $\delta = \frac{12}{n}$ gives the desired expression.

Finally, we note that $2 \geq (X_{\pi_k(n)k} - X_{\pi_k(1)k})$ since $X_k \subset [-1, 1]$. This concludes the proof for the first part of the theorem.

To prove the second and the third claims, let the interval $[-1, 1]$ be divided into 64 non-overlapping segments each of length $1/32$. Because X_k is drawn from a density with a lower bound $c_l > 0$, the probability that every segment contains some samples X_{ki}'s is at least $1 - 64\left(1 - \frac{1}{32}c_l\right)^n$. Let $\mathcal{E}_k$ denote the event that every segment contains some samples.

Define $\mathsf{gap}_i = X_{\pi_k(i+1)k} - X_{\pi_k(i)k}$ for $i = 1, ..., n-1$ and define $\mathsf{gap}_0 = X_{\pi_k(1)k} - (-1)$ and $\mathsf{gap}_n = 1 - X_{\pi_k(n)k}$.

If any $\mathsf{gap}_i \geq \frac{1}{16}$, then gap_i has to contain one of the segments. Therefore, under event $\mathcal{E}_k$, it must be that $\mathsf{gap}_i \leq \frac{1}{16}$ for all i.

Thus, we have that $\mathsf{range}_k \geq 2 - 1/8 \geq 1$ and that for all i,

$$\frac{X_{k\pi_k(i+1)} - X_{k\pi_k(i)}}{\mathsf{range}_k} \leq \frac{1/16}{2 - 1/8} \leq 1/16$$

Taking a union bound for each $k \in S^c$, the probability of that all $\mathcal{E}_k$ hold is at least $1 - p64\left(1 - \frac{1}{32}c_l\right)^n$.

$p64\left(1 - \frac{1}{32}c_l\right)^n = 64p\exp(-c'n)$ where $c' = -\log(1 - \frac{c_l}{32}) > \frac{c_l}{32}$ since $c_l < 1$. Therefore, if $p \leq exp(cn)$ for some $0 < c < \frac{c_l}{32}$ and if n is larger than some constant C, $64p\exp(-c'n) \leq 64\exp(-(c'-c)n) \leq \frac{12}{n}$.

Taking a union bound with the event that λ_n upper bounds the partial sums of $\widehat{r}$ and we establish the claim.

□

Proof of False Negative Control

We begin by introducing some notation.

Notation

If $f : \mathbb{R}^s \to \mathbb{R}$, we define $\|f\|_P \equiv \mathbb{E}f(X)^2$. Given samples $X_1, ..., X_n$, we denote $\|f\|_n \equiv \frac{1}{n}\sum_{i=1}^n f(X_i)^2$ and $\langle f, g\rangle_n \equiv \frac{1}{n}\sum_{i=1}^n f(X_i)g(X_i)$.

Let $\mathcal{C}^1$ denote the set of univariate convex functions supported on $[-1, 1]$. Let $\mathcal{C}^1_B \equiv \{f \in \mathcal{C}^1 \,:\, \|f\|_\infty \leq B\}$ denote the set of B-bounded univariate convex functions. Define $\mathcal{C}^s$ as the set of convex additive functions and $\mathcal{C}^s_B$ likewise as the set of convex additive functions whose components are B-bounded:

$$\mathcal{C}^s \equiv \{f \,:\, f = \sum_{k=1}^s f_k,\ f_k \in \mathcal{C}^1\}$$

$$\mathcal{C}^s_B \equiv \{f \in \mathcal{C}^s \,:\, f = \sum_{k=1}^s f_k,\ \|f_k\|_\infty \leq B\}.$$

Let $f^*(x) = \sum_{k=1}^s f_k^*(x_k)$ be the population risk minimizer:

$$f^* = \arg\min_{f \in \mathcal{C}^s} \|f_0 - f^*\|_P^2$$

We let sB be an upper bound on $\|f_0\|_\infty$ and B be an upper bound on $\|f_k^*\|_\infty$. It follows that $\|f^*\|_\infty \leq sB$.

We define $\widehat{f}$ as the empirical risk minimizer:

$$\widehat{f} = \arg\min \left\{\|y - f\|_n^2 + \lambda \sum_{k=1}^s \|f_k\|_\infty \,:\, f \in \mathcal{C}^s_B,\ \mathbf{1}_n^\mathsf{T} f_k = 0\right\}$$

For $k \in \{1, ..., s\}$, define g_k^* to be the decoupled concave population risk minimizer

$$g_k^* \equiv \operatorname*{arg\,min}_{g_k \in \text{-}\mathcal{C}^1} \|f_0 - f^* - g_k\|_P^2.$$

In our proof, we will analyze g_k^* for each k such that $f_k^* = 0$. Likewise, we define the empirical version:

$$\widehat{g}_k \equiv \arg\min \left\{\|f_0 - \widehat{f} - g_k\|_n^2 \,:\, g_k \in \text{-}\mathcal{C}^1_B\,, \mathbf{1}_n^\mathsf{T} g_k = 0\right\}.$$

By the definition of the AC/DC procedure, $\widehat{g}_k$ is defined only for an index k that has zero as the convex additive approximation.

Proof

By additive faithfulness of the AC/DC procedure, it is known that $f_k^* \neq 0$ or $g_k^* \neq 0$ for all $k \in S$. Our argument will be to show that the risk of the AC/DC estimators $\widehat{f}, \widehat{g}$ tends to the risk of the population optimal functions f^*, g^*:

$$\begin{aligned} \|f_0 - \widehat{f}\|_P^2 &= \|f_0 - f^*\|_P^2 + \mathrm{err}_+(n) \\ \|f_0 - f^* - \widehat{g}_k\|_P^2 &= \|f_0 - f^* - g_k^*\|_P^2 + \mathrm{err}_-(n) \quad \text{for all } k \in S \text{ where } f_k^* = 0, \end{aligned}$$

where the estimation errors $\mathrm{err}_+(n)$ and $\mathrm{err}_-(n)$ decrease with n at some rate.

Assuming this, suppose that $\widehat{f}_k = 0$ and $f_k^* \neq 0$. Then when n is large enough such that $\mathrm{err}_+(n)$ and $\mathrm{err}_-(n)$ are smaller than α_+ and α_- defined in equation (2.5.3), we reach a contradiction. This is because the risk $\|f_0 - f^*\|_P$ of f^* is strictly larger by α_+ than the risk of the best approximation whose k-th component is constrained to be zero. Similarly, suppose $f_k^* = 0$ and $g_k^* \neq 0$. Then when n is large enough, $\widehat{g}_k$ must not be zero.

Theorem 2.7.3 and Theorem 2.7.4 characterize $\mathrm{err}_+(n)$ and $\mathrm{err}_-(n)$ respectively.

Theorem 2.7.3. *Let $\widetilde{\sigma} \equiv \max(\sigma, B)$, and let $\widehat{f}$ be the minimizer of the restricted regression with $\lambda \leq 768 s\widetilde{\sigma}\sqrt{\frac{1}{n}\log^2 np}$. Suppose $n \geq c_1 s\sqrt{sB}$. Then with probability at least $1 - \frac{C}{n}$,*

$$\|f_0 - \widehat{f}\|_P^2 - \|f_0 - f^*\|_P^2 \leq cB^2\widetilde{\sigma}\sqrt{\frac{s^5}{n^{4/5}}\log^2 Cnp}, \tag{2.7.6}$$

where c_1 is an absolute constant and c, C are constants possibly dependent on b.

Proof. Our proof proceeds in three steps. First, we bound the difference of empirical risks $\|f_0 - \widehat{f}\|_n^2 - \|f_0 - f^*\|_n^2$. Second, we bound the cross-term in the bound using a bracketing entropy argument for convex function classes. Finally, we combine the previous two steps to complete the argument.

Step 1. The function $\widehat{f}$ minimizes the penalized empirical risk by definition. We would thus like to say that the penalized empirical risk of $\widehat{f}$ is no larger than that of f^*. We cannot do a direct comparison, however, because the empirical mean $\frac{1}{n}\sum_i f_k^*(x_{ik})$ is close to, but not exactly zero. We thus have to work first with the function $f^* - \bar{f}^*$. We have that

$$\|y - \widehat{f}\|_n^2 + \lambda\sum_{k=1}^s \|\widehat{f}_k\|_\infty \leq \|y - f^* + \bar{f}^*\|_n^2 + \lambda\sum_{k=1}^s \|f_k^* - \bar{f}_k^*\|_\infty$$

Plugging in $y = f_0 + w$, we obtain

$$\|f_0 + w - \widehat{f}\|_n^2 + \lambda \sum_{k=1}^{s} \Big(\|\widehat{f}_k\|_\infty - \|f_k^* - \bar{f}_k^*\|_\infty \Big) \leq \|f_0 + w - f^* + \bar{f}^*\|_n^2$$

$$\|f_0 - \widehat{f}\|_n^2 + 2\langle w, f_0 - \widehat{f}\rangle_n + \lambda \sum_{k=1}^{s} \Big(\|\widehat{f}_k\|_\infty - \|f_k^* - \bar{f}_k^*\|_\infty \Big)$$
$$\leq \|f_0 - f^* + \bar{f}^*\|_n^2 + 2\langle w, f_0 - f^* + \bar{f}^*\rangle$$

$$\|f_0 - \widehat{f}\|_n^2 - \|f_0 - f^* + \bar{f}^*\|_n^2 + \lambda \sum_{k=1}^{s} \Big(\|\widehat{f}_k\|_\infty - \|f_k^* - \bar{f}_k^*\|_\infty \Big) \leq 2\langle w, \widehat{f} - f^* + \bar{f}^*\rangle.$$

The middle term can be bounded since $\|f_k^* - \bar{f}_k^*\|_\infty \leq B$; thus,

$$\|f_0 - \widehat{f}\|_n^2 - \|f_0 - f^* + \bar{f}^*\|_n^2 \leq 2\langle w, \widehat{f} - f^* + \bar{f}^*\rangle + \lambda s B.$$

Using Lemma 2.7.2, we can remove $\bar{f}^*$ from the lefthand side. Thus with probability at least $1 - \delta$,

$$\|f_0 - \widehat{f}\|_n^2 - \|f_0 - f^*\|_n^2 \leq 2\langle w, \widehat{f} - f^* + \bar{f}^*\rangle + \lambda s B + c(sB)^2 \frac{1}{n} \log \frac{2}{\delta}. \tag{2.7.7}$$

Step 2. We now upper bound the cross term $2\langle w, \widehat{f} - f^* + \bar{f}^*\rangle$ using bracketing entropy.

Define $\mathcal{G} = \{f - f^* + \bar{f}^* : f \in \mathcal{C}_B^s\}$ as the set of convex additive functions centered around the function $f^* - \bar{f}^*$. By Corollary 2.7.3, there is an ϵ-bracketing of $\mathcal{G}$ whose log-size is bounded by $\log N_{[]}(\epsilon, \mathcal{G}, L_1(P)) \leq sK^{**} \left(\frac{4sBc_u}{\epsilon}\right)^{1/2}$, for all $\epsilon \in (0, sB\epsilon_3 c_u]$. Let us suppose condition 2.7.12 holds. Then, by Corollary 2.7.4, with probability at least $1-\delta$, each bracketing pair (h_U, h_L) is close in $L_1(P_n)$ norm, i.e., for all (h_U, h_L), $\frac{1}{n}\sum_{i=1}^n |h_U(X_i) - h_L(X_i)| \leq \epsilon + 2sB\sqrt{\frac{sK^{**}(sBc_u)^{1/2}\log\frac{2}{\delta}}{\epsilon^{1/2}n}}$. We verify at the end of the proof that condition 2.7.12 indeed holds.

For each $h \in \mathcal{G}$, there exists a pair (h_U, h_L) such that $h_U(X_i) - h_L(X_i) \geq h(X_i) - h_L(X_i) \geq 0$. Therefore, with probability at least $1 - \delta$, uniformly for all $h \in \mathcal{G}$:

$$\frac{1}{n}\sum_{i=1}^n |h(X_i) - h_L(X_i)| \leq \frac{1}{n}\sum_{i=1}^n |h_U(X_i) - h_L(X_i)| \leq \epsilon + (2sB)\sqrt{\frac{sK^{**}(sBc_u)^{1/2}\log\frac{2}{\delta}}{\epsilon^{1/2}n}}.$$

We denote $\epsilon_{n,\delta} \equiv (2sB)\sqrt{\frac{sK^{**}(sBc_u)^{1/2}\log\frac{2}{\delta}}{\epsilon^{1/2}n}}$. Let $\mathcal{E}_{[]}$ denote the event that for each $h \in \mathcal{G}$, there exists h_L in the ϵ-bracketing such that $\|h - h_L\|_{L_1(P_n)} \leq \epsilon + \epsilon_{n,\delta}$. Then $\mathcal{E}_{[]}$ has probability at most $1 - \delta$ as shown.

Let $\mathcal{E}_{\|w\|_\infty}$ denote the event that $\|w\|_\infty \leq \sigma\sqrt{2\log\frac{2n}{\delta}}$. Then $\mathcal{E}_{\|w\|_\infty}$ has probability at most $1-\delta$. We now take an union bound over $\mathcal{E}_{\|w\|_\infty}$ and $\mathcal{E}_{[]}$ and get that, with probability at most $1-2\delta$, for all h

$$|\langle w, h - h_L\rangle_n| \leq \|w\|_\infty \frac{1}{n}\sum_{i=1}^n |h(X_i) - h_L(X_i)| \leq \sigma\sqrt{2\log\frac{4n}{\delta}}\left(\epsilon + \epsilon_{n,2\delta}\right).$$

Because w is a sub-Gaussian random variable, we have that the random variables $w_i h_L(X_i)$ are independent, centered, and sub-Gaussian with scale at most $2\sigma sB$. Thus, with probability at least $1-\delta$, $|\langle w, h_L\rangle_n| \leq 2\sigma sB\sqrt{\frac{1}{n}\log\frac{2}{\delta}}$. Using another union bound, we have that the event $\sup_{h_L}|\langle w, h_L\rangle| \leq 2\sigma sB\sqrt{\frac{1}{n}\log\frac{2N_{[]}}{\delta}}$ has probability at most $1-\delta$.

Putting this together, we have that

$$\begin{aligned}
&|\langle w, h\rangle_n| \leq |\langle w, h_L\rangle_n| + |\langle w, h - h_L\rangle_n| \\
&|\sup_{h\in\mathcal{G}}\langle w, h\rangle_n| \leq |\sup_{h_L}\langle w, h_L\rangle_n| + \sigma\sqrt{2\log\frac{2n}{\delta}}(\epsilon + \epsilon_{n,2\delta}) \\
&\leq 2sB\sigma\sqrt{\frac{\log N_{[]} + \log\frac{2}{\delta}}{n}} + \sigma\sqrt{2\log\frac{2n}{\delta}}(\epsilon + \epsilon_{n,\delta}) \\
&\leq 2sB\sigma\sqrt{\frac{sK^{**}(4sBc_u)^{1/2} + \log\frac{1}{\delta}}{n\epsilon^{1/2}}} + \sigma\sqrt{2\log\frac{2n}{\delta}}(\epsilon + \epsilon_{n,\delta}) \\
&\leq 2sB\sigma\sqrt{\frac{sK^{**}(4sBc_u)^{1/2} + \log\frac{1}{\delta}}{n\epsilon^{1/2}}} + \sigma\sqrt{2\log\frac{2n}{\delta}}\epsilon + 2sB\sigma\sqrt{2\frac{sK^{**}(sBc_u)^{1/2}\log\frac{1}{\delta}}{n\epsilon^{1/2}}\log\frac{2n}{\delta}} \\
&\leq \sigma\sqrt{2\log\frac{2n}{\delta}}\epsilon + 8sB\sigma\sqrt{\frac{sK^{**}(sBc_u)^{1/2}\log^2\frac{2n}{\delta}}{n\epsilon^{1/2}}}.
\end{aligned}$$

On the last line, we have assumed that conditions 2.7.12 hold so that $\frac{sK^{**}(sBc_u)^{1/2}}{\epsilon^{1/2}} + \log 1/\delta \leq \frac{sK^{**}(sBc_u)^{1/2}}{\epsilon^{1/2}}\log 1/\delta$.

We choose $\epsilon = \left(\frac{(sB)^2(sK^{**}(sBc_u)^{1/2})}{n}\right)^{2/5}$. This choice of ϵ is a bit suboptimal but it is convenient and it is sufficient for our results. It is easy to verify that if $n \geq c_1 s\sqrt{sB}$ for some absolute constant c_1, then $\epsilon \in (0, sB\epsilon_3 c_u]$ for some absolute constant ϵ_3 as required by the bracketing number statement (Corollary 2.7.3). Furthermore, conditions (2.7.12) also hold.

In summary, we have that probability at least $1-\delta$,

$$|\sup_{h\in\mathcal{G}}\langle w, h\rangle| \leq csB\sigma\sqrt{\frac{s^{6/5}(Bc_u)^{2/5}\log^2\frac{Cn}{\delta}}{n^{4/5}}} \leq csB\sigma\sqrt{\frac{s(sBc_u)^{1/2}\log^2\frac{Cn}{\delta}}{n^{4/5}}}$$

where we absorbed K^{**} into the constant c and the union bound multipliers into the constant C.

Plugging this result into equation (2.7.7) we get that, with probability at least $1-2\delta$,

$$\begin{aligned}
\|f_0-\widehat{f}\|_n^2-\|f_0-f^*\|_n^2 &\le csB\sigma\sqrt{\frac{s(sBc_u)^{1/2}\log^2\frac{Cn}{\delta}}{n^{4/5}}}+\lambda sB+c(sB)^2\frac{1}{n}\log\frac{2}{\delta}\\
\|f_0-\widehat{f}\|_n^2-\|f_0-f^*\|_n^2 &\le cB^2\sigma\sqrt{\frac{s^4c_u^{1/2}\log^2\frac{Cn}{\delta}}{n^{4/5}}}+\lambda sB\\
&\le cB^2\sigma\sqrt{\frac{s^4c_u^{1/2}}{n^{4/5}}\log^2\frac{Cn}{\delta}}+\lambda sB \qquad (2.7.8)
\end{aligned}$$

Step 3. Continuing from equation (2.7.8), we use Lemma 2.7.1 and another union bound to obtain that, with probability at least $1-3\delta$,

$$\begin{aligned}
\|f_0-\widehat{f}\|_P^2-\|f_0-f^*\|_P^2 &\le cB^2\sigma\sqrt{\frac{s^4c_u^{1/2}}{n^{4/5}}\log^2\frac{Cn}{\delta}}+\lambda sB+cB^3\sqrt{\frac{s^5c_u^{1/2}}{n^{4/5}}\log\frac{2}{\delta}}\\
&\le cB^2\widetilde{\sigma}\sqrt{\frac{s^5c_u^{1/2}}{n^{4/5}}\log^2\frac{Cn}{\delta}}+\lambda sB
\end{aligned}$$

Substituting in $\lambda \le 768s\widetilde{\sigma}\sqrt{\frac{1}{n}\log^2 np}$ and $\delta=\frac{C}{n}$ we obtain the statement of the theorem. □

Theorem 2.7.4. *Let $\widehat{g}_k$ denote the minimizer of the concave postprocessing step with $\lambda_n \le 768s\widetilde{\sigma}\sqrt{\frac{1}{n}\log^2 np}$. Let $\widetilde{\sigma}\equiv\max(\sigma,B)$. Suppose n is sufficiently large that $\frac{n^{4/5}}{\log^2 np}\ge c'B^4\widetilde{\sigma}^2s^5c_u^{1/2}$ where $c'\ge 1$ is a constant. Then with probability at least $1-\frac{C}{n}$, for all $k=1,...,s$,*

$$\|f_0-f^*-\widehat{g}_k\|_P^2-\|f_0-f^*-g_k^*\|_P^2 \le cB^2\widetilde{\sigma}^{1/2}\sqrt[4]{\frac{s^5c_u^{1/2}}{n^{4/5}}\log^2 np}.$$

Proof. This proof is similar to that of Theorem 2.7.3; it requires a few more steps because $\widehat{g}_k$ is fitted against $f_0-\widehat{f}$ instead of f_0-f^*. We start with the following

decomposition:

$$\|f_0 - f^* - \widehat{g}_k\|_P^2 - \|f_0 - f^* - g_k^*\|_P^2 = \underbrace{\|f_0 - \widehat{f} - \widehat{g}_k\|_P^2 - \|f_0 - \widehat{f} - g_k^*\|_P^2}_{\text{term 1}} + \underbrace{\|f_0 - f^* - \widehat{g}_k\|_P^2 - \|f_0 - \widehat{f} - \widehat{g}_k\|_P^2}_{\text{term 2}} + \underbrace{\|f_0 - \widehat{f} - g_k^*\|_P^2 - \|f_0 - f^* - g_k^*\|_P^2}_{\text{term 3}}. \quad (2.7.9)$$

We now bound each of the terms. The proof proceeds almost identically to that of Theorem 2.7.3, because convex and concave functions have the same bracketing number.

Step 1. To bound term 1, we start from the definition of $\widehat{g}_k$ and obtain

$$\|y - \widehat{f} - \widehat{g}_k\|_n^2 + \lambda_n \|\widehat{g}\|_\infty \leq \|y - \widehat{f} - (g_k^* - \bar{g}_k^*)\|_n^2 + \lambda_n \|(g^* - \bar{g}_k^*)\|_\infty$$
$$\|y - \widehat{f} - \widehat{g}_k\|_n^2 \leq \|y - \widehat{f} - (g_k^* - \bar{g}_k^*)\|_n^2 + \lambda_n B$$

$$\|f_0 - \widehat{f} - \widehat{g}_k + w\|_n^2 \leq \|f_0 - \widehat{f} - (g_k^* - \bar{g}_k^*) + w\|_n^2 + \lambda_n B$$
$$\|f_0 - \widehat{f} - \widehat{g}_k\|_n^2 - \|f_0 - \widehat{f} - (g_k^* - \bar{g}_k^*)\|_n^2 \leq 2\langle w, \widehat{g}_k - (g_k^* - \bar{g}_k^*)\rangle_n + \lambda_n B.$$

$\bar{g}_k^* = \frac{1}{n}\sum_{i=1}^n g_k^*(X_{ik})$; we subtract it from g_k^* again so that g_k^* is empirically mean-zero.

Using the same bracketing analysis as in Step 2 of the proof of Theorem 2.7.3 but setting $s = 1$, we have, with probability at least $1 - \delta$,

$$\|f_0 - \widehat{f} - \widehat{g}_k\|_n^2 - \|f_0 - \widehat{f} - g_k^*\|_n^2 \leq cB^2\sigma\sqrt{\frac{c_u^{1/2}}{n^{4/5}}\log\frac{nC}{\delta}} + \lambda_n B.$$

The condition $n \geq c_1 s\sqrt{sB}$ in the proof of Theorem 2.7.3 is satisfied here because we assume that $n^{4/5} \geq c_1 B^4 \widetilde{\sigma}^2 s^5 \log^2 np$ in the statement of the theorem. Using the uniform convergence result of Lemma 2.7.1, with probability at least $1 - \delta$,

$$\|f_0 - \widehat{f} - \widehat{g}_k\|_P^2 - \|f_0 - \widehat{f} - g_k^*\|_P^2 \leq cB^2\sigma\sqrt{\frac{1}{n}\log\frac{Cn}{\delta}} + \lambda_n B + cB^3\sqrt{\frac{c_u^{1/2}}{n^{4/5}}\log\frac{2}{\delta}}$$
$$\leq cB^2\widetilde{\sigma}\sqrt{\frac{c_u^{1/2}}{n^{4/5}}\log\frac{C}{\delta}} + \lambda_n B$$

Finally, plugging in $\lambda_n \leq 768 s\widetilde{\sigma}\sqrt{\frac{1}{n}\log^2 np}$, we obtain

$$\|f_0 - \widehat{f} - \widehat{g}_k\|_P^2 - \|f_0 - \widehat{f} - g_k^*\|_P^2 \leq cB^2\widetilde{\sigma}\sqrt{\frac{c_u^{1/2}}{n^{4/5}}\log\frac{C}{\delta}} + csB\widetilde{\sigma}\sqrt{\frac{1}{n}\log^2 np}$$

$$\|f_0 - \widehat{f} - \widehat{g}_k\|_P^2 - \|f_0 - \widehat{f} - g_k^*\|_P^2 \leq cB^2\widetilde{\sigma}\sqrt{\frac{s^2 c_u^{1/2}}{n^{4/5}}\log^2\frac{Cnp}{\delta}}$$

with probability at least $1-\delta$.

Step 2. We now bound term 3.

$$\begin{aligned}
\|f_0 - \widehat{f} - g_k^*\|_P^2 - \|f_0 - f^* - g_k^*\|_P^2 &\leq \|f_0 - \widehat{f}\|_P^2 - \|f_0 - f^*\|_P^2 - 2\langle f_0 - \widehat{f}, g_k^*\rangle_P + 2\langle f_0 - f^*, g_k^*\rangle_P \\
&\leq cB^2\widetilde{\sigma}\sqrt{\frac{s^5 c_u^{1/2}}{n^{4/5}}\log^2 np} + 2|\langle \widehat{f} - f^*, g_k^*\rangle_P| \\
&\leq cB^2\widetilde{\sigma}\sqrt{\frac{s^5 c_u^{1/2}}{n^{4/5}}\log^2 np} + 2\|\widehat{f} - f^*\|_P\|g_k^*\|_P \\
&\leq cB^2\widetilde{\sigma}\sqrt{\frac{s^5 c_u^{1/2}}{n^{4/5}}\log^2 np} + cB\sqrt{B^2\widetilde{\sigma}\sqrt{\frac{s^5 c_u^{1/2}}{n^{4/5}}\log^2 np}} \\
&\leq cB^2\widetilde{\sigma}^{1/2}\sqrt[4]{\frac{s^5 c_u^{1/2}}{n^{4/5}}\log^2 np}
\end{aligned}$$

with probability at least $1 - \frac{C}{n}$, by Theorem 2.7.3. To obtain the fourth inequality, we used the fact that $\|\widehat{f} - f^*\|_P^2 \leq \|f_0 - \widehat{f}\|_P^2 - \|f_0 - f^*\|_P^2$, which follows from the fact that f^* is the projection of f_0 onto the set of additive convex functions and the set of additive convex functions is convex itself. The last inequality holds because the conditions of the theorem stipulate n is large enough such that $B^2\widetilde{\sigma}\sqrt{\frac{s^5 c_u^{1/2}}{n^{4/5}}\log^2 np} \leq 1$. The same derivation and the same bound likewise holds for term 2.

Step 3. Collecting the results and plugging them into equation (2.7.9), we have, with probability at least $1 - 2\delta$:

$$\|f_0 - f^* - \widehat{g}_k\|_P^2 - \|f_0 - f^* - g_k^*\|_P^2 \leq cB^2\widetilde{\sigma}^{1/2}\sqrt[4]{\frac{s^5 c_u^{1/2}}{n^{4/5}}\log^2\frac{4np}{\delta}}$$

Taking a union bound across the s dimensions completes the result. □

Support Lemmas

Lemma 2.7.1. *Let P be a distribution with a density $p(\mathbf{x})$ which is upper bounded by $c_u \geq 1$. Suppose $n \geq c_1 s\sqrt{sB}$ for some absolute constant c_1. Let δ be small enough such that $\log \frac{2}{\delta} \geq 2$. Then, with probability at least $1-\delta$:*

$$\sup_{f \in \mathcal{C}_B^s} \left| \|f_0 - f\|_n^2 - \|f_0 - f\|_P^2 \right| \leq cB^3 \sqrt{\frac{s^5 c_u^{1/2}}{n^{4/5}} \log \frac{2}{\delta}}$$

where c_1, c are some absolute constants.

Proof. Let $\mathcal{G}$ denote the off-centered set of convex functions, that is, $\mathcal{G} \equiv \mathcal{C}_B^s - f_0$. Note that if $h \in \mathcal{G}$, then $\|h\|_\infty = \|f_0 - f\|_\infty \leq 2sB$. There exists an ϵ-bracketing of $\mathcal{G}$, and by Corollary 2.7.3, the bracketing has log-size at most $\log N_{[]}(\epsilon, \mathcal{C}_B^s, L_1(P)) \leq sK^{**}\left(\frac{4sBc_u}{\epsilon}\right)^{1/2}$ for all $\epsilon < \epsilon_3 sBc_u$, for some constant ϵ_3.

For a particular function $h \in \mathcal{G}$, there exist an ϵ-bracket h_U, h_L. We construct $\psi_L \equiv \min(|h_U|, |h_L|)$ and $\psi_U \equiv \max(|h_U|, |h_L|)$ so that

$$\psi_L^2 \leq h^2 \leq \psi_U^2.$$

If x is such that $h_U^2(x) \geq h_L^2(x)$, then $\psi_U^2(x) - \psi_L^2(x) = h_U^2(x) - h_L^2(x)$. If x is such that $h_U^2(x) \leq h_L^2(x)$, then $\psi_U^2(x) - \psi_L^2(x) = h_L^2(x) - h_U^2(x)$. We can then bound the $L_1(P)$ norm of $\psi_U^2 - \psi_L^2$ as

$$\begin{aligned}
\int (\psi_U^2(x) - \psi_L^2(x)) p(x) dx &= \int |h_U^2(x) - h_L^2(x)| p(x) dx \\
&\leq \int |h_U(x) - h_L(x)| \, |h_U(x) + h_L(x)| p(x) dx \\
&\leq 4sB\epsilon
\end{aligned}$$

Now we can bound $\|h\|_n^2 - \|h\|_P^2$ as

$$\frac{1}{n}\sum_{i=1}^n \psi_L(X_i)^2 - \mathbb{E}\psi_U(X)^2 \leq \|h\|_n^2 - \|h\|_P^2 \leq \frac{1}{n}\sum_{i=1}^n \psi_U(X_i)^2 - \mathbb{E}\psi_L(X)^2 \qquad (2.7.10)$$

Since $\psi_L(X_i)^2$ and $\psi_U(X_i)^2$ are bounded random variables with upper bound $(2sB)^2$, Hoeffding's inequality and union bound give that, with probability at least $1-\delta$,, for all ψ_L (and likewise ψ_U)

$$\left| \frac{1}{n}\sum_{i=1}^n \psi_L(X_i)^2 - \mathbb{E}\psi_L(X)^2 \right| \leq (2sB)^2 \sqrt{\frac{sK^{**}(4sBc_u)^{1/2}}{\epsilon^{1/2} 2n} + \frac{\log \frac{2}{\delta}}{2n}}$$

To simplify the expression, we will suppose that $\frac{sK^{**}(4sBc_u)^{1/2}}{\epsilon^{1/2}} \geq 2$ and that $\log \frac{2}{\delta} \geq 2$. The second supposition holds by assumption in the theorem. Once we calculate the proper values of ϵ, we will verify that these first supposition holds under the assumption of the theorem also. Under these two suppositions, we have

$$\left| \frac{1}{n} \sum_{i=1}^{n} \psi_L(X_i)^2 - \mathbb{E}\psi_L(X)^2 \right| \leq (2sB)^2 \sqrt{\frac{sK^{**}(sBc_u)^{1/2} \log \frac{2}{\delta}}{\epsilon^{1/2} n}} \tag{2.7.11}$$

Plugging this into equation (2.7.10) above, we have that:

$$\begin{aligned} &\mathbb{E}\psi_L(X)^2 - \mathbb{E}\psi_U(X)^2 - (2sB)^2 \sqrt{\frac{sK^{**}(sBc_u)^{1/2} \log \frac{2}{\delta}}{\epsilon^{1/2} n}} \\ &\leq \|h\|_n^2 - \|h\|_P^2 \leq \mathbb{E}\psi_U(X)^2 - \mathbb{E}\psi_L(X)^2 + (2sB)^2 \sqrt{\frac{sK^{**}(sBc_u)^{1/2} \log \frac{2}{\delta}}{\epsilon^{1/2} n}}. \end{aligned}$$

Using our bound on the $L_1(P)$ norm of $\psi_U^2 - \psi_L^2$, we have

$$-4sB\epsilon - (2sB)^2 \sqrt{\frac{sK^{**}(sBc_u)^{1/2} \log \frac{2}{\delta}}{\epsilon^{1/2} n}} \leq \|h\|_n^2 - \|h\|_P^2 \leq 4sB\epsilon + (2sB)^2 \sqrt{\frac{sK^{**}(sBc_u)^{1/2} \log \frac{2}{\delta}}{\epsilon^{1/2} n}}$$

We choose $\epsilon = \left(\frac{(sB)^2 sK^{**}(sBc_u)^{1/2}}{n} \right)^{2/5}$. This choice of ϵ is a bit suboptimal but it is convenient and it is sufficient for our result.

One can easily verify that $\epsilon \leq sB\epsilon_3 c_u$ when $n \geq c_1 s\sqrt{sB}$ for some absolute constant c_1, thus, the ϵ-bracketing number we used is valid.

One can also verify that the condition above equation 2.7.11 is satisfied.

We have then that, with probability at least $1 - \delta$,

$$\sup_{h \in \mathcal{G}} \left| \|h\|_n^2 - \|h\|_P^2 \right| \leq cB^3 \sqrt{\frac{s^5 c_u^{1/2} \log \frac{2}{\delta}}{n^{4/5}}}$$

The theorem follows immediately.

□

Lemma 2.7.2. *Let f_0 and f^* be defined as in Section 2.7. Define $\bar{f}^* = \frac{1}{n} \sum_{i=1}^{n} f^*(X_i)$. Then, with probability at least $1 - 2\delta$,*

$$\left| \|f_0 - f^*\|_n^2 - \|f_0 - f^* + \bar{f}^*\|_n^2 \right| \leq c(sB)^2 \frac{1}{n} \log \frac{4}{\delta}$$

Proof. We decompose the empirical norm as

$$\begin{aligned}\|f_0 - f^* + \bar{f}^*\|_n^2 &= \|f_0 - f^*\|_n^2 + 2\langle f_0 - f^*, \bar{f}^*\rangle + \bar{f}^{*2} \\ &= \|f_0 - f^*\|_n^2 + 2\bar{f}^*\langle f_0 - f^*, \mathbf{1}\rangle_n + \bar{f}^{*2} \\ &= \|f_0 - f^*\|_n^2 + 2\bar{f}^*\bar{f}_0 - \bar{f}^{*2}.\end{aligned}$$

Now $\bar{f}^* = \frac{1}{n}\sum_{i=1}^n f^*(X_i)$ is the average of n bounded mean-zero random variables and therefore, with probability at least $1-\delta$, $|\bar{f}^*| \leq 4sB\sqrt{\frac{1}{n}\log\frac{2}{\delta}}$. The same reasoning likewise applies to $\bar{f}_0 = \frac{1}{n}\sum_{i=1}^n f_0(X_i)$.

We take a union bound and get that, with probability at least $1-2\delta$,

$$\begin{aligned}|\bar{f}^*||\bar{f}_0| &\leq c(sB)^2\frac{1}{n}\log\frac{2}{\delta} \\ \bar{f}^{*2} &\leq c(sB)^2\frac{1}{n}\log\frac{2}{\delta}\end{aligned}$$

Therefore, with probability at least $1-2\delta$,

$$\|f_0 - f^*\|_n^2 - c(sB)^2\frac{1}{n}\log\frac{2}{\delta} \leq \|f_0 - f^* + \bar{f}^*\|_n^2 \leq \|f_0 - f^*\|_n^2 + c(sB)^2\frac{1}{n}\log\frac{2}{\delta}$$

□

Supporting Technical Material

Detail for Proof of Theorem 2.3.1

Let $p(\mathbf{x}_{-k} \,|\, x_k), f(\mathbf{x}), r(\mathbf{x}_{-k})$ be defined as in the proof of Theorem 2.3.1.

We claim that

$$\partial_{x_k}\int_{\mathbf{x}_{-k}} p(\mathbf{x}_{-k} \,|\, x_k)\,(f(\mathbf{x}) - r(\mathbf{x}_{-k}))\,d\mathbf{x}_{-k} = \int_{\mathbf{x}_{-k}} \partial_{x_k}\Big(p(\mathbf{x}_{-k} \,|\, x_k)\,(f(\mathbf{x}) - r(\mathbf{x}_{-k}))\Big)d\mathbf{x}_{-k}$$

And likewise for the second derivative.

The first derivative of the integrand is

$$p'(\mathbf{x}_{-k} \,|\, x_k)\,(f(\mathbf{x}) - r(\mathbf{x}_{-k})) + p(\mathbf{x}_{-k} \,|\, x_k)f'(x_k, \mathbf{x}_{-k}).$$

$f(\mathbf{x})$ is continuous and bounded and $p'(\mathbf{x}_{-k} \,|\, x_k)$ is bounded for all $\mathbf{x}_{-k}$ and all $x_k \in [0,\epsilon) \cup (1-\epsilon, 1]$ by boundary flatness. Thus, $p'(\mathbf{x}_{-k} \,|\, x_k)f(\mathbf{x})$ is bounded for all $\mathbf{x}_{-k}$ and all $x_k \in [0,\epsilon) \cup (1-\epsilon, 1]$.

$r(\mathbf{x}_{-k})p(\mathbf{x})$ is integrable, and since $\inf_{\mathbf{x}} p(\mathbf{x}) > 0$, $r(\mathbf{x}_{-k})$ is integrable. Since $p'(\mathbf{x}_{-k} \,|\, x_k)$ is bounded, $r(\mathbf{x}_{-k})p'(\mathbf{x}_{-k} \,|\, x_k) \leq |r(\mathbf{x}_{-k})|M$ for some constant M for all $\mathbf{x}_{-k}$ and $x_k \in [0, \epsilon) \cup (1 - \epsilon, 1]$.

$f'(x_k, \mathbf{x}_{-k})$ is continuous and thus bounded, therefore, $p(\mathbf{x}_{-k} \,|\, x_k)f'(x_k, \mathbf{x}_{-k})$ is bounded for all $\mathbf{x}_{-k}$ and $x_k \in [0, \epsilon) \cup (1 - \epsilon, 1]$.

This verifies that, for all $\mathbf{x}_{-k}$ and for all $x_k \in [0, \epsilon) \cup (1 - \epsilon, 1]$, the first derivative is less than $M|r(\mathbf{x}_{-k})| + C$, which is integrable. By dominated convergence theorem, we can thus exchange the derivative with the integral.

The second derivative of the integrand is

$$p''(\mathbf{x}_{-k} \,|\, x_k)\left(f(\mathbf{x}) - r(\mathbf{x}_{-k})\right) + 2p'(\mathbf{x}_{-k} \,|\, x_k)f'(x_k, \mathbf{x}_{-k}) + p(\mathbf{x}_{-k} \,|\, x_k)f''(x_k, \mathbf{x}_{-k})$$

We just need to remember that $p''(\mathbf{x}_{-k} \,|\, x_k)$ is bounded for all $\mathbf{x}_{-k}$ and $x_k \in [0, \epsilon) \cup (1 - \epsilon, 1]$ by boundary flatness, that $f''(x_k, \mathbf{x}_{-k})$ is bounded, and the same argument applies.

Uniqueness of the Additive Components

Lemma 2.7.3. *Let $p(\mathbf{x})$ be a positive density over $[0, 1]^p$. Let $f(\mathbf{x}) = \sum_{j=1}^{p} f_j(x_j)$ and $h(\mathbf{x}) = \sum_{j=1}^{d} h_j(x_j)$ be two additive functions such that $\mathbb{E}(f(X) - h(X))^2 = 0$. Suppose also that $\mathbb{E}f_j(X_j) = 0, \mathbb{E}h_j(X_j) = 0$ for all j. Then, it must be that $\mathbb{E}(f_j(X_j) - h_j(X_j))^2 = 0$ for all j.*

Proof. Let $\phi(\mathbf{x}) = f(\mathbf{x}) - g(\mathbf{x})$ and it is clear that $\phi(\mathbf{x}) = \sum_{j=1}^{d} \phi_j(x_j)$ with $\phi_j(x_j) = f_j(x_j) - h_j(x_j)$ and $\mathbb{E}\phi_j(X_j) = 0$. It is also immediate that $E\phi(X)^2 = 0$.

Let P be the probability measure induced by the density $p(\mathbf{x})$, so that $P(A) = \int_A p(\mathbf{x})d\lambda$ where λ is the Lebesgue measure. Since $p > 0$, $\lambda(A) > 0$ implies that $P(A) > 0$ as well.

For sake of contradiction, let us assume that for some j, $\mathbb{E}\phi_j(X_j)^2 > 0$. Then, $P(A_j) > 0$ where $A_j = \{\mathbf{x} \in [0, 1]^p : \phi_j(x_j) > 0\}$. To see this, suppose that $\phi_j \leq 0$ almost surely. Then, $\mathbb{E}\phi_j = 0$ implies that $\phi_j = 0$ almost surely, contradicting the assumption that $\mathbb{E}\phi_j(X_j)^2 > 0$.

For $j' \neq j$, define $B_{j'} = \{\mathbf{x} \in [0, 1]^p : \phi_{j'}(x_{j'}) \geq 0\}$. $P(B_{j'}) > 0$ because, if not, then $\phi_{j'} < 0$ almost surely and that contradicts the $\mathbb{E}\phi_{j'}(X_{j'}) = 0$ assumption.

Since the probability measure P is absolutely continuous with respect to the Lebesgue measure on $[0,1]^p$, $\lambda(A_j) > 0$. Let λ_1 be the one dimensional Lebesgue measure on $[0,1]$ and let $A_j^1 = \{x_j \in [0,1] : \phi_j(x_j) > 0\}$. From the fact that $\lambda(A_j) > 0$, and that $A_j = A_j^1 \times [0,1]^{p-1}$, $\lambda_1(A_j^1) > 0$. Same reasoning show that $\lambda_1(B_{j'}^1) > 0$ where $B_{j'}^1$ is similarly defined.

$A_j \cap (\bigcap_{j'} B_{j'}) = A_j^1 \times \prod_{j'} B_{j'}^1$. Therefore, $\lambda(A_j \cap (\bigcap_{j'} B_{j'})) = \lambda_1(A_j^1) \prod_{j'} \lambda_1(B_{j'}) > 0$. Since the density of P is positive, $P(A_j \cap (\bigcap_{j'} B_{j'})) > 0$ and since $\phi > 0$ on this event, we conclude that $\mathbb{E}\phi(X)^2 > 0$, thus giving us the desired contradiction.

□

Concentration of Measure

A sub-exponential random variable is the square of a sub-Gaussian random variable Vershynin (2010).

Proposition 2.7.1. *(Subexponential Concentration Vershynin (2010)) Let $X_1, ..., X_n$ be zero-mean independent subexponential random variables with subexponential scale K. Then*

$$P(|\frac{1}{n}\sum_{i=1}^{n} X_i| \geq \epsilon) \leq 2\exp\left[-cn\min\left(\frac{\epsilon^2}{K^2}, \frac{\epsilon}{K}\right)\right]$$

where $c > 0$ is an absolute constant.

For uncentered subexponential random variables, we can use the following fact. If X_i subexponential with scale K, then $X_i - \mathbb{E}[X_i]$ is also subexponential with scale at most $2K$. Restating, we can set

$$c\min\left(\frac{\epsilon^2}{K^2}, \frac{\epsilon}{K}\right) = \frac{1}{n}\log\frac{1}{\delta}.$$

Thus, with probability at least $1-\delta$, the deviation is at most

$$K\max\left(\sqrt{\frac{1}{cn}\log\frac{C}{\delta}}, \frac{1}{cn}\log\frac{C}{\delta}\right).$$

Corollary 2.7.1. *Let $W_1, ..., W_n$ be n independent sub-Gaussian random variables with sub-Gaussian scale σ. Then, for all $n > n_0$, with probability at least $1 - \frac{1}{n}$,*

$$\frac{1}{n}\sum_{i=1}^{n} W_i^2 \leq c\sigma^2.$$

Proof. Using the subexponential concentration inequality, we know that, with probability at least $1 - \frac{1}{n}$,

$$\left| \frac{1}{n} \sum_{i=1}^{n} W_i^2 - \mathbb{E}W^2 \right| \leq \sigma^2 \max \left(\sqrt{\frac{1}{cn} \log \frac{C}{\delta}}, \frac{1}{cn} \log \frac{C}{\delta} \right).$$

First, let $\delta = \frac{1}{n}$. Suppose n is large enough such that $\frac{1}{cn} \log Cn < 1$. Then, we have, with probability at least $1 - \frac{1}{n}$,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} W_i^2 &\leq c\sigma^2 \Big(1 + \sqrt{\frac{1}{cn} \log Cn} \Big) \\ &\leq 2c\sigma^2. \end{aligned}$$

□

Sampling Without Replacement

Lemma 2.7.4. *(Serfling (1974)) Let $x_1, ..., x_N$ be a finite list, $\bar{x} = \mu$. Let $X_1, ..., X_n$ be sampled from x without replacement.*

Let $b = \max_i x_i$ and $a = \min_i x_i$. Let $r_n = 1 - \frac{n-1}{N}$. Let $S_n = \sum_i X_i$. Then we have that

$$\mathbb{P}(S_n - n\mu \geq n\epsilon) \leq \exp \left(-2n\epsilon^2 \frac{1}{r_n (b-a)^2} \right).$$

Corollary 2.7.2. *Suppose $\mu = 0$.*

$$\mathbb{P} \left(\frac{1}{N} S_n \geq \epsilon \right) \leq \exp \left(-2N\epsilon^2 \frac{1}{(b-a)^2} \right)$$

And, by union bound, we have that

$$\mathbb{P} \left(|\frac{1}{N} S_n| \geq \epsilon \right) \leq 2 \exp \left(-2N\epsilon^2 \frac{1}{(b-a)^2} \right)$$

A simple restatement is that with probability at least $1 - \delta$, the deviation $|\frac{1}{N} S_n|$ is at most $(b-a)\sqrt{\frac{1}{2N} \log \frac{2}{\delta}}$.

Proof.

$$\mathbb{P} \left(\frac{1}{N} S_n \geq \epsilon \right) = \mathbb{P} \left(S_n \geq \frac{N}{n} n\epsilon \right) \leq \exp \left(-2n \frac{N^2}{n^2} \epsilon^2 \frac{1}{r_n (b-a)^2} \right).$$

We note that $r_n \leq 1$ always, and $n \leq N$ always. Thus,

$$\exp\left(-2n\frac{N^2}{n^2}\epsilon^2\frac{1}{r_n(b-a)^2}\right) \leq \exp\left(-2N\epsilon^2\frac{1}{(b-a)^2}\right)$$

completing the proof. □

Bracketing Numbers for Convex Functions

Definition 2.7.1. Let $\mathcal{C}$ be a set of functions. For a given ϵ and metric ρ (which we take to be L_2 or $L_2(P)$), we define an *ϵ-bracketing* of $\mathcal{C}$ to be a set of pairs of functions $\{(f_L, f_U)\}$ satisfying (1) $\rho(f_L, f_U) \leq \epsilon$ and (2) for any $f \in \mathcal{C}$, there exist a pair (f_L, f_U) where $f_U \geq f \geq f_L$.

We let $N_{[]}(\epsilon, \mathbf{C}, \rho)$ denote the size of the smallest bracketing of $\mathcal{C}$

Proposition 2.7.2. *(Proposition 16 in Kim and Samworth (2014)) Let $\mathcal{C}$ be the set of convex functions supported on $[-1,1]^d$ and uniformly bounded by B. Then there exist constants ϵ_3 and K^{**}, dependent on d, such that*

$$\log N_{[]}(\epsilon, \mathcal{C}, L_2) \leq K^{**}\left(\frac{4B}{\epsilon}\right)^{d/2}$$

for all $\epsilon \in (0, B\epsilon_3]$.

It is easy to extend Kim and Samworth's result to the $L_1(P)$ norm for a distribution P with a bounded density $p(x)$.

Proposition 2.7.3. *Let P be a distribution with a density $p(x)$ and suppose $p(x) \leq c_u$ for some constant $c_u > 0$. Let $\mathcal{C}, B, \epsilon_3, K^{**}$ be defined as in Proposition 2.7.2. Then,*

$$\log N_{[]}(\epsilon, \mathcal{C}, L_1(P)) \leq K^{**}\left(\frac{4Bc_u}{\epsilon}\right)^{d/2}$$

for all $\epsilon \in (0, B\epsilon_3 c_u]$.

Proof. Let $\mathcal{C}_{\epsilon/c_u}$ be an ϵ/c_u-bracketing with respect to the L_2 norm. Because $\epsilon \in (0, B\epsilon_3 c_u]$, it is clear that $\epsilon/c_u \in (0, B\epsilon]$. Then, the log-size of $\mathcal{C}_{\epsilon/c_u}$ is at most $K^{**}\left(\frac{4Bc_u}{\epsilon}\right)^{d/2}$ by Proposition 2.7.3.

Let $(f_L, f_U) \in \mathcal{C}_{\epsilon/c_u}$. Then we have that:

$$\begin{aligned}\|f_L - f_U\|_{L_1(P)} &= \int |f_L(x) - f_U(x)| p(x) dx \\ &\leq \left(\int |f_L(x) - f_U(x)|^2 dx\right)^{1/2} \left(\int p(x)^2 dx\right)^{1/2} \\ &\leq \left(\int |f_L(x) - f_U(x)|^2 dx\right)^{1/2} c_u \\ &\leq \|f_L - f_U\|_{L_2} c_u \leq \epsilon\end{aligned}$$

On the third line, we used the fact that $\left(\int p(x)^2 dx\right)^2 \leq c_u$. □

It is also simple to extend the bracketing number result to additive convex functions. As before, let $\mathcal{C}_B^s$ be the set of additive convex functions with s components, each component of which is bounded by B.

Corollary 2.7.3. *Let P be a distribution with a density $p(x)$ and suppose $p(x) \leq c_u$. Let B, ϵ_3, K^{**} be defined as in Proposition 2.7.2. Then,*

$$\log N_{[]}(\epsilon, \mathcal{C}_B^s, L_1(P)) \leq sK^{**} \left(\frac{4sBc_u}{\epsilon}\right)^{1/2}$$

for all $\epsilon \in (0, sB\epsilon_3 c_u]$.

Proof. Let $f \in \mathcal{C}^s$. We can construct an ϵ-bracketing for f through ϵ/s-bracketings (with respect to the $L_1(P)$ norm) for each of the components $\{f_k\}_{k=1,\ldots,s}$:

$$f_U = \sum_{k=1}^{s} f_{Uk} \qquad f_L = \sum_{k=1}^{s} f_{Lk}$$

It is clear that $f_U \geq f \geq f_L$. It is also clear that $\|f_U - f_L\|_{L_1(P)} \leq \sum_{k=1}^{s} \|f_{Uk} - f_{Lk}\|_{L_1(P)} \leq \epsilon$. □

The following result follows from Corollary 2.7.3 directly by a union bound.

Corollary 2.7.4. *Let $X_1, ..., X_n$ be random samples from a distribution P and suppose P has a density $p(x)$ bounded by c_u. Let $1 > \delta > 0$. Let $\mathcal{C}_\epsilon^s$ be an ϵ-bracketing of $\mathcal{C}_B^s$ with respect to the $L_1(P)$-norm whose size is at most $N_{[]}(\epsilon, \mathcal{C}^s, L_1(P))$. Let $\epsilon \in (0, sB\epsilon_3 c_u]$.*

Then, with probability at least $1 - \delta$, for all pairs $(f_L, f_U) \in \mathcal{C}_\epsilon^s$, we have that

$$\frac{1}{n} \sum_{i=1}^{n} |f_L(X_i) - f_U(X_i)| \leq \epsilon + \epsilon_{n,\delta}$$

where

$$\epsilon_{n,\delta} \equiv 2sB\sqrt{\frac{\log N_{[]}(\epsilon, \mathcal{C}^s, L_1(P)) + \log\frac{2}{\delta}}{2n}} = 2sB\sqrt{\frac{sK^{**}(4sBc_u)^{1/2}}{\epsilon^{1/2}2n} + \frac{1}{2n}\log\frac{2}{\delta}}.$$

Proof. Noting that $|f_L(X_i) - f_U(X_i)|$ is at most $2sB$ and that there are at most $N_{[]}(\epsilon, \mathcal{C}^s, L_1(P))$ pairs (f_L, f_U), the inequality follows from a direct application of a union bound and Hoeffding's Inequality. □

To make the expression in this corollary easier to work with, we derive an upper bound for $\epsilon_{n,\delta}$. Suppose

$$\frac{sK^{**}(4sBc_u)^{1/2}}{\epsilon^{1/2}} \geq 2 \quad \text{and} \quad \log\frac{2}{\delta} \geq 2. \tag{2.7.12}$$

Then we have that

$$\epsilon_{n,\delta} \leq 2sB\sqrt{\frac{sK^{**}(4sBc_u)^{1/2}\log\frac{2}{\delta}}{2\epsilon^{1/2}n}} = 2sB\sqrt{\frac{sK^{**}(sBc_u)^{1/2}\log\frac{2}{\delta}}{\epsilon^{1/2}n}}$$

2.8 Gaussian Example

Let H be a positive definite matrix and let $f(x_1, x_2) = H_{11}x_1^2 + 2H_{12}x_1x_2 + H_{22}x_2^2 + c$ be a quadratic form where c is a constant such that $\mathbb{E}[f(X)] = 0$. Let $X \sim N(0, \Sigma)$ be a random bivariate Gaussian vector with covariance $\Sigma = [1, \alpha; \alpha, 1]$

Proposition 2.8.1. *Let $f_1^*(x_1) + f_2^*(x_2)$ be the additive projection of f under the bivariate Gaussian distribution. That is,*

$$f_1^*, f_2^* \equiv \underset{f_1, f_2}{argmin} \left\{\mathbb{E}\left(f(X) - f_1(X_1) - f_2(X_2)\right)^2 \,:\, \mathbb{E}[f_1(X_1)] = \mathbb{E}[f_2(X_2)] = 0\right\}$$

Then, we have that

$$f_1^*(x_1) = \left(\frac{T_1 - T_2\alpha^2}{1 - \alpha^4}\right)x_1^2 + c_1$$
$$f_2^*(x_2) = \left(\frac{T_2 - T_1\alpha^2}{1 - \alpha^4}\right)x_2^2 + c_2$$

where $T_1 = H_{11} + 2H_{12}\alpha + H_{22}\alpha^2$ and $T_2 = H_{22} + 2H_{12}\alpha + H_{11}\alpha^2$ and c_1, c_2 are constants such that $\mathbb{E}[f_1^*(X_1)] = \mathbb{E}[f_2^*(X_2)] = 0$.

Proof. By Lemma 2.3.1, we need only verify that f_1^*, f_2^* satisfy

$$f_1^*(x_1) = \mathbb{E}[f(X) - f_2^*(X_2) \,|\, x_1]$$
$$f_2^*(x_2) = \mathbb{E}[f(X) - f_1^*(X_1) \,|\, x_2].$$

Let us guess that f_1^*, f_2^* are quadratic forms $f_1^*(x_1) = a_1 x_1^2 + c_1$, $f_2^*(x_2) = a_2 x_2^2 + c_2$ and verify that there exist a_1, a_2 to satisfy the above equations. Since we are not interested in constants, we define $\simeq$ to be equality up to a constant. Then,

$$\begin{aligned}
&\mathbb{E}[f(X) - f_2^*(X_2) \,|\, x_1] \\
&\simeq \mathbb{E}[H_{11}X_1^2 + 2H_{12}X_1X_2 + H_{22}X_2^2 - a_2X_2^2 \,|\, x_1] \\
&\simeq H_{11}x_1^2 + 2H_{12}x_1\mathbb{E}[X_2 \,|\, x_1] + H_{22}\mathbb{E}[X_2^2 \,|\, x_1] - a_2\mathbb{E}[X_2^2 \,|\, x_1] \\
&\simeq H_{11}x_1^2 + 2H_{12}\alpha x_1^2 + H_{22}\alpha^2 x_1^2 - a_2\alpha^2 x_1^2 \\
&\simeq (H_{11} + 2H_{12}\alpha + H_{22}\alpha^2 - a_2\alpha^2)x_1^2.
\end{aligned}$$

Likewise, we have that

$$\mathbb{E}[f(X) - f_1^2(X_1) \,|\, x_2] \simeq (H_{22} + 2H_{12}\alpha + H_{22}\alpha^2 - a_1\alpha^2)x_2^2.$$

Thus, a_1, a_2 need only satisfy the linear system

$$T_1 - a_2\alpha^2 = a_1$$
$$T_2 - a_1\alpha^2 = a_2$$

where $T_1 = H_{11} + 2H_{12}\alpha + H_{22}\alpha^2$ and $T_2 = H_{22} + 2H_{12}\alpha + H_{11}\alpha^2$. It is then simple to solve the system and verify that a_1, a_2 are as specified. □

CHAPTER 3

HIGH DIMENSIONAL CONCAVE UTILITY ESTIMATION

3.1 Introduction

Many human behaviors can be modeled as a consumer selecting one item to purchase from among a set of alternatives. Examples include buying a product on Amazon, choosing the bus or car for commuting Ortuzar and Willumsen (1994), deciding where to buy a house Nechyba and Strauss (1998), and even choosing where to commit a crime Bernasco and Block (2009). The discrete choice model (DCM) originated in econometrics McFadden (1973) as a general method to model such finite choice problems. The DCM measures the attractiveness of item i to consumer n by a utility function $f(\mathbf{x}_i, \mathbf{s}_n)$ where $\mathbf{x}_i, \mathbf{s}_n$ are feature vectors of the item and the consumer, respectively. The consumer is more likely to pick item i over the alternatives if the utility $f(\mathbf{x}_i, \mathbf{s}_n)$ is higher. The utility function in the DCM is estimated from a dataset of purchases; each purchase consists of a consumer, a set of items, and the consumer's choice from that set. The AI and machine learning communities have in recent years rediscovered the DCM as a form of *preference learning* Fürnkranz and Hüllermeier (2010); Chu and Ghahramani (2005).

Because it has become easier to extract and store information digitally, the number of features in a modern dataset is often large, possibly larger than the number of samples. Variable selection becomes important, where an estimation technique must select and use only a small set of relevant variables to avoid the well known curse of dimensionality. Variable selection among the item features $\mathbf{x}_i$ is especially important in the DCM, as people tend to make decisions based on a few important cues or factors Shah and Oppenheimer (2008). Good variable selection methods give insight into how consumers make choices.

We assume that the utility function $V(\mathbf{x}_i, \mathbf{s}_n)$ is decomposed as $f(\mathbf{x}_i) + h(\mathbf{s}_n)$ and focus on the estimation of $f(\mathbf{x}_i)$. We suppose $f(\mathbf{x}_i)$ obeys certain shape-constraints, mainly concavity. We do not assume that f is additive, but we show that for the purpose of screening out irrelevant variables, it is safe to approximate the possibly non-additive f with an additive concave model, followed by a sequence of decoupled convex models to catch non-concave residuals.

We prove that this procedure, in the population setting, is *faithful* in that it will not erroneously mark a relevant variable as irrelevant. The assumptions we make

on the underlying density are mild, and do not restrict correlations between the variables. This is in contrast to linear models where, if the true function is non-linear, one must make stronger covariance structure assumptions in order to provide the same guarantee. While estimation of a low-dimensional concave utility function for the DCM is studied by Matzkin using parametric distributional assumptions Matzkin (1991), we are unaware of previous results on variable selection in the DCM in the high dimensional nonparametric setting that we study in this paper.

The utility function is often assumed to be linear Nechyba and Strauss (1998); McFadden et al. (1978) but a concavity assumption is less restrictive. In many economics applications, the concavity assumption is popular and natural because of the law of diminishing returns. For example, Nechyba and Strauss Nechyba and Strauss (1998) represent the attractiveness of a community in the DCM with features such as per-pupil school spending. The law of diminishing returns in this case states that once a school spends enough per pupil, further spending will be less effective and thus effect a smaller increase in a household's utility.

Though our estimation method is a nonparametric generalization of the linear model, it requires no additional tuning parameters, such as the smoothing bandwidth that makes local polynomial methods difficult. Concavity (and other similar shape-constraints) thus offers an attractive computational compromise between a parametric and fully nonparametric model. We formulate a convex optimization in the infinite dimensional constraint space of concave functions, which reduces to a finite dimensional space of piecewise linear functions.

3.2 Discrete Choice Model

In discrete choice model, each consumer n chooses one item out of a set $\mathcal{A}_n$ of alternatives based on the utility-maximization principle: each item $i \in \mathcal{A}_n$ has a utility U_{ni} and the consumer chooses item i if $U_{ni} \geq U_{nj}$ for all $j \in \mathcal{A}_n$ (ties broken arbitrarily). The utility U_{ni} is unobservable but is assumed to equal a function of some observable features of the items and of the consumer plus noise

$$U_{ni} = V_{ni} + \epsilon_{ni} = V(\mathbf{x}_i, \mathbf{s}_n) + \epsilon_{ni}.$$

The vector $\mathbf{x}_i$ denotes features of item i, $\mathbf{s}_n$ denotes features of consumer n, and ϵ_{ni} denotes the noise term. The probability of consumer n choosing item i depends on the assumptions on the distribution of the noise vector $\epsilon_n = (\epsilon_{n1}, ..., \epsilon_{n|\mathcal{A}_n|})$.

$$P_{ni} = \mathbb{P}(U_{ni} > U_{nj}, \forall j \neq i) = \mathbb{P}(V_{ni} + \epsilon_{ni} > V_{nj} + \epsilon_{nj}, \forall j \neq i)$$

For example, $\epsilon_n \sim_{iid}$ Gaussian yields the probit model, $\epsilon \sim_{iid}$ extreme value yields the logit model. In this paper, we consider the logit model, also known as the Bradley-Terry model. The probability consumer n chooses item i under the logit model has the expression:

$$P_{ni} = \frac{\exp(V_{ni})}{\sum_{j\in\mathcal{A}_n}\exp(V_{nj})} = \frac{\exp(V(\mathbf{x}_i,\mathbf{s}_n))}{\sum_{j\in\mathcal{A}_n}\exp(V(\mathbf{x}_j,\mathbf{s}_n))}$$

We follow the standard assumption that the representative utility function $V(\mathbf{x}_i,\mathbf{s}_n)$ is decomposed additively as $f(\mathbf{x}_i)+h(\mathbf{s}_n)$ and focus on the estimation of $f(\mathbf{x}_i)$, similar to Chu and Ghahramani Chu and Ghahramani (2005). The results in this paper hold regardless of how one chooses to model $h(\mathbf{s}_n)$.

We assume that $f(\mathbf{x}_i)$ is concave, which is strictly more general than the usual linear assumption. Concavity is justified by the principle of diminishing returns present in many economics applications. [1] We can then model the consumer choices by

$$\mathbb{P}(\text{consumer } n \text{ chooses } i \,|\, \mathcal{A}_n) = \frac{\exp(f(\mathbf{x}_i)+h(\mathbf{s}_n))}{\sum_{j\in\mathcal{A}_n}\exp(f(\mathbf{x}_j)+h(\mathbf{s}_n))}$$

The unknown f, α, h can be estimated from a dataset of purchases. We represent each purchase by a vector $\mathbf{y}_n = (y_{ni})_{i\in\mathcal{A}_n}$; $y_{ni}=1$ iff consumer n chooses item i. For notational simplicity, we assume that each consumer makes exactly one purchase. It is straightforward to extend the model to cases where each consumer makes multiple purchases.

Given N purchases, the likelihood under the logit DCM is $\frac{1}{N}\sum_{i=1}^{N}\ell(\mathbf{y}_n \,|\, \mathbf{X}_{\mathcal{A}_n},\mathbf{s}_n)$ where

$$\ell(\mathbf{y}_n|\mathbf{X}_{\mathcal{A}_n},\mathbf{s}_n) = \sum_{i\in\mathcal{A}_n} y_{ni}(f(\mathbf{x}_i)+h(\mathbf{s}_n)) - \log(\sum_{i\in\mathcal{A}_n}\exp\left(f(\mathbf{x}_i)+h(\mathbf{s}_n)\right)$$

3.3 Additive Faithfulness

Let d_1 and d_2 denote the number of features in $\mathbf{x}_i$ and $\mathbf{z}_i$ respectively. In the high-dimensional setting where d_1 and d_2 are large, it is necessary to select a small subset $S_1 \subset \{1,..,d_1\}$ and $S_2 \subset \{1,...,d_2\}$ such that $f(\mathbf{x}_i) \approx f(\mathbf{x}_{S_1,i})$ where $\mathbf{x}_{S_1,i}$ is the restriction of the vector $\mathbf{x}_i$ to coordinates in S_1.

[1] our results readily apply to the estimation of $h(\mathbf{s}_n)$ in the cases where h can be assumed concave.

The lasso effectively tackles high-dimensional problems by adding an ℓ_1 penalty on the linear coefficients to the likelihood maximization. The natural extension for high-dimensional concave function estimation is to add to the likelihood a group ℓ_1 penalty on the *subgradients* of f:

$$\begin{aligned} \underset{\mathbf{f},\beta,\gamma,\mathbf{h}}{\text{minimize}} \quad & -\sum_{n=1}^{N} \ell(\mathbf{y}_n|\mathbf{X}_{\mathcal{A}_n}, \mathbf{s}_n) + \lambda \sum_{k=1}^{d_1} ||\beta_{k.}||_\infty \\ \text{subject to} \quad & f_j \leq f_i + \beta_i^T(x_j - x_i), \text{for all } i,\, j \end{aligned}$$

where f_i is the estimated function value $f(\mathbf{x}_i)$ and the vector $\beta_i \in \mathbb{R}^{d_1}$ is the subgradient at $\mathbf{x}_i$.

This convex optimization problem has $O(Mp)$ variables and $O(M^2)$ constraints, leading to a potentially cumbersome and computationally inefficient method. In the following section, we will use additive functions to approximate $f(\cdot)$ and argue that concave functions are additively faithful with respect to variable selection, under very mild conditions. The resulting variable selection framework is much more computationally efficient.

Additive Faithfulness of Concave Functions

The *additive approximation* to a multivariate function f is a sum of one-dimensional functions f_k such that $\sum_{k=1}^{d} f_k(\mathbf{x}_k)$ approximates $f(\mathbf{x})$. In general, if the true model is non-additive, an additive approximation may introduce false negatives and cause potential misspecification problems. However, we show that concave functions have an unique property: as long as the true function we approximate is concave and monotone on the boundary, we can safely mark as irrelevant any variable that is zeroed out by the optimization algorithm. In other words, it is *faithful* in terms of variable selection under an additive approximation. Before giving our main result, which makes this precise, we begin with a lemma that characterizes the components of the optimal additive approximation.

For notational simplicity, we suppose that $|\mathcal{A}_n| = m$ for all n. We assume that each purchase, which comprises $\left(\{\mathbf{x}_i\}_{i=1}^{m}, \mathbf{s}_n\right)$, is iid drawn from some distribution F with density p. For this section, we use k to index features, i, j to index items, and n to index consumers.

Lemma 3.3.1. *Let F be a distribution on $[0,1]^{md}$ with a positive density p. Let $f : [0,1]^d \to \mathbb{R}$ be an integrable true function of the items. Define the following for*

any fixed h:

$$\{f_k^*\}_{k=1}^d = \arg\min_{\{f_k\}_{k=1}^{d_1}} \mathbb{E}\left[-\sum_{i=1}^m Y_i V(\mathbf{x}_i, \mathbf{s}_n) + \log\left(\sum_{j=1}^m \exp\left(V(\mathbf{x}_j, \mathbf{s}_n)\right)\right)\right] \tag{3.3.1}$$

$$\textit{where} \qquad V(\mathbf{x}_i, \mathbf{s}_n) = \sum_{k=1}^{d_1} f_k(x_{ki}) + h(\mathbf{s}_n)$$

$$Y_i \,|\, \mathbf{x}, \mathbf{s}_n \sim \text{Bernoulli}\left(\frac{\exp(f(\mathbf{x}_i) + h(\mathbf{s}_n))}{\sum_{j=1}^m \exp(f(\mathbf{x}_j) + h(\mathbf{s}_n))}\right).$$

Then f_k^* satisfies

$$\mathbb{E}\left[\frac{\exp(f_k^*(x_{ki}) + \phi(\mathbf{x}_{-k,i}, \mathbf{s}_n))}{\sum_{j=1}^m \exp(f_k^*(x_{kj}) + \phi(\mathbf{x}_{-k,j}, \mathbf{s}_n))} - \frac{\exp\left(f(x_{ki}, \mathbf{x}_{-k,i}) + h(\mathbf{s}_n)\right)}{\sum_{j=1}^m \exp\left(f(x_{kj}, \mathbf{x}_{-k,j}) + h(\mathbf{s}_n)\right)}\middle| x_{ki}\right] = 0, \tag{3.3.2}$$

where $\phi(\mathbf{x}_{-k,i}, \mathbf{s}_n) = \sum_{k' \neq k} f_{k'}(x_{k'i}) + h(\mathbf{s}_n)$. Furthermore, this solution is unique.

This lemma follows from the fact that

$$\mathbb{E}\left[-\sum_{i=1}^m Y_i V(\mathbf{x}_i, \mathbf{s}_n) + \log\left(\sum_{j=1}^m \exp\left(V(\mathbf{x}_j, \mathbf{s}_n)\right)\right)\right] =$$
$$\int -\sum_{i=1}^m \left(\sum_{k=1}^d f_k(x_{ik})\right) \frac{\exp(f(\mathbf{x}_i))}{\sum_{i=1}^m \exp(f(\mathbf{x}_i))} p(\{\mathbf{x}_i\}_{i=1}^m) d\{\mathbf{x}_i\}_{i=1}^m$$
$$+ \int \log \sum_{i=1}^m \exp\left(\sum_{k=1}^d f_k(x_{ik})\right) p(\{\mathbf{x}_i\}_{i=1}^m) d\{\mathbf{x}_i\}_{i=1}^m$$

and a standard KKT argument.

Lemma 3.3.1 states the intuitive fact that the first moment conditional on x_{ki} under the true model must equal that of the optimally fitted f_k^*. We now give our main result and the accompanying assumptions.

Definition 1. Let $f : [0,1]^d \to \mathbb{R}$ be an integrable function, f is boundary monotone if for all k and all $\mathbf{x}_{-k}$

$$\partial_{x_k} f \geq 0 \quad \text{or} \quad \partial_{x_k} f \leq 0 \qquad \text{at the boundary } x_k = 0 \text{ and } x_k = 1$$

Definition 2. Let $p(\mathbf{x})$ be a density supported on $[0,1]^{md}$, p satisfies the boundary-points condition, if

$$\frac{\partial}{\partial x_{ki}} p(\mathbf{x}_{-k,i}, \{\mathbf{x}_l\}_{l \neq i} \,|\, x_{j,i}) = 0 \quad \text{at } x_{ji} = 0 \text{ and } x_{ji} = 1, \quad \text{for any } \mathbf{x}_{-j,i}, \{\mathbf{x}_l\}_{l \neq i}$$

Theorem 3.3.1. *(Additive Faithfulness) Let p be a positive mixed density supported on $[0,1]^{md}$ that satisfies the boundary-points property (definition 2). Suppose f is concave, boundary-monotone (definition 1) and differentiable.*

Fix arbitrary h, let $\{f_k^*\}_{k=1}^{d_1}$ be the optimal additive components as defined in equation 3.3.1. Then $f_k^* = 0$ implies that $\partial_{x_k} f(\mathbf{x}) = 0$, that is, f does not depend on feature k.

Theorem 3.3.1 is the main theoretical result of this paper. It states that even if the true function f is not additive, the additive approximation yields no false negatives. We defer the proof to section 3.6 of the appendix.

It is important to note that additive faithfulness does not rely on any restrictions of the correlation structure between the covariates. The only distributional assumption we make is the mild boundary-point condition (definition 2). We allow the density to behave arbitrarily in the interior of the support. In contrast, in linear regression where $\beta^* = \Sigma^{-1}\mathbb{E}[Xf(X)]$, we would need to restrict the covariance to make the same faithfulness guarantee.

The boundary monotone condition (definition 1) is reasonable in applications where the concavity assumption is natural. With respect to some features, such as the per-pupil school spending in Nechyba and Strauss Nechyba and Strauss (1998), the utility function is monotone and thus boundary monotone as well. Boundary monotone condition also holds for features of which people want more when there is too little (one boundary point) and less when there is too much (the other boundary point). For instance, people distrust extremely cheap items and refrain from extremely expensive items.

Theorem 3.3.1 does not give a way to estimate whether $f_k^* = 0$. The next section tackles this problem.

Concave Additive Model

Since the true function f is concave, it is natural to consider a concave additive model. For notational simplicity, we omit $h(\mathbf{s}_n)$ in this section.

$$\{\widetilde{f}_k^*\}_{k=1}^d = \arg\min_{\widetilde{f}_k \in -\mathcal{C}^1} \mathbb{E}\left[-\sum_{i=1}^m Y_i \left(\sum_{k=1}^d \widetilde{f}_k(x_{ki})\right) + \log \sum_{j=1}^m \exp\left(\sum_{k=1}^d \widetilde{f}_k(x_{kj})\right)\right]$$

where we use $\mathcal{C}^1, -\mathcal{C}^1$ to denote the set of univariate convex and concave functions respectively.

Concave additive components $\widetilde{f}_k^*$ are not additively faithful, but we can restore faithfulness by coupling the $\widetilde{f}_k^*$'s with a set of convex functions:

$$g_k^* = \arg\min_{g_k \in \mathcal{C}^1} \mathbb{E}\left[-\sum_{i=1}^m Y_i \left(g_k(x_{ki}) + \phi(\mathbf{x}_{-k,i}) \right) + \log\left(\sum_{j=1}^m \exp\left(g_k(x_{kj}) + \phi(\mathbf{x}_{-k,j}) \right) \right) \right]$$

where $\phi(\mathbf{x}_{-k,i}, \mathbf{z}_i) = \sum_{k' \neq k} \widetilde{f}_{k'}^*(x_{k'i})$.

Theorem 3.3.2. *Suppose $p(\mathbf{x})$ is a mixed positive density on C, where $C \subset \mathbb{R}^{md}$ is a compact set and $p(\mathbf{x})$ satisfies the boundary-points condition. Suppose that $\partial_{x_{ki}} f(\mathbf{x}_i)$, $\partial^2_{x_{ki}} f(\mathbf{x}_i)$, $\partial_{x_{ki}} p(\mathbf{x}_{-k,i}, \{\mathbf{x}_j\}_{j \neq i} \,|\, x_{ki})$, and $\partial^2_{x_{ki}} p(\mathbf{x}_{-k,i}, \{\mathbf{x}_j\}_{j \neq i} \,|\, x_{ki})$ are all continuous on C. Then $\widetilde{f}_k^* = 0$ and $g_k^* = 0$ only if f does not depend on x_k, i.e. $\partial_{x_k} f(\mathbf{x}) = 0$ with probability 1.*

We defer the proof of theorem 3.3.2 to section 3.6 of the appendix. Theorem 3.3.2 states that if a covariate is relevant, then at least one of the optimal concave and convex functions that minimizes the negative likelihood should be nonzero. Therefore, if we fit 0 for both the convex and concave component, we can safely zero out the corresponding variable and claim it as irrelevant. Intuitively, the convex $\widehat{g}_k^*$ "catches" any non-concave residual that $\widetilde{f}_k^*$ could not capture.

3.4 Estimation Procedure

Theorem 3.3.2 motivates a two stage procedure for variable selection. In the first stage, we fit a sparse additive concave function under the logistic DCM framework. We then separately fit a convex function on the residuals for each dimension.

Importantly, we do not introduce tuning parameters for smoothing the function. Such smoothing parameters are essential to most nonparametric estimation methods, but are typically very difficult to set. In particular, there is no easy way to optimally adjust smoothing parameters in a traditional additive model, based for example on kernel regression or smoothing splines. This is a key attraction of the shape-constrained approach.

Given sample $\{\mathbf{x}_{ni}, \mathbf{s}_n, \mathbf{y}_n\}_{i \in \mathcal{A}_n}^{n=1,\ldots N}$, the following procedure, referred to as AC/DC(additively concave/decoupled convex), is performed.

AC Stage: Compute, jointly

$$\widehat{f}_1, \ldots, \widehat{f}_{d_1}, \widehat{\gamma}, \widehat{h} = \underset{f_1,\ldots,f_{d_1} \in -\mathcal{C}^1, \gamma, h}{\arg\min} -\frac{1}{N} \sum_{n=1}^N \ell(\mathbf{y}_n \,|\, \mathbf{X}_{\mathcal{A}_n}, \mathbf{s}_n) + \lambda \sum_{k=1}^p \|\beta_{k\cdot}\|_\infty \qquad (3.4.1)$$

$$\text{where} \qquad \ell(\mathbf{y}_n|\mathbf{X}_{\mathcal{A}_n}, \mathbf{s}_n) = \sum_{i\in\mathcal{A}_n} y_{ni}\widehat{V}(\mathbf{x}_i, \mathbf{s}_n) - \log(\sum_{j\in\mathcal{A}_n} \exp(\widehat{V}(\mathbf{x}_j, \mathbf{s}_n))$$

$$\widehat{V}(\mathbf{x}_i, \mathbf{s}_n) = \sum_k f_k(x_{ki}) + h(\mathbf{s}_n)$$

and $\beta_{k\cdot}$ are the corresponding subgradients of $f_k(\cdot)$.
DC Stage: Compute, separately, for each k where $||\beta_{k\cdot}||_\infty = 0$

$$\widehat{g}_k = \underset{g_k\in\mathcal{C}^1}{\arg\min} -\frac{1}{N}\sum_{n=1}^{N} \ell(\mathbf{y}_n \,|\, \mathbf{X}_{\mathcal{A}_n}, \mathbf{s}_n) + \lambda\|\widetilde{\beta}_{k\cdot}\|_\infty \tag{3.4.2}$$

where $\widetilde{\beta}_{k\cdot}$ are the corresponding subgradients of $g_k(\cdot)$. We then output as the set of continous relevant variables $\{k : \|\beta_{k\cdot}\|_\infty > 0 \text{ or } \|\widetilde{\beta}_{k\cdot}\|_\infty > 0\}$ and of discrete relevant variables $\{k' : \gamma_{k'} \neq 0\}$

We adopted an ℓ_∞/ℓ_1 penalty in 3.4.1 and the ℓ_∞ penalty in 3.4.2 to encourage sparsity. In the AC stage (3.4.1), any estimation method for h can be used.

Optimization

We describe the optimization algorithm only for the additive concave logistic regression stage, the second decoupled convex logistic regression stage is a straightforward modification. We observe that a univariate concave function is characterized by non-increasing subgradients. So we form our optimization problem as

$$\begin{aligned} \underset{\mathbf{f},\beta,\gamma,\mathbf{h}}{\text{minimize}} \quad & -\sum_{n=1}^{N} \ell(\mathbf{y}_n \,|\, \mathbf{X}_{\mathcal{A}_n}, \mathbf{s}_n) + \lambda_1 \sum_{k=1}^{d_1} ||\beta_{k\cdot}||_\infty \\ \text{subject to} \quad & f_{k(i+1)} = f_{k(i)} + \beta_{k(i)}(x_{k(i+1)} - x_{k(i)}) \\ & \sum_{i=1}^{M} f_{ki} = 0, \qquad \beta_{k(i+1)} \leq \beta_{k(i)}, (\forall k, i) \end{aligned} \tag{3.4.3}$$

$$\ell(\mathbf{y}_n \,|\, \mathbf{X}_{\mathcal{A}_n}, \mathbf{s}_n) \equiv \sum_{i\in\mathcal{A}_n} y_{ni}(\sum_{k=1}^{d_1} f_k(x_{ki}) - \log(\sum_{j\in\mathcal{A}_n} \exp(\sum_{k=1}^{d_1} f_k(x_{kj}) + h_n))$$

where $\{(1), (2), \ldots, (M)\}$ is a reordering of $\{1, 2, \ldots, M\}$ such that $x_{k(1)} \leq x_{k(2)} \leq \ldots \leq x_{k(n)}$. We use the centering constraints $\sum_{i=1}^{M} f_k(x_{ki}) = 0$ for identifiability.

Motivated by the shooting algorithm for the lasso Friedman et al. (2010), we solve

optimization 3.4.3 with block coordinate descent. When estimating $\mathbf{f}$ (or γ), we iteratively select a dimension k, fix all $\mathbf{f}_{k'}$ (or $\gamma_{k'}$) for $k' \neq k$, and optimize $\{f_k(x_{ki})\}_{i=1,\ldots,M}$ (or γ_k). For each iteration, we apply Newton's method and solve a sequence of quadratic programs. We use the optimization software MOSEK to solve the intermediate QPs in our implementation. In the cases where $h_n = h(\mathbf{s}_n)$ must be estimated as well, we would iterate between $(\mathbf{f}, \beta), \mathbf{h}$ in an outer loop and, depending on choice of model for $h(\mathbf{s}_n)$, any appropriate optimization algorithm can be used to optimize $\mathbf{h}$ in a step of the outer loop.

The estimated function can be evaluated on an input item $\mathbf{x}_j$ with the equation $f(\mathbf{x}_j) = \sum_{k=1}^{d_1} f_k(x_{kj}) = \sum_{k=1}^{d_1} \min_i\{f_{ki} + \beta_{ki}(x_{kj} - x_{ki})\}$ For univariate convex function estimation, we modify the linear inequality so that the subgradients are non-decreasing: $\beta_{k(i+1)} \geq \beta_{k(i)}$.

3.5 Experiment

We evaluate AC/DC on both synthetic data experiments as well as a novel survey dataset. For all of our experiments we do not consider consumer features, i.e., we omit the $h(\mathbf{s}_n)$ term.

Simulation

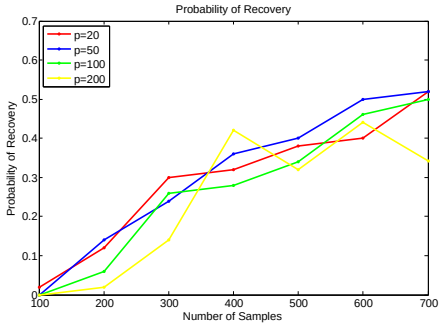
For the M items, we generate continuous feature vectors $\mathbf{x}_1, ..., \mathbf{x}_M \sim P$ where the distribution $P = c\bar{N}(0, \Sigma) + (1 - c)U$ is a mixture between a multivariate Gaussian distribution thresholded to lie in $[-b, b]^{p_1}$ and an uniform distribution supported on $[-b', b']^{p_1}$ where $b' > b$ to fulfill the boundary condition. By "thresholded", we mean that if the Gaussian sample is greater than b, then we set it equal to b. The discrete feature vectors $\mathbf{z}_1, ..., \mathbf{z}_M$ are generated similarly, except that we discretize the vectors by setting a continuous value to zero if it is less than 0.

The true utility function is taken to be the sum of a piecewise linear function of the continuous features $\mathbf{x}_{S_1}$ and a linear function of the discrete features $\mathbf{z}_{S_2}$, where the piecewise linear function is guaranteed to be concave and S_1, S_2 represent the corresponding active feature sets with $|S_1| = |S_2| = 3$. In the simulations, we take $\Sigma_{ij} = \nu^{|i-j|}$ for various ν's, and pick the set of active features at random to create varying amounts of correlation between relevant and irrelevant variables. In addition, we always set $\lambda_1 = \sqrt{\frac{\log(Np_1)}{N}}$ and $\lambda_2 = 0.3\sqrt{\frac{\log(Np_1)}{N}}$.

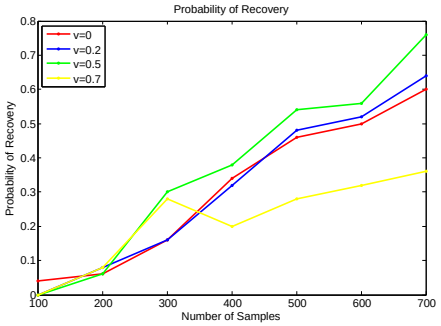
In the first simulation, we fix $\nu = 0.3$. We vary $N = 100, 200, \ldots, 700$, $p_1 = 20, 50, 100, 200$. For each (N, p_1), we generate 50 independent datasets and apply

AC/DC procedure to infer the function estimates $\mathbf{f}$,subgradients β, and discrete coefficients γ. We declare correct support recovery, if for $\forall k \in S_1$, $\|\beta_{k\cdot}\|_\infty > 10^{-6}$, $\forall k \notin S_1$, $\|\beta_{k\cdot}\|_\infty < 10^{-6}$ and for $\forall k' \in S_2$,$|\gamma_{k'}| > 10^{-6}$, $\forall k' \notin S_2$,$|\gamma_{k'}| < 10^{-6}$. We show the plot of correct support recovery probability versus differnt combinations of N and p_1 in figure 3.1(a). As can be seen, the ACDC algorithm achieves higher support recovery rate as sample size increases even when p is large.

In the second simulation, we fix $p_1 = 15$ and investigate the robustness of the ACDC algorithm over different correlation structures. N varies from $100, 200, \ldots, 700$ and ν varies from $0, 0.2, 0.5, 0.7$. As before, we generate 50 data sets and compute the probability of correct support recovery for each combination of N and ν. The results are shown in figure 3.1(b) and demonstrate that ACDC can still select relevant variables well for design of moderate correlation.



(a) Support recovery results for various p

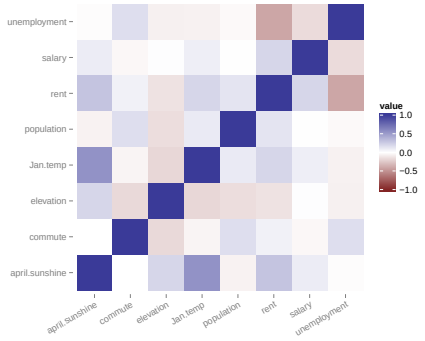


(b) Support recovery results for various ν

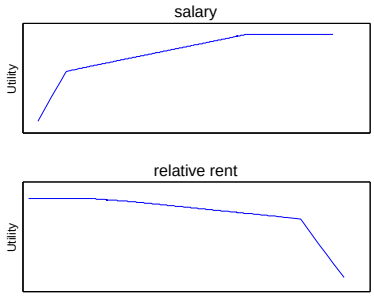
Survey Data

This dataset consists of 530 surveys we gave to the students and staff at our university. Each survey contains three options, each of which is a hypothetical living arrangement that consists of a job and an apartment in some neighborhood within some city. We ask the respondents to choose the one they most prefer. Each living arrangement is described on the survey by three types of information: personal level, city level, and neighborhood level. The personal level information are *yearly salary, monthly rent, and commute time.* The city level information are *year round temperature, population, robbery rate, and diabetes rate.* The neighborhood level information are *average income, unemployment rate, and the percent of college graduates.*

We created the surveys by gathering information on 68 US cities and a total of 148 zipcode regions (which we call neighborhoods) within those cities. The informa-



(a) Correlation among some features shown and not shown on the survey.



(b) Two examples of the fitted component functions. Relative rent is rent divided by average income of the neighborhood.

	ACDC	linear
Features in Survey	**83.3%**	76.9%
Features in Survey (minus *% diabetes*)	**80.8%**	75.7 %

(a) Percentage of features selected that are among the features given on the surveys.

	concave	linear
8 features	**0.670**	0.680
3 features	**0.680**	0.686

(b) Negative log-likelihood of models fitted using either the top 8 features or the top 3 features of the feature selection process.

Figure 3.1: Variable selection accuracy on survey data.

tion is gathered from `www.city-data.com`. We generate each living arrangement by randomly selecting a city and a zipcode region and then generating a random salary, rent, and commute time based on the average in that zipcode neighborhood. The reader can find examples of the survey as well as more detail about how we made the surveys in section 3.7.

Feature selection evaluation. In addition to the features shown on the survey, we collected various other features of the cities and zipcode regions we used. These additional features are *July humidity level, January snowfall, April sunshine rate, % households gay/lesbian, % households unmarried, elevation, air quality index, and % voted Obama in 2008*. Because these features were not shown on the survey and not known to the respondents, they are by construction irrelevant to the survey responses. These irrelevant features are, however, correlated with the survey features (Figure 3.1(a)).

We evaluate AC/DC by taking random subsamples of the data, performing variable selection, and measuring how often the features shown on the survey are marked as relevant. We compare AC/DC against the sparse standard logistic DCM where we let $f(\mathbf{x}_i) = \beta^T \mathbf{x}_i$ and apply the ℓ_1 penalty on β. The regularization parameters are selected so that on average 9.5 features are selected. From the 400 surveys in the training data, we took 94 random subsamples of 200 surveys and performed both AC/DC and sparse linear model on these subsamples. In addition to the raw features, we also added some interaction terms among the relevant variables.

The results are shown in Figure 3.2 and Table 3.1(a). AC/DC outperforms the linear model in choosing features that are relevant to the survey. We included *%diabetes* as a feature on the survey though it is unlikely to play a part in a respondent's decision process. Thus, in the second row of Table 3.1(a), we exclude *%diabetes* as a relevant variable. Not surprisingly, the two features selected with 100% frequency are salary and a rent-times-commute interaction term.

Heldout likelihood evaluation. To ensure that the concavity assumption is reasonable and that we are not overfitting to the training data, we also evaluate the log-likelihood of our estimated model on a heldout dataset of 114 surveys. These surveys use information only from cities that *do not appear in any of the training data surveys*. We use the top 3 or the top 8 features in selection process and refit an additive concave model, unregularized, on the training data, using only those features (likewise with the sparse linear model). For features whose monotoncity in the utility is obvious, we also add a monotone constraint when refitting. Table 3.1(b) shows that concave monotone model performs slightly better. Though the improvement is small, the concave monotone model using 3 features achieves the same likelihood as the linear model using 8 features. We show two examples of the fitted functions

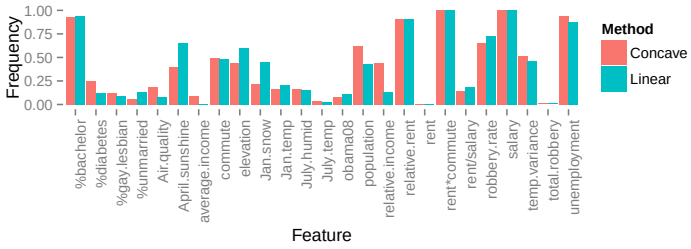


Figure 3.2: Variable selection frequency for survey data

in Figure 3.1(b). Both salary and the relative rent (rent / average income) exhibit concavity.

3.6 Proofs

Proof of Theorem 3.3.1

For simplicity, we omit the consumer function $h(\mathbf{s}_n)$ in the proofs.

Proof. (of Theorem 3.3.1) We want to show that $f_k^* = 0 \Rightarrow \frac{\partial}{\partial x_k} f = 0$.

Now we assume that for all x_{ki}, (3.3.2) holds:

$$\mathbb{E}\left[\frac{\exp(f_k^*(x_{ki}) + \phi(\mathbf{x}_{-k,i}))}{\sum_{i=1}^m \exp(f_k^*(x_{ki}) + \phi(\mathbf{x}_{-k,i}))} - \frac{\exp(f(x_{ki}, \mathbf{x}_{-k,i}))}{\sum_{i=1}^m \exp(f(x_{ki}, \mathbf{x}_{-k,i}))} \,|\, x_{ki}\right] = 0$$

Note that $f_k^*(x_{ki}) = f_k^*(x_{kj})$ for all i, j. Differentiating with respect to x_{ki} under the integral gives:

$$\int p'(\mathbf{x}_{-k,i}, \{\mathbf{x}_j\}_{j\neq i} \,|\, x_{ki}) \left[\frac{\exp(\phi(\mathbf{x}_{-k,i}))}{\sum_{j=1}^m \exp(\phi(\mathbf{x}_{-k,j}))} - \frac{\exp\left(f(x_{ki}, \mathbf{x}_{-k,i})\right)}{\sum_{j=1}^m \exp\left(f(\mathbf{x}_j)\right)}\right]$$
$$+ p(\mathbf{x}_{-k,i}, \{\mathbf{x}_j\}_{j\neq i} \,|\, x_{ki}) \frac{\exp(f(x_{ki}, \mathbf{x}_{-k,i})) f'(x_{ki}, \mathbf{x}_{-k,i}) \sum_{j\neq i} \exp\left(f(x_{kj}, \mathbf{x}_{-k,j})\right)}{\left(\sum_{j=1}^m \exp\left(f(x_{kj}, \mathbf{x}_{-k,j})\right)\right)^2} d\mathbf{x}_{-(k,i)} = 0$$

We use the shorthand $d\mathbf{x}_{-(k,i)}$ to represent $\prod_{k',j' \,:\, (k,j)\neq(k,i)} d\mathbf{x}_{k'j'}$. That is, we integrate with respect to all variables except x_{ki}.

If p satisfies the boundary-points condition, then, at $x_{ki} = 0/1$, the integral equation reduces to:

$$\int p(\mathbf{x}_{-k,i}, \{\mathbf{x}_j\}_{j\neq i} \,|\, x_{ki}) \frac{\exp(f(x_{ki}, \mathbf{x}_{-k,i})) f'(x_{ki}, \mathbf{x}_{-k,i}) \sum_{j\neq i} \exp(f(x_{kj}, \mathbf{x}_{-k,j}))}{\left(\sum_{j=1}^m \exp(f(x_{kj}, \mathbf{x}_{-k,j}))\right)^2} d\mathbf{x}_{-(k,i)} = 0$$

Recall that f is boundary-monotone, so without loss of generality we can assume that $f'(x_{ki}, \mathbf{x}_{-k,i}) \geq 0$ for $x_{ij} = 0/1$. Also, since we assumed that the density p is positive, $p(\mathbf{x}_{-k,i}\{\mathbf{x}_j\}_{j\neq i} \,|\, x_{ki}) > 0$. So we have $f'(x_{ki}, \mathbf{x}_{-k,i}) = 0$ at $x_{ki} = 0/1$ for all $\mathbf{x}_{-k,i}$.

Because $f(x_{ki}, \mathbf{x}_{-k,i})$ as a function of x_{ki} is concave, it must be that, for all $x_{ki} \in (0, 1)$ and for all $\mathbf{x}_{-k,i}$:

$$0 = f'(1, \mathbf{x}_{-k,i}) \leq f'(x_{ki}, \mathbf{x}_{-k,i}) \leq f'(0, \mathbf{x}_{-k,i}) = 0$$

Therefore, f does not depend on x_k. □

Proof of Theorem 3.3.2

Proof. From Theorem 3.3.1, it suffices to show that $f_k^* = 0$.
Now suppose $\widetilde{f}_k^{*'} = g_k^{*'} = 0$. First consider the univariate function $h_k(x_{ki}) = \delta e^{-x_{ki}}$, where $\delta \in \mathbb{R}$. $h_k(x_{ki})$ is convex and decreasing if $\delta > 0$, concave and increasing if $\delta < 0$. Since $\widetilde{f}_k^* = g_k^* = 0$, then

$$\begin{aligned}
&\arg\min_{\delta\in\mathbb{R}} \left\{ \mathbb{E}\left[-\sum_{i=1}^m Y_i \left(\delta e^{-x_k} + \phi(\mathbf{x}_{-k,i})\right) + \log\left(\sum_{i=1}^m \exp(\delta e^{-x_{ki}} + \phi(\mathbf{x}_{-k,i}))\right)\right]\right\} \\
&= \arg\min_{\delta\in\mathbb{R}} \left\{ \mathbb{E}\left[\log\left(\sum_{i=1}^m \exp(\delta e^{-x_{ki}} + \phi(\mathbf{x}_{-k,i}))\right) - \sum_{i=1}^m p_i(\delta e^{-x_{ki}} + \phi(\mathbf{x}_{-k,i}))\right]\right\} \\
&= 0
\end{aligned}$$

where

$$p_i = \frac{\exp(f(\mathbf{x}_i))}{\sum_{j=1}^m \exp(f(\mathbf{x}_j))}$$

Recall that the objective function is convex in δ, the stationary condition gives us:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^m e^{-x_{ki}} \left(\frac{\exp(\delta e^{-x_{ki}} + \phi(\mathbf{x}_{-k,i}))}{\sum_{j=1}^m \exp(\delta e^{-x_{kj}} + \phi(\mathbf{x}_{-k,i}))} - \frac{\exp(f(\mathbf{x}_i))}{\sum_{j=1}^m \exp(f(\mathbf{x}_j))}\right)\right] = 0 \\
&\overset{\delta^*=0}{\Longrightarrow} \mathbb{E}\left[\sum_{i=1}^m e^{-x_{ki}} \left(\frac{\exp(\phi(\mathbf{x}_{-k,i}))}{\sum_{j=1}^m \exp(\phi(\mathbf{x}_{-k,i}))} - \frac{\exp(f(\mathbf{x}_i))}{\sum_{j=1}^m \exp(f(\mathbf{x}_j))}\right)\right] = 0
\end{aligned}$$

It is not hard to prove that $f_k^*(x_{ki})$ has lower bounded derivatives $f_k^{*'}(x_{ki})$ and $f_k^{*''}(x_{ki})$. Then we can always find an η such that $e^{-x_{ki}} + \eta f_k^*(x_{ki})$ is convex and non-increasing. Therefore, by a similar argument, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^m (e^{-x_{ki}} + \eta f_k^*(x_{ki})) \left(\frac{\exp(\phi(\mathbf{x}_{-k,i}))}{\sum_{j=1}^m \exp(\phi(\mathbf{x}_{-k,j}))} - \frac{\exp\left(f(\mathbf{x}_i)\right))}{\sum_{j=1}^m \exp(f(\mathbf{x}_j))}\right)\right] = 0 \\
&\Longrightarrow \mathbb{E}\left[\sum_{i=1}^m f_k^*(x_{ki}) \left(\frac{\exp(\phi(\mathbf{x}_{-k,i}))}{\sum_{j=1}^m \exp(\phi(\mathbf{x}_{-k,j}))} - \frac{\exp\left(f(\mathbf{x}_i)\right))}{\sum_{j=1}^m \exp(f(\mathbf{x}_j))}\right)\right] = 0 \\
&\Longrightarrow \mathbb{E}\left[\sum_{i=1}^m f_k^*(x_{ki}) \mathbb{E}\left[\left(\frac{\exp(\phi(\mathbf{x}_{-k,i}))}{\sum_{j=1}^m \exp(\phi(\mathbf{x}_{-k,j}))} - \frac{\exp\left(f(\mathbf{x}_i)\right))}{\sum_{j=1}^m \exp(f(\mathbf{x}_j))}\right) \middle| x_{ki}\right]\right] = 0
\end{aligned}$$

Recall that $f_k^*(x_{ki})$ is a unique function that satisfies

$$\mathbb{E}\left[\frac{\exp(f_k^*(x_{ki}) + \phi(\mathbf{x}_{-k,i}))}{\sum_{j=1}^m \exp(f_k^*(x_{kj}) + \phi(\mathbf{x}_{-k,j}))} - \frac{\exp\left(f(\mathbf{x}_i)\right))}{\sum_{j=1}^m \exp(f(\mathbf{x}_j))} \middle| x_{ki}\right] = 0.$$

Then we have

$$\mathbb{E}\left[\sum_{i=1}^{m} f_k^*(x_{ki})\mathbb{E}\left[\left(\frac{\exp(\phi(\mathbf{x}_{-k,i}))}{\sum_{j=1}^m \exp(\phi(\mathbf{x}_{-k,j}))} - \frac{\exp(f_k^*(x_{ki})+\phi(\mathbf{x}_{-k,i}))}{\sum_{j=1}^m \exp(f_k^*(x_{kj})+\phi(\mathbf{x}_{-k,j}))}\right)\middle| x_{ki}\right]\right] = 0$$

$$\Longrightarrow \mathbb{E}\left[\sum_{i=1}^{m} f_k^*(x_{ki})\left(\frac{\exp(\phi(\mathbf{x}_{-k,i}))}{\sum_{j=1}^m \exp(\phi(\mathbf{x}_{-k,j}))} - \frac{\exp(f_k^*(x_{ki})+\phi(\mathbf{x}_{-k,i}))}{\sum_{j=1}^m \exp(f_k^*(x_{kj})+\phi(\mathbf{x}_{-k,j}))}\right)\right] = 0$$

$$\Longrightarrow \mathbb{E}\left[\sum_{i=1}^{m} \frac{f_k^*(x_{ki})\exp(\phi(\mathbf{x}_{-k,i}))\sum_{j\neq i}\exp(\phi(\mathbf{x}_{-k,j}))\left(\exp(f_k^*(x_{kj})-\exp(f_k^*(x_{ki}))\right)}{\sum_{j=1}^m \exp(\phi(\mathbf{x}_{-k,j}))\sum_{j=1}^m \exp(f_k^*(x_{kj})+\phi(\mathbf{x}_{-k,j}))}\right] = 0$$

Note that

$$\begin{aligned}
&\sum_{i=1}^{m} f_k^*(x_{ki})\exp(\phi(\mathbf{x}_{-k,i}))\sum_{j\neq i}\exp(\phi(\mathbf{x}_{-k,j}))\left(\exp(f_k^*(x_{kj})-\exp(f_k^*(x_{ki}))\right)\\
&=\sum_{i=1}^{m}\sum_{j\neq i} f_k^*(x_{ki})\exp(\phi(\mathbf{x}_{-k,i})+\phi(\mathbf{x}_{-k,j}))\left(\exp(f_k^*(x_{kj})-\exp(f_k^*(x_{ki}))\right)\\
&=\sum_{i<j}\exp(\phi(\mathbf{x}_{-k,i})+\phi(\mathbf{x}_{-k,j}))\Big(f_k^*(x_{ki})-f_k^*(x_{kj})\Big)\Big(\exp(f_k^*(x_{kj})-\exp(f_k^*(x_{ki}))\Big)\\
&\leq 0 \quad \text{since } \Big(f_k^*(x_{ki})-f_k^*(x_{kj})\Big)\Big(\exp(f_k^*(x_{kj})-\exp(f_k^*(x_{ki})\Big)\leq 0
\end{aligned}$$

and

$$\sum_{j=1}^{m}\exp(\phi(\mathbf{x}_{-k,j}))\sum_{j=1}^{m}\exp(f_k^*(x_{kj})+\phi(\mathbf{x}_{-k,j}))>0.$$

Thus we have

$$\Big((f_k^*(x_{ki})-f_k^*(x_{kj})\Big)\Big(\exp(f_k^*(x_{kj}))-\exp(f_k^*(x_{ki}))\Big)=0, \text{for all } i\neq j$$

i.e. $f_k^*(x_{ki})=f_k^*(x_{kj}), \text{for all } i\neq j.$ □

```
 [1] bridgeport   orlando      columbus     jacksonville dallas
 [6] charlotte    reno         portland     durham       denver
[11] jersey_city  paradise     spokane      rockford     chesapeake
[16] chicago      cambridge    austin       seattle      raleigh
[21] allentown    berkeley     philadelphia pittsburgh   boston
[26] san_diego    las_vegas    lynn         atlanta      richmond
[31] cincinnati   warren       madison      houston      san_antonio
[36] miami        fremont      nyc          albany       la
[41] newark       vancouver    sf           detroit      aurora
[46] stamford     ann_arbor    springfield  grand_rapids elizabeth
[51] eugene       milwaukee    cleveland    new_haven    dc
[56] boulder      henderson    buffalo
```

Figure 3.3: List of the cities used in the training dataset surveys.

3.7 Survey Detail

Figure 3.5 shows examples of surveys we handed out to respondents. Each survey contains three random living arrangements. The living arrangements are sampled randomly from a large collection that we generated from actual city and zipcode region data. To generate a living arrangement, we take the following steps:

1. We select a random city and a random zipcode region. All features of the living arrangement except *salary, commute time, and rent* are simply the features of the selected city and the zipcode region.

2. We generate a random salary $s = s_{base} \cdot c_{adj} \cdot c_{noise}$ where $s_{base} \sim Unif\{69K, 80K\}$. $c_{adj} = \left(\frac{\text{national average income}}{\text{region average income}}\right)^{0.15}$ is the regional wealth adjustment; richer regions yield higher salaries. $c_{noise} \sim N(0, 0.15^2)$ is a Gaussian multiplier noise.

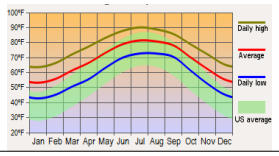
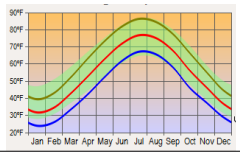
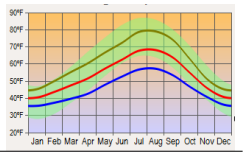
3. We generate a random rent $r = r_{base} c_{noise}$ where r_{base} is the average rent of the zipcode region and $c_{noise} \sim N(0, 0.15^2)$ is a Gaussian multiplier noise.

```
[1] kansas       minneapolis  baltimore    phoenix      fort_wayne
[6] indianapolis manchester   st_louis     st_paul      norfolk
```

Figure 3.4: List of the cities used in the test dataset surveys.

- **Gender**: (circle one): M F Other N/A
- **Major**: Social Science Physical Science Math/Stats Language Biology Economics Other Undecided

You have just completed your Bachelor's degree. You have a choice of three different jobs in three different locations in the US. All other considerations being equal, which one of the following living situations do you most prefer?

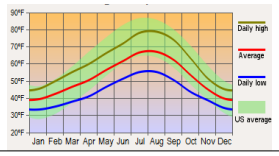
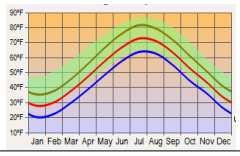
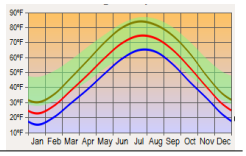
	Situation A	Situation B	Situation C
Salary	69,000$	79,000$	88,000$
Rent	690$	710$	760$
Commute Time	28 min	24 min	26 min
Temperature			
City Population (Chicago: 3,000,000)	837,000	1,550,000	603,000
City Robbery Rate (per 100,000) (New York City: 250)	163	519	159
Diabetes Rate (national: 11%)	10.4%	10.4%	7.0%
Neighborhood Income (national: 44,000)	41,000$	23,000$	56,000$
Neighbor. %College Grad (national: 32%)	10.6%	11.5%	54.7%
Neighbor. %Unemploy (national: 8.1%)	11.2%	17.6%	10.3%
Response (check one)			

Survey ID: 1

1

- **Gender**: (circle one): M F Other N/A
- **Major**: Social Science Physical Science Math/Stats Language Biology Economics Other Undecided

You have just completed your Bachelor's degree. You have a choice of three different jobs in three different locations in the US. All other considerations being equal, which one of the following living situations do you most prefer?

	Situation A	Situation B	Situation C
Salary	84,000$	76,000$	64,000$
Rent	1,100$	1,500$	870$
Commute Time	12 min	34 min	26 min
Temperature			
City Population (Chicago: 3,000,000)	165,000	636,000	2,710,000
City Robbery Rate (per 100,000) (New York City: 250)	103	303	498
Diabetes Rate (national: 11%)	8.2%	8.0%	8.4%
Neighborhood Income (national: 44,000)	57,000$	61,000$	54,000$
Neighbor. %College Grad (national: 32%)	22.1%	49.7%	36.1%
Neighbor. %Unemploy (national: 8.1%)	10.2%	6.5%	7.5%
Response (check one)			

Survey ID: 2

1

Figure 3.5: Example surveys

CHAPTER 4

SHAPE-CONSTRAINT PATTERN SELECTION

4.1 Introduction

A major advantage of shape-constrained estimation is that there are no smoothing parameters. But, one does have to choose between two possible orientations. For instance, either an increasing function could be used or a decreasing one, either a convex function could be used or a concave one. The appropriate choice of orientation is clear in some applications but in general non-obvious. This choice becomes especially difficult in situations with model misspecification where the true underlying function may not be exactly increasing or decreasing but may be well approximated by either an increasing or a decreasing function.

In this section, we study the problem of finding the orientation pattern in shape-constrained estimation. We focus on regression with an additive model where the true condition mean $f_0(\mathbf{x}) = \mathbb{E}[y \,|\, X = \mathbf{x}]$ is modeled as a sum of p univariate functions $\sum_{j=1}^{p} f_j(x_j)$. The problem is to determine, for each j, whether f_j should be a monotone increasing function or a decreasing one. We consider monotoncity in our analysis but some of the ideas will apply to other shape-constraints as well.

There is an easy solution to this problem in the low-dimensional regime where p is small–try everything! There are 2^p possible patterns in total and one could try them all and select the pattern that provides the best fit. This brute force search strategy becomes inpractical once p becomes moderate (≥ 20).

We propose, in this chapter, a convex relaxation approach that scales to high dimensions. The idea is this: the brute search can be thought of an optimization with discrete variables where each component $f_j(x_j) = c_j f_j^{\text{incr}}(x_j) + (1 - c_j) f_j^{\text{decr}}(x_j)$ where $c_j \in \{0, 1\}$, f_j^{incr} is an increasing function and f_j^{decr} is a decreasing function. Our approach is motivated by the observation that we can relax c_j to be a continuous variable in $[0, 1]$ and add a regularization that encourages c_j to tend toward either 0 or 1. This approach is equivalent to additive trend filtering and also naturally arise out of our analysis of the identifiability of the additive shape-constraint model.

Related Work

Several works have studied additive shape-constrained estimation. Mammen and Yu (2007) propose an additive isotone model and show its estimation consistency. Chen and Samworth (2014) give an estimator for general shape-constraints and general linear model. Pya and Wood (2014) propose a smooth additive shape-constrained model based on constraining the basis coefficients of an additive B-spline model. Fang and Meinshausen (2012) study high-dimensional additive isotone regression and derive a backfitting optimization scheme where each iteration is a soft-thresholded PAVA. They also propose additive trend filtering but they do not analyze its performance in the pattern selection problem.

4.2 Setting

Definition 4.2.1. Let $f : \mathbb{R}^d \to \mathbb{R}$. We say that f is an *additive mixed monotone function* (Add-MM) if $f(\mathbf{x}) = \sum_{j=1}^d c_j f_j(x_j)$ where $c_j \in \{-1, +1\}$ and f_j is a monotone increasing function.

The vector $\mathbf{c} = (c_1, ..., c_d)$ represent the increasing/decreasing pattern. $c_j = 1$ implies that the j-th component is increasing and -1 if decreasing.

Let Y be a response. We want to model Y by the best additive mixed monotone function. In population setting, the problem is

$$\begin{aligned} &\min_{c_j, f_j} \mathbb{E}\left(Y - \sum_{j=1}^d c_j f_j(X_j)\right)^2 \\ &\text{s.t. } f_j \text{ monotone increasing, } \mathbb{E} f_j(X_j) = 0 \\ &\quad c_j \in \{-1, +1\} \end{aligned}$$

We have added a mean-zero constraint to eliminate an obvious source of non-identifiability: clearly, for any constant a, $f_1 + f_2 = (f_1 - a) + (f_2 + a)$. In finite sample, the optimization becomes

$$\min_{c_j, f_j} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{d} c_j f_j(X_{ij}) \right)^2 \tag{4.2.1}$$

$$\text{s.t. } f_j \text{ monotone increasing, } \sum_{i=1}^{n} f_j(X_{ij}) = 0 \tag{4.2.2}$$

$$c_j \in \{-1, +1\} \tag{4.2.3}$$

We can change the monotone increasing constraint to be the convexity constraint to get an *additive mixed convex/concave* model. We will focus on the monotone case for this section.

We can state optimization 4.2.1 in a more convenient way. We observe that f_j depends on the X_{ij}'s only through their ordering–any two vectors X_j, X'_j give rise to the same f_j so long as they have the same ordering. We will denote $f_{ij} = f_j(X_{\sigma(i)j})$ where $\sigma(i)$ is the index of the i-th smallest element of X_j.

Let $P_j \in \mathbb{R}^{n \times n}$ be a permutation matrix that corresponds to the σ^{-1} permutation so that $P_j(i, i') = 1$ iff $i = \sigma(i')$. Therefore, $(P_j f_j)_i = \sum_{i'=1}^{n} P_j(i, i') f_j(X_{\sigma(i')j}) = f_j(X_{ij})$. With this new notation, we can write

$$\min_{c_j, f_j} \|Y_i - \sum_{j=1}^{d} c_j P_j f_j\|_n^2 \tag{4.2.4}$$

$$\text{s.t. for all } j, \text{ for all } i = 1, ..., n-1, \ f_{ij} \leq f_{i+1,j} \tag{4.2.5}$$

$$\text{for all } j, \ \sum_{i=1}^{n} f_{ij} = 0 \tag{4.2.6}$$

$$c_j \in \{-1, +1\} \tag{4.2.7}$$

This optimization program is nonconvex because the variable c_j is discrete. The brute force approach is to try all 2^p combinations of patterns. Much effort will go toward an efficient method to reformulate and solve this problem. But first, it is worthwhile to understand the properties of this model more.

4.3 Identifiability

We want to know whether the model is identifiable and whether it is tractable. Both of these questions turned out to be difficult.

Identifiability is the following question: could there exist two solutions $\mathbf{c}, f$ and $\mathbf{c}', f'$ such that $\mathbf{c} \neq \mathbf{c}'$ but $\sum_{j=1}^{p} c_j P_j f_j = \sum_{j=1}^{p} c'_j P_j f'_j$. Another way of asking the same question is to say, suppose that the linear system

$$\sum_{j=1}^{p} c_j P_j f_j = y \tag{4.3.1}$$

$$\sum_{i=1}^{n} f_{ij} = 0 \quad \text{for all } j \tag{4.3.2}$$

is solvable for one set of $\mathbf{c}, f$, could there be another solution with a different pattern $\mathbf{c}$?

It is not hard to see that the model is non-identifiable. We will need to again set up some new notations before proceeding.

For any f_j satisfying the monotone increasing constraint, we can write $f_j = S\beta_j$ where β_j is a $(n-1)$-dimensional vector and $\beta_j \geq 0$ and S is an $n \times n-1$ matrix. Intuitively, β_j is the discrete first derivative and S is the zero-mean partial sum operator. $(Sv)_i = v_1 + v_2 + ... + v_i - \frac{1}{n}\sum_{i=1}^{n}\sum_{i'=1}^{i} v_{i'}$, where we subtract the mean in the last term.

As a side note, S is the pseudo-inverse to the first difference matrix D where D is $(n-1) \times n$ and $(Dv)_i = v_{i+1} - v_i$. If S does not have the centering adjustment, then $DS = I_{n-1}$ still, but SD is no longer symmetric.

In this notation, the linear system of equations 4.3.1 is equivalent to $\sum_{j=1}^{p} P_j S\beta_j = y$ where $\beta_j \geq 0$ if $c_j = 1$ and $\beta_j \leq 0$ if $c_j = -1$, which is in turn equivalent to

$$\sum_{j=1}^{p} P_j S\beta_j = y \quad \Leftrightarrow \quad \sum_{j=1}^{p} DP_j S\beta_j = Dy. \tag{4.3.3}$$

The new system has $n-1$ equations–one fewer than the original–because the f_j's and y must have mean zero and thus there is one redundant degree of freedom. With $n-1$ equations and $p(n-1)$ unknowns, the system is therefore hopelessly under-determined without the sign constraints imposed by $\mathbf{c}$. Even with the sign constraints, the system is still ill-defined and even worse, the system may admit multiple solutions with inconsistent sign patterns.

Example 4.3.1. Suppose $p = 2$. Let the rank of X_1 be $(1, 2, 3)$ and the rank of X_2 be $(3, 1, 2)$. We let $y = (-4, 0, 4)$. For reader's convenience, we represent f_{ij} by the notation f_i^j.

The linear system, in the form of Equation 4.3.1, is

$$\begin{aligned} f_1^1 + f_3^2 &= -4 \\ f_1^2 + f_1^2 &= 0 \\ f_1^3 + f_2^2 &= 4 \end{aligned}$$

We transform to the first difference variables β^1 and β^2 using the identity $f_1^1 = -\mu_1, f_2^1 = \beta_1^1 - \mu_1, f_3^1 = \beta_1^1 + \beta_2^1 - \mu_1$ and $\mu_1 = \frac{1}{3}(\beta_1^1 + (\beta_1^1 + \beta_2^1))$.

$$\begin{aligned} -\mu_1 + (\beta_1^2 + \beta_2^2 - \mu_2) &= -4 \\ \beta_1^1 - \mu_1 + (-\mu_2) &= 0 \\ \beta_1^1 + \beta_2^1 - \mu_1 + (\beta_1^2 - \mu_2) &= 4 \end{aligned}$$

The linear system, in the form of Equation 4.3.3, is

$$\begin{aligned} \beta_1^1 + (-\beta_1^2 - \beta_2^2) &= 4 \\ \beta_2^1 + \beta_1^2 &= 4 \end{aligned}$$

There are two sets of solutions to this system that have different sign patterns.

$$\begin{aligned} \beta^1 = (2, 5) \qquad \beta^1 = (6, 3) \\ \beta^2 = (-1, -1) \qquad \beta^2 = (1, 1) \end{aligned}$$

Not all hope is lost with this example however. We can still give a form of identifiability in the $p = 2$ case.

Theorem 4.3.1. *Let P_1, P_2 be arbitrary permutation matrices. Suppose*

$$P_1 S\beta_1 + P_2 S\beta_2 = P_1 S\beta_1' + P_2 S\beta_2'$$

. If $\beta_1 \geq 0$ and $\beta_1' \leq 0$, then there exists a solution (β_1'', β_2'') such that $\beta_1'' = 0$. Furthermore, $\|\beta_1''\|_1 + \|\beta_2''\|_1 \leq \|\beta_1\|_1 + \|\beta_2\|_1$ and $\|\beta_1'\|_1 + \|\beta_2'\|_1$.

Intuitively, the theorem states that the solution that minimizes L_1 norm of the difference vector β_1, β_2 does have an identifiable pattern.

This theorem follows immediately from the following Lemma.

Lemma 4.3.1. *Let P_1, P_2 be arbitrary permutation matrices. Suppose there exists a solution (β_1', β_2'), $\beta_1', \beta_2' \geq 0$, to the linear system*

$$P_1 S\beta_1 + P_2 S\beta_2 = y$$

Then, the solution (β_1'', β_2'') *to the following optimization program:*

$$\begin{aligned} &\min_{\beta_1,\beta_2} \|\beta_1\|_1 + \|\beta_2\|_1 \\ &s.t.\ P_1 S\beta_1 + P_2 S\beta_2 = y \end{aligned} \tag{4.3.4}$$

satisfies $\beta_1'' \geq 0$ *and* $\beta_2'' \geq 0$.

The lemma assumes that both component functions are increasing. This can be done without loss of generality because if $\beta_2' \leq 0$, then we can replace P_2 with the its reverse permutation matrix P_2'. $P_2 S\beta_2' = P_2'\text{reverse}(S\beta_2')$ and $\text{reverse}(S\beta_2')$ is an increasing sequence.

Lemma 4.3.1 readily implies Theorem 4.3.1. We can take β_1'', β_2'' to be the solution that minimizes the L_1 norm and Lemma 4.3.1 would imply that $\beta_1'' \geq 0$ and $\beta_1'' \leq 0$.

We will need some preparation before we can prove Lemma 4.3.1.

Definition 4.3.1. $v \in \mathbb{R}^n$ is an *alternating sign vector* if all entries of v are either $+c, -c, 0$ for some constant c, and, if $v_i, v_{i'} = -c$, then there must exist some $i < i'' < i'$ such that $v_{i'} = +c$ and vice versa. If $c = 1$, then we call v the *alternating sign unit vector*.

Let v, v' be two alternating sign vectors. We say that v, v' are *concordant* if there does not exist any i such that $v_i > 0$ and $v_i' < 0$ or $v_i < 0$ and $v_i' > 0$.

Intuitively, v is an alternating sign vector if it looks like $(+c, -c, +c, -c, ...)$ with the zeroes removed. The sum of all the v_i's is either $+c, -c$, or 0.

Proposition 4.3.1. *For any permutation matrix* P*, the columns of* DPS *are concordant alternating sign vectors. Furthermore, if* v *is an alternating sign vector, then* $DPSv$ *is also an alternating sign vector.*

Proof. (of Proposition 4.3.1)
Suppose P is associated with permutation π. We first prove that

$$(DPS)_{ij} = \begin{cases} 1 & \text{if } \pi(i) \leq j \text{ and } \pi(i+1) > j \\ -1 & \text{if } \pi(i) > j \text{ and } \pi(i+1) \leq j \\ 0 & \text{else} \end{cases}$$

. From here, it is not hard to see that a column of DPS must be an alternating sign vector. Suppose the columns are not concordant, then there exists i, j, j' such that $(DPS)_{ij} > 0$ and $(DPS)_{ij'} < 0$. Then, $\pi(i) \leq j$ and $\pi(i) > j'$ and so $j' < j$. But, $\pi(i+1) > j$ and $\pi(i+1) \leq j'$ and so $j' > j$. This is a contradiction.

To prove above identity, note that row i of DP is of the form $-\mathbf{e}_{\pi(i)} + \mathbf{e}_{\pi(i+1)}$ and column j of S is a of the form $(-\frac{n-j}{n}, ..., -\frac{n-j}{n}, \frac{j}{n}, ..., \frac{j}{n})$ with j negative elements followed by $n-j$ positive elements.

Therefore,

$$(DPS)_{ij} = (\mathbf{e}_i^\mathsf{T} DP)(S\mathbf{e}_j) = (-\mathbf{e}_{\pi(i)} + \mathbf{e}_{\pi(i+1)})^\mathsf{T}(-\frac{n-j}{n}, ..., -\frac{n-j}{n}, \frac{j}{n}, ..., \frac{j}{n})$$

The identity follows immediately.

We now focus on the second part of the proposition. Let $v \in \mathbb{R}^n$ be an alternating sign vector. Let the non-zero entries of v be $v_{i_1}, ..., v_{i_m}$. We can partition $\{1, ..., n\} = S_+ \cup S_-$ where we allocate a contiguous block $v_{i_t+1}, ..., v_{i_{t+1}} \in S_+$ if $v_{i_{t+1}} > 0$ and $v_{i_t} < 0$, else we allocate it in S_-.

For the first block $(1, ..., v_{i_1})$, we allocate it in S_+ if $v_{i_1} > 0$, else S_-. For the last block $(v_{i_m+1}, ..., v_n)$, we allocate it in S_+ if $v_{i_m} < 0$, else S_-.

Let $k = |S_-|$ and let τ be any permutation such that $\tau(i) \leq k$ for all $i \in S_-$ and $\tau(i) > k$ for all $i \in S_+$. At least one such τ exists because we defined $k = |S_-|$. Let P_τ be the permutation matrix associated with τ, then, by our previous identity, it is easy to verify that v is the k-th column of $DP_\tau S$.

$$\begin{aligned} DPS(DP_\tau S) &= DP(SD)P_\tau S \\ &= DP(I_n + \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T})P_\tau S \\ &= DPP_\tau S + \frac{1}{n}D\mathbf{1}_n\mathbf{1}_n^\mathsf{T} S \\ &= D(PP_\tau)S \end{aligned}$$

Since PP_τ is another permutation matrix, the columns of $DPS(DP_\tau S)$ must be altnerating sign vectors. Therefore, $DPSv$ is an alternating sign vector. □

Proof. (of Lemma 4.3.1)
Suppose the conditions of the theorem hold. We will show that the solution to the following optimization program is equivalent to that of Optimization 4.3.4:

$$\begin{aligned} &\min_{\beta_1, \beta_2} \|\beta_1\|_1 + \|\beta_2\|_1 \\ &\text{s.t. } P_1 S\beta_1 + P_2 S\beta_2 = y \\ &\quad \beta_1 \geq 0, \beta_2 \geq 0 \end{aligned} \tag{4.3.5}$$

Let (β_1^*, β_2^*) be the solution 4.3.5. Suppose it is not a solution to 4.3.4, then, because 4.3.4 is a convex program, there exists an arbitrarily small perturbation (v_1, v_2), both $\mathbb{R}^{n-1}$, such that

$$\|\beta_1^* + v_1\|_1 + \|\beta_2^* + v_2\|_1 < \|\beta_1^*\|_1 + \|\beta_2^*\|_1 \tag{4.3.6}$$

This is because we can set $v_1 = \gamma(\beta_1'' - \beta_1^*)$ and $v_2 = \gamma(\beta_2'' - \beta_2^*)$ where (β_1'', β_2'') is a solution to 4.3.4 and γ is an arbitrarily small positive constant. We will suppose that γ is small enough such that $|v_{1i}| < |\beta_{1i}^*|$ and $|v_{2i}| < |\beta_{2i}^*|$.

The proof will proceed in three steps. In the first step, we show that 4.3.6 cannot hold if we choose v_1, v_2 to be alternating sign vectors (Definition 4.3.1).

In the second step, we show that if v_1, v_2 are positive linear combinations of concordant alternating sign vectors, then 4.3.6 cannot hold. In the third step, we show that any vector can be written as a positive linear combination of concordant alternating sign vectors and thus show that 4.3.6 cannot hold for any (v_1, v_2).

Step 1. We let S_1 and S_2 denote the indices of the non-zero coordinates of β_1^* and β_2^* respectively. Let $\bar{v}_1, \bar{v}_2$ be alternating sign unit vectors and let $v_1 = c\bar{v}_1, v_2 = c\bar{v}_2$ where $c < |\beta_{i1}^*|$ for all $i \in S_1$ and $c < |\beta_{i2}^*|$ for all $i \in S_2$.

$$\begin{aligned}
\|\beta_1^* + v_1\|_1 &= \sum_{i \in S_1} (\beta_{i1}^* + v_{i1}) + \sum_{i \notin S_1} |v_{i1}| \\
&= \|\beta_1^*\|_1 + \sum_{i \in S_1} v_{i1} + \sum_{i \notin S_1} v_{i1} + 2 \sum_{i \notin S_1,\, v_{i1} < 0} (-v_{i1}) \\
&= \|\beta_1^*\|_1 + \mathbf{1}^\mathsf{T} v_1 + 2\mathbf{1}_{S_1^c}^\mathsf{T} v_{1-}
\end{aligned}$$

We define v_{1-} as a vector representing the negative part of v_1. $v_{i,1-} = |v_{i1}|$ if $v_{i1} < 0$, else $v_{i,1-} = 0$. The first equality of the derivation follows because $\|v_1\|_\infty < \min_{i \in S_1} |\beta_{i1}^*|$ by assumption. We can perform similar reasoning on β_2^* and get that

$$\|\beta_1^*\|_1 + \|\beta_2^*\|_1 - \|\beta_1^* + v_1\|_1 - \|\beta_2^* + v_2\|_1 = -\mathbf{1}^\mathsf{T} v_1 - \mathbf{1}^\mathsf{T} v_2 - 2\mathbf{1}_{S_1^c}^\mathsf{T} v_{1-} - 2\mathbf{1}_{S_2^c}^\mathsf{T} v_{2-}$$

First, suppose both $\mathbf{1}_{S_1^c}^\mathsf{T} v_{1-}$ and $\mathbf{1}_{S_2^c}^\mathsf{T} v_{2-}$ are both 0. Then $\beta_1^* + v_1$ and $\beta_2^* + v_2$ are both non-negative vectors. By definition of β_1^*, β_2^* as a solution to 4.3.5, it must be then that $\|\beta_1^*\|_1 + \|\beta_2^*\|_1 - \|\beta_1^* + v_1\|_1 - \|\beta_2^* + v_2\|_1 \leq 0$.

Therefore, we may assume without loss of generality that $\mathbf{1}_{S_1^c}^\mathsf{T} v_{1-} \geq c$. Because v_1 is an alternating sign vector, $\mathbf{1}^\mathsf{T} v_1 \geq -c$. Therefore,

$$-\mathbf{1}^\mathsf{T} v_1 - \mathbf{1}^\mathsf{T} v_2 - 2\mathbf{1}_{S_1^c}^\mathsf{T} v_{1-} - 2\mathbf{1}_{S_2^c}^\mathsf{T} v_{2-} \leq 2c - 2c \leq 0$$

Step 2.

Let $v_1 = \lambda u + (1-\lambda)w$ where $\lambda \in [0,1]$ and u, w are concordant alternating sign vectors. Therefore, if $v_{i1} < 0$, then $u_i \leq 0$ and $w_i \leq 0$. If $v_{i1} > 0$, then $u_i \geq 0$ and $w_i \geq 0$. Thus, $|v_{i1}| = \lambda|u_i| + (1-\lambda)|w_i|$.

Let us suppose as before that $\|v_1\|_\infty < \min_{i \in S_1} |\beta^*_{i1}|$, then,

$$\begin{aligned}
\|\beta^*_1 + v_1\|_1 &= \sum_{i \in S_1} (\beta^*_{i1} + v_{i1}) + \sum_{i \notin S_1} |v_{i1}| \\
&= \lambda \sum_{i \in S_1} (\beta^*_{i1} + u_i) + (1-\lambda) \sum_{i \in S_1} (\beta^*_{i1} + w_i) + \lambda \sum_{i \notin S_1} |u_i| + (1-\lambda) \sum_{i \notin S_1} |w_i| \\
&= \lambda \|\beta^*_1 + u\|_1 + (1-\lambda)\|\beta^*_1 + w\|_1 \\
&\geq \|\beta^*_1\|_1
\end{aligned}$$

It is straightforward to do the same analysis for a convex combination of any finite number of concordant components. The same analysis also holds for β^*_2.

Step 3. Let v_1 and v_2 be arbitrary vectors such that

$$\|v_1\|_\infty \vee \|v_2\|_\infty < \min_{i \in S_1} |\beta^*_{i1}| \wedge \min_{i \in S_2} |\beta^*_{i2}| \tag{4.3.7}$$

Let $DP_1 S v_1 + DP_2 S v_2 = 0$ so that $(\beta^*_1 + v_1, \beta^*_2 + v_2)$ is a feasible solution. Then, $v_1 = -DP_1^\mathsf{T} P_2 S v_2$ by multiplying both sides by the full ranked matrix $DP_1^\mathsf{T} S$.

Let τ be the permutation that represent the ordering of Sv_2. That is, $(Sv_2)_{\tau(1)}$ is the smallest entry, $(Sv_2)_{\tau(2)}$ is the second smallest entry, etc.

Then, $DP_\tau S v_2 = \vec{\lambda} \geq 0$. Therefore, $v_2 = DP_\tau^\mathsf{T} S \vec{\lambda}$. Because the columns of $DP_\tau^\mathsf{T} S$ are concordant alternating sign vectors, v_2 can be written as a convex combination of concordant alternating sign vectors.

v_1 shares a similar decomposition $v_1 = -DP_1^\mathsf{T} P_2 S v_2 = -DP_1^\mathsf{T} P_2 P_\tau^\mathsf{T} S \vec{\lambda}$.

Let E_{i2} and E_{i1} be the i-th column of $DP_\tau^\mathsf{T} S$ and $-DP_1^\mathsf{T} P_2 P_\tau^\mathsf{T} S$ respectively.

$$DP_1 S E_{i1} + DP_2 S E_{i2} = DP_1 S(-DP_1^\mathsf{T} P_2 P_\tau^\mathsf{T} S)\mathbf{e}_i + DP_2 S(DP_\tau^\mathsf{T} S)\mathbf{e}_i = 0$$

We have shown that $v_1 = E_1 \vec{\lambda}$ and $v_2 = E_2 \vec{\lambda}$ where each pair of columns of E_{i1}, E_{i2} satisfy $DP_1 S E_{i1} + DP_2 S E_{i2} = 0$. Therefore, by what we have shown in Step 1 and Step 2, it must be that

$$\|\beta^*_1\|_1 + \|\beta^*_2\|_1 \leq \|\beta^*_1 + v_1\|_1 + \|\beta^*_2 + v_2\|_1$$

□

We have established identifiability for the $d = 2$ case: the pattern is unique and is recovered by the sign pattern of the solution that minimizes L_1-norm of the differences, as shown in Lemma 4.3.1.

Remark 4.3.1. Lemma 4.3.1 does not exactly apply when $d > 2$. Let us be more precise. Suppose there exists $\beta'_1, ..., \beta'_p \geq 0$ such that $\sum_{j=1}^{p} DP_j S\beta'_j = y$. If $\beta''_1, \beta''_2, ...\beta''_p$ satisfies $\sum_{DP_jS} \beta''_j = y$ and minimzies $\sum_{j=1}^{p} \|\beta''_p\|_1$, it is no longer guaranteed that $\beta''_j \geq 0$ as well. Instead, using randomly generated permutations P_j's, what we observe empirically is that β''_j may have small negative entries but $\mathbf{1}^\mathsf{T}\beta''_j$ is still overwhelmingly positive. A rigorous analysis of this is an important direction of future work.

4.4 Estimation

Motivated by Theorem 4.3.1 and Remark 4.3.1 where L_1 norm of the difference is an important quantity to minimize, we propose the following generalized Lasso estimation procedure for pattern selection:

$$\min_{\beta_j} \|y - \sum_{j=1}^{d} P_j S\beta_j\|_2^2 + \lambda \sum_{j=1}^{d} \|\beta_j\|_1 \tag{4.4.1}$$

Let us reformulate this optimization to get a better intuitive understanding. Observe that we can decompose $S\beta_j = S\beta_{j,+} + S\beta_{j,-} = f_j + g_j$ where $\beta_{j,+}$ contains the positive entries of β_j and $\beta_{j,-}$ contains the negative entries. Therefore, f_j is increasing and g_j is decreasing. It is clear then that $\|\beta_j\|_1 = f_{nj} - f_{1j} + g_{1j} - g_{nj}$ because f_{nj}, g_{1j} are the largest entries of the sequences f_j, g_j and f_{1j}, g_{nj} are the smallest entries.

Therefore, the optimization is equivalent to

$$\begin{aligned} &\min_{f_j, g_j} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d} (P_j(f_j + g_j))_i \right)^2 + \lambda \sum_{j=1}^{d} (f_{nj} - f_{1j} + g_{1j} - g_{nj}) \\ &\text{s.t. for } i = [1, n-1], \text{ for all } j,\ f_{i+1,j} \geq f_{ij},\ g_{i+1,j} \leq g_{ij} \\ &\quad \text{for all } j, \sum_{i=1}^{n} f_{ij} = 0, \sum_{i=1}^{n} g_{ij} = 0 \end{aligned} \tag{4.4.2}$$

BACKFITTING ALGORITHM FOR ADDITIVE TREND FILTERING

Input: $(X_1, y_1), ..., (X_n, y_n)$, regularization parameter λ

Initialization: Set vectors $\widehat{f}_j, \widehat{g}_j$ as all 0 for $j = 1, ..., p$.

Repeat until convergence, for $j = 1, ..., p$:

Set $y_{\text{res}} = y - \sum_{j' \neq j} P_{j'}(\widehat{f}_{j'} + \widehat{g}_{j'})$.

Do trend filter update:

$$\widehat{f}_j, \widehat{g}_j = \underset{f_j, g_j}{\arg\min} \frac{1}{n} \left\| y_{\text{res}} - P_j(f_j + g_j) \right\|_2^2 + \lambda_t (f_{nj} - f_{1j} + g_{1j} - g_{nj}) \qquad (4.4.3)$$
$$\text{s.t.} \forall i,\ f_{ij} \leq f_{i+1,j},\ g_{ij} \geq g_{i+1,j}$$

Output: $\text{pattern}_j = +1$ if $\|\widehat{f}_j\|_\infty > \|\widehat{g}_j\|_\infty$, and -1 if $\|\widehat{g}_j\|_\infty > \|\widehat{f}_j\|_\infty$.

Figure 4.1: Backfitting algorithm for additive trend filtering. Any solver can be used at 4.4.3. λ_t can be iteration dependent so long as $\lambda_t \to \lambda$. We suggest $\lambda_t = \lambda\left(1 + e^{-at+b}\right)$ for $0 < a \leq 1/2$ and $b \geq 5$.

This procedure is an additive 0-th order trend filtering (Tibshirani et al., 2014). (0-th order trend filtering is precisely this optimization with $d = 1$) A solution from 4.4.2 is also a solution to 4.4.1 with the transformation $\beta_{j,+} = Df_j$ and $\beta_{j,-} = Dg_j$.

4.4.2 can be interpreted as the convex relaxation of the mixed integer Add-MM program (4.2.1). Instead of forcing either f_j or g_j to be zero, we instead allow both to be non-zero but promote sparsity with the Lasso like penalty.

Optimization

To solve optimization 4.4.2, we propose a backfitting scheme. At every iteration, we update a single f_j, g_j and fix all other $f_{j'}, g_{j'}$'s. The optimization at each iteration is a quadratic program and can be solved via QP software such as MOSEK. For our experiments, we instead use the R package `glmgen` (Arnold et al. (2014)), which implements an ADMM algorithm described by Ramdas and Tibshirani (2014).

4.5 Pattern Selection Consistency

In this section, we analyze the pattern selection consistency of additive trend filtering (figure 4.1). More precisely, we want to show that the estimated pattern is equal to the true pattern with probability convergent to 1.

under the following stochastic assumptions:

A1 We suppose that the true regression function is additive

$$y_i = f_1^*(x_{i1}) + f_2^*(x_{i2}) + ...f_d^*(x_{id}) + \varepsilon_i$$

where each additive component f_j^* is, without loss of generality, monotone increasing. If f_j^* is decreasing, we can analyze $-x_j$ instead so that f_j is increasing.

A2 We suppose that X is drawn from a positive product density. That is, $X_j, X_{j'}$ are independent for $j \neq j'$.

A3 ε_i is an independent subgaussian random variable with scale σ.

A4 $\|f_j^*\|_\infty \leq B$ for some constant B. We suppose that the constraints $\|f_j\|_\infty, \|g_j\|_\infty \leq B$ are also added onto optimization 4.4.2.

A5 Let $\alpha \leq \min_j Ef_j^*(X_j)^2$. We suppose that $\alpha > 0$.

A1 can be weakened: if the true regression function $f(\mathbf{x}) = \mathbb{E}[y \mid X = \mathbf{x}]$ is non-additive, then we can compare our finite sample estimate against the population setting additive projection of $f(\mathbf{x})$ instead.

A2 is a strong assumption. However, our experiments show that our estimator is effective even when X has significant correlation. Extending the consistency analysis to correlated data is an important direction of future work.

In A4, we suppose that the estimation procedure has extra B-bounded constraints. This is entirely to make the theoretical analysis convenient; we do not use these constraints in the experiments.

A5 defines the signal level of this problem. Larger α implies easier pattern selection.

Our analysis considers a restricted form of optimization 4.4.2 where we force $g_j = 0$:

$$\begin{aligned} &\min_{f_j} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d} (P_j f_j)_i \right)^2 + \lambda \sum_{j=1}^{d} (f_{nj} - f_{1j}) \\ &\text{s.t. for } i = 1, ..., n-1, \text{ for all } j,\ f_{i+1,j} \geq f_{ij} \\ &\quad \text{for all } j,\ \sum_{i=1}^{n} f_{ij} = 0 \end{aligned} \tag{4.5.1}$$

Our analysis proceeds in two parts. In the first part, we show that the solution $\widehat{f}_j$ of the restricted optimizatio 4.5.1, along with $\widehat{g}_j = 0$, is also the solution to the full optimization 4.4.2. This shows that $\|\widehat{g}_j\|_\infty$ is not greater than $\|\widehat{f}_j\|_\infty$. In the second part, we show that $\|\widehat{f}_j\|_\infty > 0$.

Part One

In this part, our goal is to show that the output $\widehat{f}_j, \widehat{g}_j$ of optimization 4.4.2 satisfy that $\widehat{g}_j = 0$ with high probability.

The KKT theorem gives a standard set of conditions for optimality. But, given the special structure of Optimization 4.4.2, it will be convenient to transform the KKT conditions into an equivalent set of conditions on the partial sums of the derivatives.

Theorem 4.5.1. *Suppose y has mean zero. (f_j, g_j) are the output of optimization 4.4.2 if and only if for all $j = [1, p]$, for all $t < n$,*

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^{t} \left(P_j^\mathsf{T} \left(\sum_{k=1}^{d} P_k (f_k + g_k) - y \right) \right)_i \leq \lambda \quad \textit{equal if } f_{tj} < f_{t+1,j} \\ &\frac{1}{n} \sum_{i=1}^{t} \left(P_j^\mathsf{T} \left(\sum_{k=1}^{d} P_k (f_k + g_k) - y \right) \right)_i \geq -\lambda \quad \textit{equal if } g_{tj} > g_{t+1,j} \end{aligned}$$

For $t = n$, $\frac{1}{n} \sum_{i=1}^{n} \left(P_j^\mathsf{T} \left(\sum_{k=1}^{d} P_k (f_k + g_k) - y \right) \right)_i = 0$.

Proof. We take the Lagrangian of Optimization 4.4.2. Let α_{ij} and α'_{ij} be non-negative

Lagrangian multipliers.

$$\mathcal{L}(f,g,\alpha,\alpha') = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{d}(P_j(f_j+g_j))_i\right)^2 + \lambda\sum_{j=1}^{d}(f_{nj} - f_{1j} + g_{1j} - g_{nj}) + \sum_{j=1}^{p}\sum_{i=1}^{n-1}\alpha_{ij}(f_{ij} - f_{i+1,j}) + \sum_{j=1}^{p}\sum_{i=1}^{n-1}\alpha'_{ij}(g_{i+1,j} - g_{ij}) + \mu_j\sum_{i=1}^{n} f_{ij} + \mu'_j\sum_{i=1}^{n} g_{ij}$$

We differentiate this with respect to f_j:

$$\frac{\partial\mathcal{L}(f,g,\alpha,\alpha')}{\partial f_j} = \frac{1}{n}P_j^{\mathsf{T}}\left(\sum_{k=1}^{d}P_k(f_k+g_k) - y\right) + \lambda(\mathbf{e}_n - \mathbf{e}_1) + \bar{\alpha}_j - \underline{\alpha}_j + \mu_j\mathbf{1}_n$$

The vector $\bar{\alpha}_j = (\alpha_j, 0)$ and $\underline{\alpha}_j = (0, \alpha_j)$. For illustration, the first couple of terms are

$$\begin{aligned}(\bar{\alpha}_j - \underline{\alpha}_j)_1 &= \alpha_{1j}\\ (\bar{\alpha}_j - \underline{\alpha}_j)_2 &= \alpha_{2j} - \alpha_{1j}\\ &\ldots\\ (\bar{\alpha}_j - \underline{\alpha}_j)_n &= -\alpha_{nj}\end{aligned}$$

Therefore, the t-th partial sum $\sum_{i=1}^{t}(\bar{\alpha}_j - \underline{\alpha}_j)_i = \alpha_{tj}$ for $t < n$. The partial sum is 0 for $t = n$.

KKT states that $\frac{\partial\mathcal{L}(f,g,\alpha,\alpha')}{\partial f_j} = 0$, therefore, all partial sums of $\frac{\partial\mathcal{L}(f,g,\alpha,\alpha')}{\partial f_j}$ must also be zero. The partial sums of $\lambda(\mathbf{e}_n - \mathbf{e}_1)$ is $(-\lambda, -\lambda, ..., 0)$. The partial sums of the derivative is then, for $t < n$,

$$\frac{1}{n}\sum_{i=1}^{t}\left(P_j^{\mathsf{T}}\left(\sum_{k=1}^{d}P_k(f_k+g_k) - y\right)\right)_i - \lambda + \alpha_{tj} = -\mu_j = 0$$

$\mu_j = 0$ because, if we let $t = n$, we have that $\frac{1}{n}\sum_{i=1}^{n}\left(P_j^{\mathsf{T}}\left(\sum_{k=1}^{d}P_k(f_k+g_k) - y\right)\right)_i = -\mu_j$. y is assumed to have mean zero in the theorem and f_k, g_k for k all have mean zero because of primal feasibility and therefore, $\mu_j = 0$.

Since $\alpha_{ij} = 0$ if $f_{i+1,j} > f_{ij}$, the part of the theorem regarding f_j follows. The statement regarding g_j can be worked out in exactly the same fashion.

□

Let us without loss of generality focus on $j = 1$ and suppose that $P_1 = I$. Let us also plug in the expression for y. Theorem 4.5.1 then states that

$$\frac{1}{n}\sum_{i=1}^{t}\left(f_1 + g_1 - f_1^* + \sum_{k=2}^{d} P_k(f_k + g_k - f_k^*) + \varepsilon\right)_i \leq \lambda$$
$$\frac{1}{n}\sum_{i=1}^{t}\left(f_1 + g_1 - f_1^* + \sum_{k=2}^{d} P_k(f_k + g_k - f_k^*) + \varepsilon\right)_i \geq -\lambda$$

$\frac{1}{n}\sum_{i=1}^{t}\varepsilon_i$ can be shown to be at most $\sqrt{\frac{1}{n}}$ with high probability. If we assume that the X_k's are independent and that $f_k + g_k - f_k^*$ are bounded, then $\sum_{k=2}^{d} P_k(f_k + g_k - f_k^*)$ behaves like an independent noise and its partial sums can be shown to have magnitude at most $\sqrt{\frac{d}{n}}$.

If we treating $\sum_{k=2}^{d} P_k(f_k + g_k - f_k^*)$ as noise, we can show that f_1, g_1 is pattern consistent.

Proposition 4.5.1. *Suppose that* $\max_t |\frac{1}{n}\sum_{i=1}^{t}\varepsilon_i| \leq \sigma\sqrt{\frac{\log(1/\delta)}{2n}}$ *with probability at least* $1-\delta$.

Suppose also that $\max_t \left|\frac{1}{n}\sum_{i=1}^{t}\left(\sum_{k=2}^{d} P_k(f_k + g_k - f_k^*)\right)_i\right| \leq dB\sqrt{\frac{\log(1/\delta)}{2n}}$ *with probability at least* $1-\delta$.

Then, if $\lambda \geq (dB+\sigma)\sqrt{\frac{\log(2/\delta)}{2n}}$, *we have that* $g_1 = 0$ *with probability at least* $1-\delta$.

Proof. Take optimization (4.4.2) and hold f_j, g_j fixed for $j > 1$. Suppose we now hold $g_1 = 0$ and solve for f_1. If the solution we get satisfies the KKT condition for jointly optimizing both f_1, g_1, then we can say that $g_1 = 0$ even if we do not hold it to be zero.

If we solve for f_1 while holding $g_1 = 0$, we know that [todo! elaborate] the first KKT statement in Theorem 4.5.1 holds, that for all $t < n-1$.

$$\frac{1}{n}\sum_{i=1}^{t}\left(f_1 - f_1^* + \sum_{k=2}^{d} P_k(f_k + g_k - f_k^*) + \varepsilon\right)_i \leq \lambda \quad \text{equal if } f_{tj} > f_{t+1,j}$$

For $t = n$, we have that $\frac{1}{n}\sum_{i=1}^{n}\left(f_1 - f_1^* + \sum_{k=2}^{d} P_k(f_k + g_k - f_k^*) + \varepsilon\right)_i = 0$. Since $f_k, g_k, f_k^*, \varepsilon$ are all assumed to have mean zero, it must be that $\sum_{i=1}^{n}(f_1 - f_1^*)_i = 0$ as well.

We need to prove the second statement:

$$\frac{1}{n}\sum_{i=1}^{t}\left(f_1 - f_1^* + \sum_{k=2}^{d} P_k(f_k + g_k - f_k^*) + \varepsilon\right)_i > -\lambda$$

We claim that for all t, $\frac{1}{n}\sum_{i=1}^{t}(f_1 - f_1^*)_i \geq 0$. Suppose for sake of contradiction that this is not true and let t be the first instance on which $\frac{1}{n}\sum_{i=1}^{t}(f_1 - f_1^*)_i < 0$. It is obvious then that $f_{1t} - f_{1t}^* < 0$.

Suppose $f_{1t'} = f_{1t}$ for all $t' > t$, then $f_{1t'} - f_{1t'}^* < f_{1t} - f_{1t}^* < 0$. This implies that $\sum_{i=1}^{n}(f_1 - f_1^*)_i < 0$. This is a contradiction because KKT implies that $\sum_{i=1}^{n}(f_1 - f_1^*)_i = 0$.

So, there must exist a $t' > t$ such that t' is the smallest index where $f_{1t'} > f_{1t}$. Then, $f_{1t'} > f_{1,t'-1}$ and therefore, $\frac{1}{n}\sum_{i=1}^{t'-1}\left(f_1 - f_1^* + \sum_{k=2}^{d} P_k(f_k + g_k - f_k^*) + \varepsilon\right)_i = \lambda$. Moving terms around, we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{t'-1}(f_1 - f_1^*)_i &= \lambda - \frac{1}{n}\sum_{i=1}^{t'-1}\left(\sum_{k=2}^{d} P_k(f_k + g_k - f_k^*) + \varepsilon\right)_i \\
&\geq \lambda - d\sqrt{\frac{c_2 \log(2/\delta)}{n}} - \sqrt{\frac{c_1 \log(2/\delta)}{n}} \quad \text{w.p. at least } 1-\delta \\
&\geq 0
\end{aligned}$$

Since $f_{1,t'-1} = f_{1t}$ by definition of t', we have that $\sum_{i=1}^{t'-1}(f_1 - f_1^*)_i \leq \sum_{i=1}^{t}(f_1 - f_1^*)_i < 0$. We have reached another contradiction.

□

Proposition 4.5.2. *Suppose the estimated f_j, g_j satisfies $\|f_j\|_\infty, \|g_j\|_\infty \leq B$ for some constant B. Then, we have, with probability at least $1 - \frac{1}{n}$ that*

$$\max_t \left|\frac{1}{n}\sum_{i=1}^{t}\varepsilon_i\right| \leq \sigma\sqrt{\frac{1}{2n}\log^2 2n}$$

$$\max_t \left|\frac{1}{n}\sum_{i=1}^{t}\left(\sum_{j'\neq j} P_{j'}\left(f_j + g_j - f_j^*\right)\right)_i\right| \leq 3dB\sqrt{\frac{1}{2n}\log 2np} \quad \textit{for all } j$$

Proof. We first prove the first inequality. Because ε_i is subgaussian with scale σ, with probability at least $1 - \frac{1}{n}$, we have that $|varepsilon_i| \leq \sigma\sqrt{\log 2n}$ for all i.

Then, we apply Serfling's concentration for replacing without replacement (Corollary 2.7.2) with a union bound and immediately derive the first inequality.

The second inequality follows similarly. Because $\|f_j\|_\infty, \|g_j\|_\infty \leq B$,

$$\max_i \left|\left(\sum_{j'\neq j} P_{j'}(f_{j'} + g_{j'} - f_{j'}^*)\right)_i\right| \leq 3dB$$

. We apply Serfling's theorem again with a union bound and get the second inequality.

□

Combining Proposition 4.5.1 and Proposition 4.5.2, we have the following pattern selection consistency result.

Theorem 4.5.2. *Suppose assumptions A1-A4 hold. Let* $\widetilde{\sigma} = \max(\sigma, B)$.

Suppose $\lambda \geq 3d\widetilde{\sigma}\sqrt{\frac{1}{2n}\log^2 2np}$. *Let* $\widehat{f}_j, \widehat{g}_j$ *be the output of 4.4.2, then, we have, with probability at least* $1 - \frac{1}{n}$, *that*

$$\text{for all } j,\ \widehat{g}_j = 0$$

Part Two

Our goal in this part is to show that the output $\widehat{f}_j$ of optimization 4.5.1 satisfies $\|\widehat{f}_j\|_\infty > 0$ with high probability.

The key is to observe that the set of bounded additive monotone functions has a bounded bracketing entropy and therefore, similar to the false positive analysis of AC/DC (Theorem 2.7.3, we will show that the population risk $\mathbb{E}(\widehat{f}) - f^*)^2 \to 0$ as $n \to \infty$. If $\widehat{f}_j = 0$ for some j, then $\mathbb{E}(\widehat{f} - f^*)^2 \geq \alpha > 0$ where α is the signal level defined in Assumption A5. This gives us a contradiction and we can thus conclude that $\widehat{f}_j \neq 0$ for all j.

Proposition 4.5.3. *Suppose assumptions A1-A4 hold. Let* $\widetilde{\sigma} = \max(\sigma, B)$.

Let $\widehat{f}_j$ *be the output of the restricted optimization 4.5.1 with* $\lambda \leq cd\widetilde{\sigma}\sqrt{\frac{1}{n}\log^2 c'np}$. *Then, with probability at least* $1 - \frac{1}{n}$, *we have that*

$$\mathbb{E}\left(\sum_{j=1}^{d} f_j^*(X_j) - \widehat{f}_j(X_j)\right)^2 \leq c'' B^2 \widetilde{\sigma}\sqrt{\frac{d^6}{n^{2/3}}\log^2 c'nd}$$

where c, c', c'' *are absolute constants.*

The proof of Proposition 4.5.3 is identical to that of Theorem 2.7.3 in Chapter 2. We need only the following bracketing number result.

Proposition 4.5.4. *Let* $\mathcal{M}_B^1$ *be the set of univariate monotonic increasing functions bounded by* B *and let* P *be a distribution over* $\mathbb{R}$. *Then, we have that the bracketing entropy of* $\mathcal{M}_B^1$ *is bounded by*

$$\log N_{[]}(\epsilon, \mathcal{M}_B^1, L_1(P)) \leq \frac{2KB}{\epsilon}$$

for some constant K.

Let $\mathcal{M}_B^d = \{f = \sum_{j=1}^d f_j : f_j \in \mathcal{M}_B^1\}$ be the set of additive bounded monotone functions and let P be a distribution over $\mathbb{R}^d$. Then,

$$\log N_{[]}(\epsilon, \mathcal{M}_B^d, L_1(P)) \leq \frac{2KBd^2}{\epsilon}$$

Proof. The first result is well known (Van der Vaart and Wellner (1996)). The second result is derived from the first result. We observe that we can construct an ϵ bracketing of $\mathcal{M}_B^d$ by taking d different $\frac{\epsilon}{d}$-bracketings of $\mathcal{M}_B^1$. □

Proposition 4.5.3 along with Lemma 2.7.3 give the following theorem:

Theorem 4.5.3. *Suppose assumptions A1-A5 hold. Let $\widetilde{\sigma} = \max(\sigma, B)$.*
Let $\widehat{f}_j$ be the output of 4.5.1 with $\lambda \leq cd\widetilde{\sigma}\sqrt{\frac{1}{n}\log^2 c'np}$.
Suppose n is large enough such that

$$c''\frac{d^6}{n^{2/3}}\log^2 c'nd \leq \alpha$$

for some constants c', c''. Then, we have that, with probability at least $1 - \frac{1}{n}$, $\|\widehat{f}_j\|_\infty > 0$.

Combining the Two Steps

Suppose Theorem 4.5.2 holds, then the solution to 4.5.1 is also a solution to 4.4.2. Therefore, we can combine Theorem 4.5.2 and Theorem 4.5.3 to get the following result:

Corollary 4.5.1. *Suppose assumptions A1-A5 hold. Let $\widetilde{\sigma} = \max(\sigma, B)$. Let $\widehat{f}_j, \widehat{g}_j$ be the output of 4.4.2 with $\lambda = \Theta\left(d\widetilde{\sigma}\sqrt{\frac{1}{n}\log^2 nd}\right)$. Suppose n is large enough such that*

$$c''\frac{d^6}{n^{2/3}}\log^2 c'nd \leq \alpha.$$

for some constants c', c''. Then, with probability at least $1 - \frac{1}{n}$, we have that $\widehat{g}_j = 0$ and $\widehat{f}_j \neq 0$ and therefore,

$$\|\widehat{f}_j\|_\infty > \|\widehat{g}_j\|_\infty$$

The above result is a preliminary theoretical justification of why additive trend filtering can be used for pattern selection. The rate $\frac{d^6}{n^{2/3}}$ is clearly suboptimal and many of the assumptions are clearly stronger than they need to be, as shown in our experiments.

4.6 Experiment

We perform experiments with the proposed pattern selection procedure on both synthetic and real data.

Synthetic Data

We first test pattern selection accuracy and predictive accuracy with simulated data. Our X is Gaussian with a covariance Σ. y is generated as $y_i = f(\mathbf{x}_i) + \varepsilon_i$ where ε_i is an independent normal noise and f is additive $f = \sum_{j=1}^{d} f_j$. Each f_j is randomly selected as one of the following four increasing functions or its negative, which is a decreasing function.

1. Exponential. $f_j(x_j) = ae^{x_j} - b$, a, b are set so that f_j is mean zero, $\|f_j\|_\infty \leq 3$.
2. Negative reverse of the exponential. $f_j(x_j) = -(ae^{-x_j} - b)$
3. Single step. $f_j(x_j) = -a$ if $x_j < 0.7$ and $f_j(x_j) = 1 - a$ if $x_j \geq 0.7$ where a is chosen so that f_j has mean zero.
4. Double step. $f_j(x_j)$ is a sum of two step functions, one with threshold at 0 and one with threshold at 0.7.

For our experiment, Σ is generated as $cDZ^\mathsf{T}ZD + (1-c)\mathbf{1}_d\mathbf{1}_d^\mathsf{T}$ where Z is a Gaussian random matrix (Z_{ij}'s are iid Gaussian), and D is a diagonal matrix such that $DZ^\mathsf{T}ZD$ has 1's on the diagonal. $c \in [0, 1]$ is a knob we can turn to adjust the level of correlation; larger c leads to a lower level of correlation. We set $c = 2/3$ in all of our experiments. The standard deviation of noise ε_i is chosen so that the signal-to-noise ratio is 4.

We compare additive trend filtering against two baseline methods. The first baseline method is the OLS **linear** fit: we take simply the sign of the linear fit $(X^\mathsf{T}X)^{-1}X^\mathsf{T}y$ as the estimated pattern. The second baseline method is a **naive** method based on marginal regression: for each j, we compare $y_{(n)}$ and $y_{(1)}$ where (n) is the index of the largest entry of X_j and (1) is the index of the smallest entry. We say that f_j is increasing if $y_{(n)} \geq y_{(1)}$ and decreasing otherwise.

For each trial, we generate X, y as described and run additive trend filtering as well as the linear estimator and the naive estimator. We select λ for additive trend filtering via 4-fold cross validation where we choose the λ that minimizes the CV predictive error. We plot the percentage of pattern recovery errors of all

three methods as well as the predictive errors of additive trend filtering and linear regression. Each point on the plot shows the average of 15 independent trials.

Figure 4.2 shows how the pattern recovery error and predictive error vary with the sample size. As can be seen, additive trend filtering overperforms the baseline methods. Furthermore, the pattern recovery errors of the linear method and the naive method do not decrease significantly with increasing sample size. Figure 4.3 shows how pattern recovery error and predictive error vary with the dimensionality.

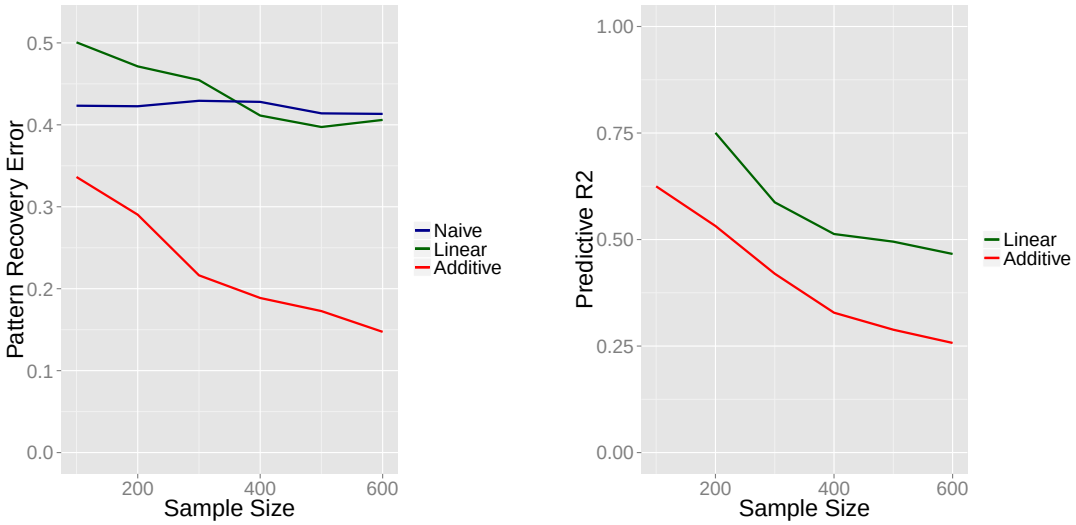


Figure 4.2: Experimental result where n varies from 100 to 600. $p = 100$. The left plot shows pattern recovery error (random guess yields 0.5). The right plot shows predictive R2 error.

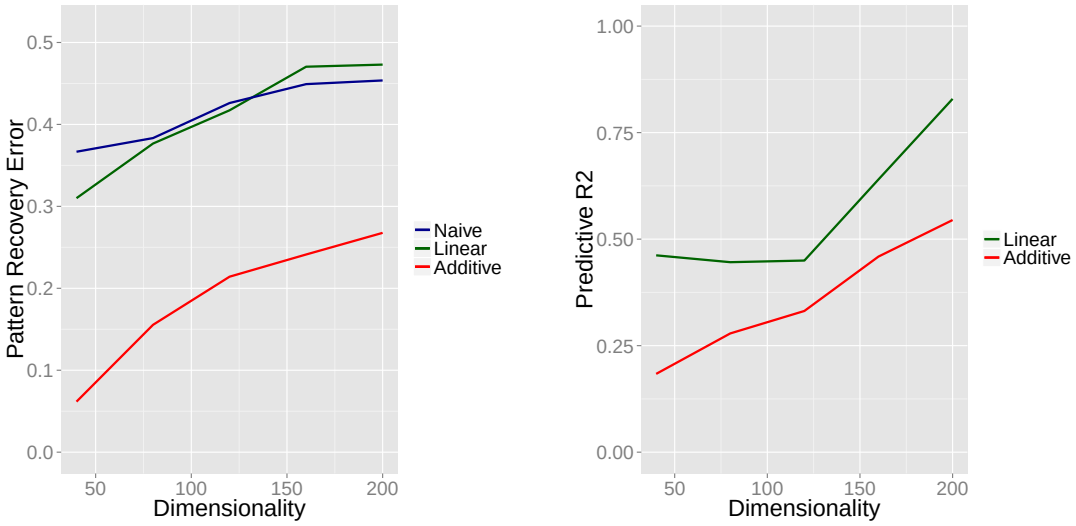


Figure 4.3: Experimental result where p varies from 40 to 200. $n = 400$. The left plot shows pattern recovery error (random guess yields 0.5). The right plot shows predictive R2 error.

Real Data

We use the Boston housing dataset from the UC Irvine repository to evaluate the effectiveness of our proposed pattern selection method. Because we do not the "true" orientation pattern in the Boston housing data, we cannot report pattern recovery accuracy. We instead report the predictive accuracy of a model built from the learned orientation. More precisely, we first perform pattern selection and then refit a shape-constrained model according to the selected pattern. We compare against the predictive accuracy of an additive spline model. The predictive accuracy is measured through R2 over 50 random subsamples.

The dataset itself has about 500 data points and 13 features, not including the median housing price–the quantity we predict. We discard 2 discrete features and keep the remaining 11. We perform two sets of experiments, one with training data size $n = 400$ and one with training data size $n = 100$. In each trial, we randomly select the n training data pointsto learn the model and evaluate the model on the remaining data points to measure predictive R2. We plot the accuracy averaged from all the trials.

We perform the experiments with both the monotone shape-constraint and the convex/concave shape-constraint. Figure 4.4 shows the result for $n = 400$ training

data points; shape-constrained methods perform slightly worse but are more stable. The real advantage of shape-constrained methods appears in the $n = 100$ case, shown in Figure 4.5, where additive B-spline model overfits and the shape-constrained models are still able to achieve reasonable predictive accuracy.

	R2 +/- sd
monotone	0.75 +/- 0.05
convex/concave	0.77 +/- 0.065
B-splines	0.79 +/- 0.09

Figure 4.4: Boston experiment results for $n = 400$ training data.

	R2 +/- sd
monotone	0.66 +/- 0.06
convex/concave	0.65 +/- 0.012
B-splines	0.35 +/- 0.750

Figure 4.5: Boston experiment results for $n = 100$ training data.

CHAPTER 5

DISCUSSION

We have studied variable selection for a sparse convex/concave function. We show, for both the regression setting and the discrete choice model setting, that a procedured based on a shape-constrained additive model is effective theoretically and practically. Our analysis supposes that the underlying function satisfies the shape-constraint assumptions but our method is useful even with model misspecification. Indeed, the goal of the dissertation is to demonstrate that estimators based on high dimensional shape-constrained models can be a practical generalization of popular methods like the lasso. Shape-constrained estimators have no smoothing bandwidth and are easily interpretable; they are, in a sense, as easy to use as the lasso. In instances where the underlying function is significantly nonlinear, shape-constrained estimator presents a good alternative to linear models–it is an easy way to trade off model simplicity to improve model fitness.

One caveat of using shape-constrained estimation under model misspecification is that one can no longer choose an orientation (increasing vs decreasing, convex vs concave, etc) by prior knowledge and intuition. This becomes a problem if the number of possible orientation patterns explode exponentially with the dimensionality. Chapter 4 of the dissertation aims to rectify this problem by proposing a method to automatically select an orientation pattern for additive shape-constrained models.

In this concluding chapter, we first discuss specific open questions relevant to our work. We then list two general directions of future work in line with our goal of making high dimensional shape-constrained estimation a practical alternative to linear models.

5.1 Loose Ends

Our analyses in each of the chapters have many loose ends. These questions are important toward establishing a more complete theoretical justification of high dimensional shape-constrained estimation.

Chapter 2

A minor question in our population level analysis is whether one could prove additive faithfulness without the twice-differentiability asumption on the convex function f_0. A more important question is whether the converse to additive faithfulness holds under conditions more general than those discussed in section 2.3.

Our finite-sample analysis has much room for improvement. The current rate, $\frac{s^5 \log^2 np}{n^{4/5}}$, is likely suboptimal. More delicate proof technique is necessary to achieve a better rate. The independent assumption (Assumption A1 in section 2.5 is restrictive. It remains an open question what is the nonparametric analogue of mutual incoherence (Wainwright, 2009). Such a condition would remove the reliance on the independence assumption from Theorem 2.5.2. Our assumption that $\|\widehat{f}_k\|_\infty, \|\widehat{g}\|_\infty \leq B$ is unsatisfactory. We conjecture that these bounds hold automatically with high probability but a mathematical proof is still beyond our reach.

Chapter 3

Real decision problems often involve discrete varibles such as binary variables and count variables. To make our estimator more practical, we need to incorporate discrete variables into our model and we need to analyze conditions under which additive faithfulness holds even with extraneous discrete variables.

On the theoretical side, the finite sample properties of our estimator is still unknown. It would be reasonable to assume that n, the number of consumers, and m, the number of items, must both be larger in order to achieve variable selection consistency. But, it is not obvious how s the sparsity level and p the ambient dimensionality feature in the rate.

Chapter 4

As described in the last paragraph of section 4.3, we do not yet understand identifiability for dimension greater than two. Theorem 4.3.1 does not hold exactly when $d > 2$: we observe empirically that the L_1 minimizing solution $\widehat{\beta}_j$'s may not have the perfect sign pattern but the output functions $\widehat{f}_j = S\widehat{\beta}_j$ approximately follow the correct orientation. This observation is based on randomly generated covariate permutation matrices P_j's. Adversarially generated P_j's may not exhbit any degree of identifiability.

Our pattern selection consistency result is preliminary. Our proof technique critically relies on the assumption that the covariates P_j's are independent. Independence

is an easy setting under which naive pattern selection methods such as marginal regression are also consistent and effective. The truly interesting setting–one which demonstrates the advantages of our proposed method–is that of correlated covariates. Indeed, simulations show that additive trend filtering can retrieve the correct pattern even with moderate degree of correlation.

5.2 Future Directions

We propose two directions of future work, the goal of which is to provide shape-constrained estimation with greater applicability. The first is to study the estimation of high-dimensional log-concave densities–which is a very general class of densities. The second is to improve the computational efficiency, possibly making the computation adaptive to the complexity of the output.

Log-concave Density Estimation

Log-concave density is of the form $p(\mathbf{x}) = \exp f(\mathbf{x})$ where f is a concave function. This class of densities include many commonly seen models such as the Gaussian, Laplace, and (for certain parameter settings) Dirichlet distributions. Initially proposed by Grenander (1956) for univariate densities, log-concave density estimation has since been extended to the multivariate regime (Cule et al., 2010b), with an established rate of convergence (Cule et al., 2010a; Kim and Samworth, 2014).

Log-concave density estimation suffers from the curse of dimensionality. The maximum likelihood estimation algorithm proposed by Cule et al. (2010b) becomes inpractical when d becomes moderately large–on the order of $d = 10, n = 500$. The rate of convergence of the MLE is $n^{-1/(d-1)}$ for $d \geq 4$ (Kim and Samworth, 2014) and slows down exponentially with the dimensionality. Minimax rate, established by Kim and Samworth (2014) also, shows the high-dimensional estimation is impossible without some notion of sparsity.

A reasonable notion of sparsity for a density is that of a sparse conditional independence graph. This notion of sparsity, in the case of the multivariate Gaussian distribution, corresponds to a sparse inverse covariance matrix. For log-concave density, one possible way to enforce this sparsity assumption is to say that $f(\mathbf{x}) = \sum_j f_j(x_j) + \sum_{j,k} f_{jk}(x_j, x_k)$ where $f_{jk} = 0$ for most (j, k) pairs; Clifford-Hammersley theorem directly implies that the conditional independence graph must be sparse. The difficulty is to have such a decomposition, enforce concavity, and promote sparsity all at the same time.

One approach we attempted is to generalize the regression based neighborhood search methods of Meinshausen and Bühlmann (2006). The idea is to replace a high-dimensional linear regression with a high-dimensional convex or concave regression instead. Unfortunately, this procedure is difficult to justify theoretically. It is both unclear whether the condition first moments satisfy any natural shape-constraints and whether the conditional first moments of a log-concave density correspond to the conditional graph structure.

Computational Efficiency

Lasso is the standard method for high-dimensional predictive modeling in a large part because it is fast. Rapid advances in clever algorithms and technologies have made computation feasible for millions and even billions of variables. In light of the popularity of lasso, high-dimensional shape-constrained estimation–though practical for n, p on the order of thousands–must be faster before it can be useful.

One important open question is whether a fast algorithm exists for univariate convex regression, like the pool-adjacent violators algorithm (PAVA) for monotone regression. PAVA has a $O(n)$ runtime because, for monotone function fitting, the data can be segmented in a greedy way. It is unclear whether a $O(n)$ algorithm exists for fitting a convex function.

With the exception of the linear runtime of PAVA, it seems impossible for non-parametric models to have computational efficiency on par with parametric models. Additive convex models for instance have $O(np)$ variables in the optimization vs. $O(p)$ variabels for linear models. Yet, there is hope because the fitted functions in shape-constrained estimation are often simple: in case of convex regression, the fitted function is piece-wise linear often with few number of pieces. This is a form of sparsity. And just as lasso algorithms are faster if the fitted parameter vector is sparse, it is an interesting to ask whether optimization algorithms for shape-constrained estimation can be made faster if the fitted function is simple.

BIBLIOGRAPHY

Aleksandorov, A. (1939). Almost everywhere existence of the second differential of a convex function and some properties of convex functions. *Leningrad Univ. Ann.*, 37:3–35.

Arnold, T. B., Tibshirani, R., and Arnold, M. T. (2014). Package glmgen.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W., Silverman, E., et al. (1955). An empirical distribution function for sampling with incomplete information. *The annals of mathematical statistics*, 26(4):641–647.

Barlow, R. E., Bartholomew, D., Bremner, J., and Brunk, H. (1972). *Statistical Inference under Order Restrictions*. John Wiley & Sons, New York.

Bernasco, W. and Block, R. (2009). Where offenders choose to attack: A discrete choice model of robberies in Chicago. *Criminology*, 47(1):93–130.

Bertin, K. and Lecué, G. (2008). Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electronic Journal of Statistics*, 2:1224–1241.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Cai, T. T. and Low, M. G. (2011). A framework for estimation of convex functions. Technical report, Technical report.

Chen, H. and Yao, D. D. (2001). *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. Springer-Verlag.

Chen, Y. and Samworth, R. J. (2014). Generalised additive and index models with shape constraints. *arXiv preprint arXiv:1404.2957*.

Chu, W. and Ghahramani, Z. (2005). Preference learning with gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*, pages 137–144. ACM.

Comminges, L. and Dalalyan, A. S. (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *The Annals of Statistics*, 40(5):2667–2696.

Cule, M., Samworth, R., et al. (2010a). Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electronic Journal of Statistics*, 4:254–270.

Cule, M., Samworth, R., and Stewart, M. (2010b). Maximum likelihood estimation of a multi-dimensional log-concave density. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(5):545–607.

Cule, M., Samworth, R., and Stewart, M. (2010c). Maximum likelihood estimation of a multi-dimensional log-concave density (with discussion). *J. R. Stat. Soc. Ser. B, Stat. Methodol.*, 72:545–600.

DeVore, R., Petrova, G., and Wojtaszczyk, P. (2011). Approximation of functions of few variables in high dimensions. *Constructive Approximation*, 33:125–143.

Dykstra, R. L. (1981). An isotonic regression algorithm. Technical report, DTIC Document.

Fang, Z. and Meinshausen, N. (2012). Lasso isotone for high-dimensional additive isotonic regression. *Journal of Computational and Graphical Statistics*, 21(1):72–91.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Fürnkranz, J. and Hüllermeier, E. (2010). *Preference learning*. Springer.

Goldenshluger, A. and Zeevi, A. (2006). Recovering convex boundaries from blurred and noisy observations. *Ann. Statist*, 34:1375–1394.

Grenander, U. (1956). On the theory of mortality measurement: part ii. *Scandinavian Actuarial Journal*, 1956(2):125–153.

Groeneboom, P., Jongbloed, G., and Wellner, J. A. (2001). Estimation of a convex function: characterizations and asymptotic theory. *Annals of Statistics*, pages 1653–1698.

Guntuboyina, A. and Sen, B. (2013a). Global risk bounds and adaptation in univariate convex regression. *arXiv:1305.1648*.

Guntuboyina, A. and Sen, B. (2013b). Global risk bounds and adaptation in univariate convex regression. *arXiv:1305.1648*.

Hannah, L. A. and Dunson, D. B. (2012). Ensemble methods for convex regression with applications to geometric programming based circuit design. In *International Conference on Machine Learning (ICML)*.

Hanson, D. and Pledger, G. (1976). Consistency in concave regression. *The Annals of Statistics*, pages 1038–1050.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., and Tibshirani, R. (2009). *The elements of statistical learning*, volume 2. Springer.

Hildreth, C. (1954). Point estimates of ordinates of concave functions. *Journal of the American Statistical Association*, 49(267):598–619.

Horn, R. and Johnson, C. (1990). *Matrix Analysis*. Cambridge University Press; Reprint edition.

Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of statistics*, 38(4):2282.

Kim, A. K. and Samworth, R. J. (2014). Global rates of convergence in log-concave density estimation. *arXiv preprint arXiv:1404.2298*.

Koltchinskii, V. and Yuan, M. (2010). Sparsity in multiple kernel learning. *The Annals of Statistics*, 38(6):3660–3695.

Lafferty, J. and Wasserman, L. (2008). Rodeo: sparse, greedy nonparametric regression. *The Annals of Statistics*, 36(1):28–63.

Lele, A. S., Kulkarni, S. R., and Willsky, A. S. (1992). Convex-polygon estimation from support-line measurements and applications to target reconstruction from laser-radar data. *Journal of the Optical Society of America, Series A*, 9:1693–1714.

Lim, E. and Glynn, P. W. (2012). Consistency of multidimensional convex regression. *Operations Research*, 60(1):196–208.

Liu, H. and Chen, X. (2009). Nonparametric greedy algorithm for the sparse learning problems. In *Advances in Neural Information Processing Systems*.

Mair, P., Hornik, K., and de Leeuw, J. (2009). Isotone optimization in r: pool-adjacent-violators algorithm (pava) and active set methods. *Journal of statistical software*, 32(5):1–24.

Mammen, E. (1991). Nonparametric regression under qualitative smoothness assumptions. *The Annals of Statistics*, pages 741–759.

Mammen, E. and Yu, K. (2007). Additive isotone regression. *Lecture Notes-Monograph Series*, pages 179–195.

Matzkin, R. L. (1991). Semiparametric estimation of monotone and concave utility functions for polychotomous choice models. *Econometrica: Journal of the Econometric Society*, pages 1315–1327.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior.

McFadden, D. et al. (1978). *Modelling the choice of residential location*. Institute of Transportation Studies, University of California.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462.

Meyer, R. F. and Pratt, J. W. (1968). The consistent assessment and fairing of preference functions. *IEEE Trans. Systems Sci. Cybernetics*, 4(3):270–278.

Mossel, E., O'Donnell, R., and Servedio, R. (2004). Learning functions of k relevant variables. *Journal of Computer and System Sciences*, 69(3):421–434.

Nechyba, T. J. and Strauss, R. P. (1998). Community choice and local public services: A discrete choice approach. *Regional Science and Urban Economics*, 28(1):51–73.

Ortuzar, J. d. and Willumsen, L. G. (1994). *Modelling transport*.

Prince, J. L. and Willsky, A. S. (1990). Reconstructing convex sets from support line measurements. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:377–389.

Pya, N. and Wood, S. N. (2014). Shape constrained additive models. *Statistics and Computing*, pages 1–17.

Ramdas, A. and Tibshirani, R. J. (2014). Fast and flexible admm algorithms for trend filtering. *arXiv preprint arXiv:1406.2082*.

Raskutti, G., Wainwright, M. J., and Yu, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.*, 13:389–427.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society, Series B, Methodological*, 71(5):1009–1030.

Ravikumar, P., Liu, H., Lafferty, J., and Wasserman, L. (2007). Spam: Sparse additive models. In *Advances in Neural Information Processing Systems*.

Seijo, E. and Sen, B. (2011). Nonparametric least squares estimation of a multivariate convex regression function. *The Annals of Statistics*, 39(3):1633–1657.

Serfling, R. J. (1974). Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, 2(1):39–48.

Shah, A. K. and Oppenheimer, D. M. (2008). Heuristics made easy: an effort-reduction framework. *Psychological bulletin*, 134(2):207.

Tibshirani, R. J. et al. (2014). Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323.

Van der Vaart, A. and Wellner, J. (1996). Weak convergence and empirical processes.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. arXiv:1011.3027.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *Information Theory, IEEE Transactions on*, 55(5):2183–2202.

Yuille, A. L. and Rangarajan, A. (2003). The concave-convex procedure. *Neural computation*, 15(4):915–936.