DISSERTATION

PENALIZED ISOTONIC REGRESSION AND AN APPLICATION IN SURVEY SAMPLING

Submitted by

Jiwen Wu

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2016

Doctoral Committee:

Advisors: Jean D. Opsomer
Co-Advisor: Mary C. Meyer

F. Jay Breidt
Paul Doherty Jr

ABSTRACT

PENALIZED ISOTONIC REGRESSION AND AN APPLICATION IN SURVEY SAMPLING

In isotonic regression, the mean function is assumed to be monotone increasing (or decreasing) but otherwise unspecified. The classical isotonic least-squares estimator is known to be inconsistent at boundaries; this is called the spiking problem. A penalty on the range of the regression function is proposed to correct the spiking problem for univariate and multivariate isotonic models. The penalized estimator is shown to be consistent everywhere for a wide range of sizes of the penalty parameter. For the univariate case, the optimal penalty is shown to depend on the derivatives of the true regression function at the boundaries. Pointwise confidence intervals are constructed using the penalized estimator and bootstrapping ideas; these are shown through simulations to behave well in moderate sized samples. Simulation studies also show that the power of the hypothesis test of constant versus increasing regression function improves substantially compared to the power of the test with unpenalized alternative, and also compares favorably to tests using parametric alternatives.

The application of isotonic regression is also considered in the survey context where many variables contain natural orderings that should be respected in the estimates. For instance, the National Compensation Survey estimates mean wages for many job categories, and these mean wages are expected to be non-decreasing according to job level. In this type of situation, isotonic regression can be applied to give constrained estimators satisfying the monotonicity. We combine domain estimation and the pooled adjacent violators algorithm to construct new design-weighted constrained estimators. The resulting estimator is the classical design-based domain estimator but after adaptive pooling of neighboring domains, so that it is both readily implemented in large-scale surveys and easy to explain to data users. Under mild conditions on the sampling design and the population, the estimators are shown to be

design consistent and asymptotically normal. Confidence intervals for domain means using linearization-based and replication-based variance estimation show marked improvements compared to survey estimators that do not incorporate the constraints.

Furthermore, a cone projection algorithm is implemented in the domain mean estimate to accommodate qualitative constraints in the case of two covariates. Theoretical properties of the constrained estimators have been investigated and a simulation study is used to demonstrate the improvement of confidence interval when using the constrained estimate. We also provide a relaxed monotone constraint to loosen the qualitative assumptions, where the extent of departure from monotonicity can be controlled by a weight function and a chosen bandwidth. We compare the unconstrained estimate, constrained estimate without penalty, constrained estimate with penalty, and the relax constrained estimate. Improvements are found in the confidence interval with higher coverage rates and smaller confidence size when incorporating the constraints, and the penalized version fixes the spiking problem at the boundary.

ACKNOWLEDGEMENTS

I would first like to thank my advisors Dr. Jean Opsomer and Dr. Mary Meyer for their generous support in my research work and valuable mentoring throughout these past few years. They provided me great help whenever I got stuck and encouraged me to overcome the difficulties. I would also like to thank Dr. Jay Breidt and Dr. Paul Doherty Jr for being my committee members and for their research suggestions, comments and guidance. I am grateful to the Department of Statistics of Colorado State University for providing me great statistic courses and valuable experience of teaching and consulting. I appreciate all the faculty, the staff and my fellow graduate students for their generous help.

Additionally, I want to express my gratitude to my parents and my husband. They have been very supportive throughout my Ph.D. study and always stand by me when I encounter any difficulties. Also, I would like to acknowledge my little boy, Lucas, who gives me motivations to work hard and cheers me up with his smile and laughter.

This research was supported by the American Statistical Association/National Science Foundation/Bureau of Labor Statistics Fellowship Program and National Science Foundation MMS-1533804.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND BACKGROUND

1.1 Isotonic Regression

Consider the problem of fitting a function μ to data (x_i, y_i), where the distinct x-coordinates are ordered as $0 \leq x_1 < \cdots < x_n \leq 1$. The model is

$$y_i = \mu(x_i) + \varepsilon_i, \qquad i = 1, 2, \cdots, n$$

and ε_i's are independent normal errors with mean zero and variance σ^2. Assume μ is isotonic on the x-coordinates, namely, $\mu(x_i) \leq \mu(x_j)$ whenever $x_i < x_j$.

Brunk (1958) proposed the least squared estimator of $\mu(x_i)$ as

$$\hat{\mu}_i = \hat{\mu}(x_i) = \max_{s \leq i} \min_{t \geq i} \frac{y_s + \cdots + y_t}{t - s + 1}.$$

Barlow et al. (1972) and Robertson et al. (1988) reviewed different algorithms for computations of the isotonic estimators. For a simple ordering, the pooled adjacent algorithm (PAVA) is efficient to find $\hat{\mu}_i$. Hanson et al. (1973) proved the consistency of the estimators except at the end points 0 and 1. If we look at the estimator at the first coordinate, where

$$\hat{\mu}_1 \;=\; \min \left\{ y_1, \frac{y_1 + y_2}{2}, \cdots, \frac{y_1 + y_2 + \cdots + y_{n-1}}{n-1}, \frac{y_1 + \cdots + y_n}{n} \right\}$$

it can be shown that $\mathrm{P}(\hat{\mu}_1 < \mu(x_1) - \sigma) \geq \mathrm{P}(y_1 < \mu(x_1) - \sigma) = .1587$ under normality assumption even as n increases without bound. The estimator at the lower end point tends to be too small, while similarly the estimator of the upper end point will be too large. It is

called spiking problem, and it actually gets worse when the sample size n increases. More discussion of the bias issues can be found in Sampson et al. (2003).

The isotonic regression model with two covariates is

$$y_i = \mu(\boldsymbol{x}_i) + \varepsilon_i,$$

where $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ for $i = 1, 2, \cdots, n$. Let $\mathcal{X}$ be the set of points $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ in $\mathbb{R}^2$. The partial orderings of $\boldsymbol{x}_i$ are reflexive, transitive, anti-symmetric and defined as $\boldsymbol{x}_i \preceq \boldsymbol{x}_j$. The only *a priori* assumption of μ is isotonic with respect to the partial ordering such that $\mu(\boldsymbol{x}_i) \leq \mu(\boldsymbol{x}_j)$ if $\boldsymbol{x}_i \preceq \boldsymbol{x}_j$. Brunk (1955) presented a max-min formula based on lower and upper sets, and the resulting estimator is

$$\begin{aligned} \hat{\mu}(\boldsymbol{x}) &= \min_{L:\boldsymbol{x}\in L} \max_{U:\boldsymbol{x}\in U} Av_y(L \cap U) \\ &= \max_{U:\boldsymbol{x}\in U} \min_{L:\boldsymbol{x}\in L} Av_y(L \cap U), \end{aligned}$$

where a set L is called a lower set if $\boldsymbol{x} \in L, \boldsymbol{z} \in \mathcal{X}$, then $\boldsymbol{z} \preceq \boldsymbol{x}$ implies $\boldsymbol{z} \in L$, and an upper set U can be defined in a dual manner. Av_y is the weighted average function of y. For a large $m \times n$ grids, the number of lower sets is $\binom{m+n}{m}$, which is too large to compute even for the minimum point of the grids. PAVA does not work in this partial ordering and a cone projection algorithm (Meyer, 1999) will be applied to solve the problem with computational efficiency and inference implementation.

The spiking problem also exists in the case of two covariates. The estimator at the smallest cell of the grid is $\min_{L:L\in\mathcal{L}} Av_y(L)$ with $\mathcal{L}$ being the collection of all lower sets, and it is biased small regardless of increasing of the sample size. In this dissertation, one of our goals is to fix the spiking problem at the boundaries of isotonic regression in both univariate case and the two covariate case. We add a penalty on the range to improve the estimation at both ends. In Chapter 2, we first give the penalized isotonic regression estimator and

prove the consistency at end points, second, the optimal rate of the penalty parameter is discussed, third, simulation study shows the improvement of confidence interval estimate and the power of hypothesis test of constant versus increasing regression function. A shorter version of Chapter 2 is published as Wu et al. (2015).

For univariate case of isotonic regression, PAVA is easy to implement and efficient to find the solution. However, in the case of two covariates, the max-min formula involves searching through a large number of lower and upper sets and several algorithms have been developed to find the least squared estimator with more efficiency. A simple iterative technique is proposed by Dykstra and Robertson (1982) as applying any algorithm of univariate isotonic regression on rows and columns of a grid alternatively. Block et al. (1994) provided the isotonic block class with recursion method (IBCR) for partially ordered isotonic regression and the isotonic block class with stratification (IBCS) algorithm for matrix-ordered isotonic regression. Qian and Eddy (1996) showed a "sandwich isotonic block class" (SIBC) algorithm to solve bivariate isotonic regression on ordered rectangular grids. Chepoi et al. (1997) presented new polynomial algorithms for some types of isotonic regression problem. Spouge et al. (2003) solved the isotonic regression as a network flow problem by searching a special partition of the partially ordered covariates. Burdakov et al. (2006) gave a generalized pooled adjacent algorithm to handle isotonic regression with large sample size and two or more covariates. In this dissertation, we apply the cone projection algorithm from Meyer (1999) and Meyer (2013), which gives the unique exact solution with computational efficiency. This algorithm is easy to extend to higher dimension and the penalized isotonic regression. A review of the cone projection algorithm is given in 4.1. The code for cone projection algorithm is available in the R package `coneproj` and the penalized isotonic regression for one and two-dimension can be found in the R package `isotonic.pen`.

1.2 Survey Estimators with Constraints

The monotonicity assumption may arise naturally in some survey data. For example, one of the goals of the National Compensation Survey is to estimate mean wage by location, job type and job level, where the mean wage can be expected to be increasing with respect to job level. The 'cell' defined by location, job type and job level can be treated as a domain in survey context, and isotonic regression will be implemented to create adaptive domain estimate satisfying qualitative constraints.

Following the typical framework as given by Särndal et al. (1992), a finite population is denoted as $U_N = \{1, \ldots, k, \ldots, N\}$ and it is partitioned by T domains. Let $U_{t,N}$ denote a domain with N_t elements for $t = 1, 2, \ldots, T$. Assume the domains $U_{t,N}$ are disjoint and $\sum_{t=1}^{T} N_t = N$. To simplify the notation, domains are denoted as U_t with omission of N. The indicator variable for domains is defined as

$$z_{tk} = \begin{cases} 1 & \text{for} \quad k \in U_t \\ 0 & \text{for} \quad k \notin U_t \end{cases}$$

then the population mean of a study variable y in domain U_t is

$$\bar{y}_{U_t} = \frac{\sum_{k \in U_t} y_k}{N_t} = \frac{\sum_{k \in U_N} z_{tk} y_k}{\sum_{k \in U_N} z_{tk}}.$$

Given a sampling design $p_N(\cdot)$, a probability sample s is drawn from U_N and $p_N(s)$ is the probability of drawing the sample s. The sample size is denoted by n_N. Let $\pi_{iN} = \mathrm{P}(i \in s) = \sum_{s: i \in s} p_N(s) > 0$ and $\pi_{ijN} = \mathrm{P}(i, j \in s) = \sum_{s: i,j \in s} p_N(s) > 0$ for all $i, j \in U_N$. For simplicity of notation, we will eliminate the subscript N and write as π_i and π_{ij} instead.

Denote the sample indicator

$$I_k = \begin{cases} 1 & \text{for} \quad k \in s \\ 0 & \text{for} \quad k \notin s \end{cases}$$

then $E_p(I_k) = E_p(I_k^2) = \pi_k$, $Var_p(I_k) = \pi_k(1-\pi_k)$, and $Cov_p(I_k, I_l) = \Delta_{kl} = \pi_{kl} - \pi_k\pi_l$ where E_p, Var_p, and Cov_p are the expectation, variance, and covariance with respect to the sampling design $p_N(\cdot)$.

Let $s_t = U_t \bigcap s$ and assume the sample contains at least one observation from each domain. Without any constraints, the Horvitz-Thompson estimator

$$\bar{y}_{s_t} = \frac{\sum_{k \in s_t} y_k/\pi_k}{N_t} = \frac{\sum_{k \in U_N} z_{tk} y_k I_k/\pi_k}{\sum_{k \in U_N} z_{tk}}$$

is unbiased for $\bar{y}_{U_t}$ and has variance

$$V(\bar{y}_{s_t}) = \frac{1}{N_t^2} \sum \sum_{k,l \in U_t} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

and variance estimator

$$\hat{V}(\bar{y}_{s_t}) = \frac{1}{N_t^2} \sum \sum_{k,l \in s_t} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}.$$

If N_t is unknown, then the Hájek estimator gives an approximately unbiased estimator of the domain mean $\bar{y}_{U_t}$ as

$$\tilde{y}_{s_t} = \frac{\sum_{s_t} y_k/\pi_k}{\sum_{s_t} 1/\pi_k} = \frac{\sum_U z_{tk} y_k I_k/\pi_k}{\sum_U z_{tk} I_k/\pi_k}.$$

The variance of the linearized approximation to $\tilde{y}_{s_t}$ (Särndal et al., 1992, Chapter 5.5) is

$$AV(\tilde{y}_{s_t}) = \frac{1}{N_t^2} \sum \sum_{U_t} \Delta_{kl} \left(\frac{y_k - \bar{y}_{U_t}}{\pi_k} \right) \left(\frac{y_l - \bar{y}_{U_t}}{\pi_l} \right).$$

A variance estimator we will consider is given by

$$\hat{V}(\tilde{y}_{s_t}) = \frac{1}{\hat{N}_t^2} \sum \sum_{s_t} \frac{\Delta_{kl}}{\pi_{kl}} \left(\frac{y_k - \tilde{y}_{s_t}}{\pi_k} \right) \left(\frac{y_l - \tilde{y}_{s_t}}{\pi_l} \right)$$

where $\hat{N}_t = \sum_{s_t} 1/\pi_k$.

To take advantage of the monotonicity assumption, isotonic regression can be incorporated in the Horvitz-Thompson estimator or the Hájek estimator. Our focus will be the Hájek estimator since it does not require the known values of N_t and it tends to be more efficient than the Horvitz-Thompson estimator in practice. In the case of a single covariate, PAVA is applied to construct the constrained domain estimates by adaptively collapsing neighboring domains. Chapter 3 presents the theoretical properties of the PAVA-based constrained estimate and simulation study of its confidence interval using linearization-based and replication-based variance estimation. A shorter version of Chapter 3 was submitted to Canadian Journal of Statistics. In Chapter 4, we provide a more general application of isotonic regression in the case of two covariates by incorporating cone projection algorithm (Meyer, 1999) in the domain mean estimate. PAVA is limited to the univariate case, while cone projection algorithm can deal with complete ordering, partial ordering, and a relax ordering we proposed in Chapter 4. We show the design consistency and asymptotic normality of the constrained domain estimator and use simulation results to present the improvements of confidence interval estimate.

CHAPTER 2

PENALIZED ISOTONIC REGRESSION

2.1 Introduction

We consider the isotonic regression model

$$y_i = \mu(\boldsymbol{x}_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

where ε_i are iid random errors with mean zero and finite variance σ^2, and $\mathcal{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\} \subseteq \mathbb{R}^d$, for $d \geq 1$. A partial ordering is defined as $\boldsymbol{x}_i \preceq \boldsymbol{x}_j$ if $\boldsymbol{x}_i \leq \boldsymbol{x}_j$ coordinate-wise. We assume without loss of generality that the $\boldsymbol{x}_i$ points are distinct; otherwise the y values may be averaged at each distinct $\boldsymbol{x}_i$ and a weighted regression performed. The only *a priori* assumption for the regression function is that μ is isotonic with respect to the partial ordering; i.e., $\mu(\boldsymbol{x}_i) \leq \mu(\boldsymbol{x}_j)$ if $\boldsymbol{x}_i \preceq \boldsymbol{x}_j$. Brunk (1955) presented a max-min formula for the least-squares estimator of μ based on lower and upper sets, where L is a lower set if $\boldsymbol{x} \in L, \boldsymbol{z} \in \mathcal{X}, \boldsymbol{z} \preceq \boldsymbol{x}$ implies $\boldsymbol{z} \in L$, and a set U is an upper set if $\boldsymbol{x} \in U, \boldsymbol{z} \in \mathcal{X}, \boldsymbol{x} \preceq \boldsymbol{z}$ implies $\boldsymbol{z} \in U$. Define $Av_y(A)$ to be the mean of y_i for $\boldsymbol{x}_i \in A$; then

$$\hat{\mu}(\boldsymbol{x}) = \max_{U:\boldsymbol{x}\in U} \min_{L:\boldsymbol{x}\in L} Av_y(L \cap U). \tag{2}$$

See also Brunk (1958), Barlow et al. (1972), and Robertson et al. (1988, Chapter 1). Hanson et al. (1973) proved the consistency of the estimators except at the extreme large or small design points. Some bias issues of the isotonic estimators were discussed in Sampson et al. (2003).

For $d = 1$, the ordering is complete, the max-min formula becomes

$$\hat{\mu}_i = \hat{\mu}(x_i) = \max_{s \le i} \min_{t \ge i} \frac{y_s + \cdots + y_t}{t - s + 1}, \tag{3}$$

and the pooled adjacent algorithm (PAVA) is an efficient way to compute $\hat{\mu}_i$. For $d = 2$ and $\boldsymbol{x}$-values arranged in $m \times n$ grids, the number of lower sets is $(m+n)!/[m!n!]$, which is computationally intensive even for the estimation at the minimum point of the grid. Direct application of (2) is not practical except for quite small sample sizes. Several algorithms have been developed to find the least squares estimator for bivariate isotonic regression. Dykstra and Robertson (1982) wanted to avoid searching through a large number of lower sets, so they proposed an iterative technique: apply PAVA alternatively on the two independent variables in the model. Block et al. (1994) generalized the pooled adjacent violators algorithm and provided the "isotonic block class" algorithm (IBC) for partially ordered isotonic regression. Subsequently, Qian and Eddy (1996) improved the algorithm to a "sandwich isotonic block class" algorithm which is more efficient on ordered rectangular grids. Spouge et al. (2003) developed an algorithm to solve the problem by searching a special partition of the partially ordered covariates and reduced the computation time to $O(n^3)$. Burdakov et al. (2006) gave a fast algorithm to approximate the estimation for very large sample size.

For the examples and simulations in this article, the `coneA` function in the CRAN package `coneproj` is used to compute the multiple isotonic regression. The function uses the algorithm of Meyer (2013), where a constraint matrix $\boldsymbol{A}$ is constructed so that each comparable pair $\boldsymbol{x}_i \preceq \boldsymbol{x}_j$ corresponds to a row of $\boldsymbol{A}$ that is all zeros but -1 in the ith column and $+1$ in the jth column; then redundant rows are removed so that $\boldsymbol{A}$ is "irreducible" as defined in Meyer (1999).

In this chapter, we address a long-standing issue in penalized regression, which is often referred to as the 'spiking' problem. Suppose $\boldsymbol{x}_1$ is a minimal point in $\mathcal{X}$, so that $\boldsymbol{x}_i \not\preceq \boldsymbol{x}_1$ for all $i \neq 1$. The estimator at $\boldsymbol{x}_1$ is too small: $\mathrm{P}\left(\hat{\mu}_1 < \mu(\boldsymbol{x}_1) - \sigma\right) \geq \mathrm{P}\left(y_1 < \mu(\boldsymbol{x}_1) - \sigma\right) = .1587$

if the errors are normal, even as n increases without bound. Similarly, the estimator at a maximal point will be too large. This "spiking problem" actually gets worse when the sample size n increases. Meyer (2006) proposed a modified version of monotone ($d = 1$) and convex regression estimators to fix the inconsistency at the endpoints, and the modified estimators will increase the power of likelihood ratio tests of constant vs. monotone regression or linear vs. convex regression. Pal (2008) used a penalty to adjust the estimator at the lower end point for $0 \leq x_1 < \cdots < x_n \leq 1$; he derived the limit distribution for the residual sum of squares and constructed confidence intervals when $\mu'(0)$ is positive.

Here, we instead investigate adding a penalty on the range of $\boldsymbol{\mu}$ to fix the spiking problem at both edges at the same time. We will show theoretically that this solution solves the spiking problem for both $d = 1$ and $d = 2$ cases. We give the optimal penalty parameter for $d = 1$, which depends on the derivatives of the true regression function at the boundaries. Examples of the improvement of the penalized estimator over the original are shown in Figures 2.1 and 2.2 for univariate and bivariate isotonic regression, respectively. Through simulation experiments, we provide guidance on how to select the penalty in practice. We also propose new inference tools for the penalized estimator, applying a parametric bootstrap approach for construction of pointwise confidence intervals and for hypothesis testing, for both univariate and bivariate regression. The spiking problem is also reflected in the confidence intervals near end points even when the sample size is increasing, while the penalty can mostly correct it and adjust the interval estimation. Simulations show that the power will be improved for the likelihood ratio test for constant versus monotone regression given different underlying models and sample sizes.

2.2 Penalized Isotonic regression: $d = 1$ case

Let $\hat{\boldsymbol{\mu}}$ be the original least squares estimator minimizing $\|\boldsymbol{y}-\boldsymbol{\mu}\|^2$ over isotonic $\boldsymbol{\mu}$. Adding a penalty $\lambda > 0$ on the range, we want to minimize $\|\boldsymbol{y}-\boldsymbol{\mu}\|^2+\lambda(\mu_n-\mu_1) = \|\boldsymbol{y}-\boldsymbol{\mu}\|^2+\lambda\mathbf{b}^T\boldsymbol{\mu}$ over isotonic $\boldsymbol{\mu}$, where $\mathbf{b} = (-1, 0, \ldots, 0, 1)^T$. This is equivalent to the unpenalized isotonic

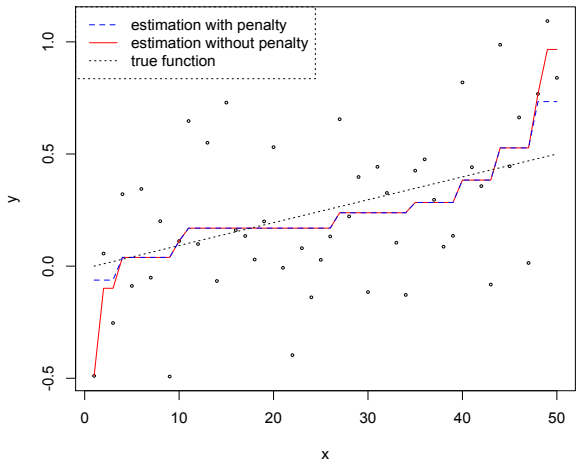

Figure 2.1: Penalized and unpenalized univariate isotonic regression estimator for $n = 50$ and penalty parameter $\lambda = .5$.

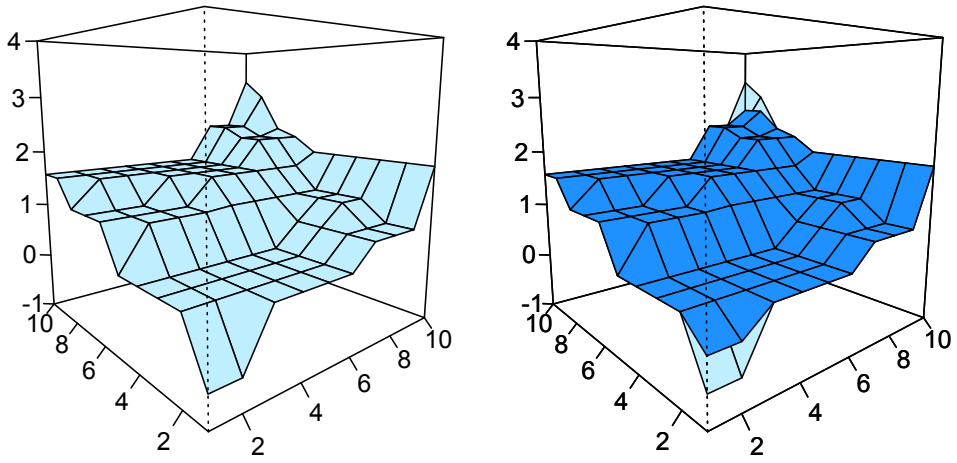

Figure 2.2: Penalized and unpenalized bivariate isotonic regression on a 10×10 grid. The lighter surface is the original estimator while the darker surface is the penalized estimator with penalty parameter $\lambda = 1$.

regression with data $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \lambda \mathbf{b}/2 = (y_1 + \lambda/2, y_2, \ldots, y_{n-1}, y_n - \lambda/2)$. The penalized estimator is denoted by $\tilde{\boldsymbol{\mu}}$, and from (3), the estimator at the left end point is

$$\tilde{\mu}_1 \;=\; \min\left\{ y_1 + \frac{\lambda}{2}, \cdots, \frac{y_1 + y_2 + \cdots + y_{n-1}}{n-1} + \frac{\lambda}{2(n-1)}, \frac{y_1 + \cdots + y_n}{n} \right\}$$

$$= \min\left\{y_1 + \frac{\lambda}{2}, \bar{y}_2 + \frac{\lambda}{4}, \cdots, \bar{y}_{n-1} + \frac{\lambda}{2(n-1)}, \bar{y}_n\right\},$$

where $\bar{y}_k = (y_1 + \cdots + y_k)/k$.

Intuitively, the penalty "pulls up" the estimator at the left, and similarly "pulls down" the estimator on the right. Consistency requires that $\lambda = \lambda_n \to \infty$ as $n \to \infty$. For given $\delta > 0$, $\mathrm{P}(\tilde{\mu}_1 < \mu(0) - \delta) \geq \mathrm{P}(y_1 + \lambda/2 < \mu(0) - \delta) = \mathrm{P}(y_1 < \mu(0) - \delta - \lambda/2)$, so if the support for y_1 is the real line, consistency requires that the penalty parameter λ increase without bound.

2.2.1 Consistency

We assume that as n increases, the values of x_i follow a distribution that is continuous on $[0, 1]$. Without loss of generality, we assume that the x_i are equally spaced on $[0, 1]$; otherwise there is a transformation to equally spaced x that preserves monotonicity. In Theorem 1, we show the consistency of the penalized estimator at the boundaries if the regression function is continuous in a small neighborhood of the end point, and if λ is proportional to n^α for $0 < \alpha < 1$. Following the idea of Hanson et al. (1973), Theorem 2 proves the consistency of $\hat{\mu}$ at points in $(0, 1)$ where μ is continuous. The following two lemmas are given first and they will be used in the proof of Theorems. The expression $a \asymp b$ is used to denote that a/b is bounded away from zero and infinity.

Lemma 1. *Let ε_i be independent and identically distributed random variables with mean 0 and finite variance σ^2, and define $S_k = \varepsilon_1 + \cdots + \varepsilon_k$. For $c > 0$,*

$$\lim_{m\to\infty} \mathrm{P}\left(\frac{\max_{k=1,\ldots,m} S_k}{\sigma\sqrt{m}} \leq c\right) = \sqrt{\frac{2}{\pi}} \int_0^c e^{-\frac{u^2}{2}} du. \tag{4}$$

Furthermore, if c_m is a sequence of positive real numbers converging to zero, then

$$\lim_{m\to\infty} \mathrm{P}\left(\frac{\max_{k=1,\ldots,m} S_k}{\sigma\sqrt{m}} \leq c_m\right) = 0. \tag{5}$$

and if c_m is a sequence of positive real numbers diverging to infinity, then

$$\lim_{m\to\infty} \mathrm{P}\left(\frac{\max_{k=1,\ldots,m} S_k}{\sigma\sqrt{m}} \leq c_m\right) = 1, \tag{6}$$

Proof. The result in (4) is Corollary 6.5.6 from Resnick (1992, p.501). When c_m is a sequence of positive real numbers converging to zero, then for any $\delta > 0$, there exists a positive integer N_δ such that $c_m \leq \delta$ when $m > N_\delta$. By (4), for all $\epsilon > 0$, there is a positive integer N_ϵ such that $\mathrm{P}(\max_{k=1,\ldots,m} S_k/(\sigma\sqrt{m}) \leq \delta) \leq \sqrt{2/\pi}\int_0^\delta e^{-u^2/2}du + \epsilon/2$ for all $m > N_\epsilon$. Choose δ satisfying $\sqrt{2/\pi}\int_0^\delta e^{-u^2/2}du \leq \epsilon/2$, then for all $m > \max\{N_\delta, N_\epsilon\}$,

$$\mathrm{P}\left(\frac{\max_{k=1,\ldots,m} S_k}{\sigma\sqrt{m}} \leq c_m\right) \leq \mathrm{P}\left(\frac{\max_{k=1,\ldots,m} S_k}{\sigma\sqrt{m}} \leq \delta\right) \leq \epsilon.$$

The proof for (6) is similar.

■

Lemma 2. *Following the conditions in Lemma 1 and suppose ε_i also has finite 3rd and 4th moments, then for all $\delta > 0$ and $\epsilon > 0$, there is a positive integer N such that*

$$\mathrm{P}\left(\bigcup_{k\geq N}\{S_k > k\delta\}\right) < \epsilon \tag{7}$$

Proof. Denote $E(\varepsilon_i^4) = M_4$, then $E(S_k^4) = kM_4 + 3k(k-1)\sigma^4$ and

$$\mathrm{P}\left(\bigcup_{k\geq N}\{S_k > k\delta\}\right) \leq \sum_{k=N}^{\infty} \mathrm{P}\left(\frac{S_k}{\sqrt{k}} > \sqrt{k}\delta\right) \leq \sum_{k=N}^{\infty} \frac{kM_4 + 3k(k-1)\sigma^4}{\delta^4 k^4}$$

which is less than ϵ for large enough N due to the convergent series.

■

Theorem 1. *Consider model (1) with $d = 1$ and the conditions in Lemma 2. If $\mu(x)$ is isotonic and continuous in a neighborhood of the left endpoint 0, then for any $\delta > 0$,*

$$\lim_{n\to\infty} \mathrm{P}(|\tilde{\mu}_1 - \mu(0)| > \delta) = 0 \tag{8}$$

if $\lambda \asymp n^\alpha$ where $0 < \alpha < 1$.

Proof. For simplicity we will assume that variance $\sigma^2 = 1$ in the following proof. Choose any $\delta > 0$; by continuity there is a $\xi > 0$ such that $x_i \le \xi$ implies $\mu(x_i) - \mu(0) < \delta/2$. Define $k' = \max\{i : x_i \le \xi\}$, so $k' \sim \xi n$, and

$$\begin{aligned}
\mathrm{P}\left(\tilde{\mu}_1 \ge \mu(0) + \delta\right) &= \mathrm{P}\left(\min\left\{y_1 + \frac{\lambda}{2}, \cdots, \bar{y}_{n-1} + \frac{\lambda}{2(n-1)}, \bar{y}_n\right\} \ge \mu(0) + \delta\right) \\
&\le \mathrm{P}\left(\bar{y}_{k'} + \frac{\lambda}{2k'} \ge \mu(0) + \delta\right) \\
&= \mathrm{P}\left(\bar{\varepsilon}_{k'} \ge \mu(0) - \bar{\mu}_{k'} + \delta - \frac{\lambda}{2k'}\right) \\
&\le \mathrm{P}\left(\bar{\varepsilon}_{k'} \ge \frac{\delta}{2} - \frac{\lambda}{2k'}\right),
\end{aligned} \tag{9}$$

where $\bar{\varepsilon}_k = (\varepsilon_1 + \cdots + \varepsilon_k)/k$ and $\bar{\mu}_k = (\mu_1 + \cdots + \mu_k)/k$. By the Law of Large Numbers, $\bar{\varepsilon}_{k'}$ converges in probability towards the mean 0. Since δ is positive and k' is proportional to n, the probability will go to 0 if $\lambda/(2k') \to 0$ as $n \to \infty$, that is $\lambda \asymp n^\alpha$ for $\alpha < 1$. Next, plug in $\lambda \asymp n^\alpha$, so that

$$\begin{aligned}
\mathrm{P}\left(\tilde{\mu}_1 \le \mu(0) - \delta\right) &= \mathrm{P}\left(\min\left\{y_1 + \frac{\lambda}{2}, \cdots, \bar{y}_{n-1} + \frac{\lambda}{2(n-1)}, \bar{y}_n\right\} \le \mu(0) - \delta\right) \\
&= \mathrm{P}\left(\bigcup_{k=1}^{n}\left\{S_k \ge k\delta + \frac{n^\alpha}{2} + k\left(\bar{\mu}_k - \mu(0)\right)\right\}\right) \\
&\le \mathrm{P}\left(\bigcup_{k=1}^{n}\left\{S_k \ge k\delta + \frac{n^\alpha}{2}\right\}\right) \\
&\le \mathrm{P}\left(\bigcup_{k=1}^{m}\left\{S_k \ge k\delta + \frac{n^\alpha}{2}\right\}\right) + \mathrm{P}\left(\bigcup_{k=m+1}^{n}\left\{S_k \ge k\delta + \frac{n^\alpha}{2}\right\}\right)
\end{aligned}$$

$$= \quad (I) + (II),$$

where m is the smallest integer larger than $n^{2\alpha-\xi}$ for some $\xi \in (0, 2\alpha)$.

For part (I),

$$\begin{aligned} \mathrm{P}\left(\bigcup_{k=1}^{m}\left\{S_k \geq k\delta + \frac{n^\alpha}{2}\right\}\right) &\leq \mathrm{P}\left(\bigcup_{k=1}^{m}\left\{S_k \geq \frac{n^\alpha}{2}\right\}\right) \\ &= \mathrm{P}\left(\max_{k=1,\ldots,m}\frac{S_k}{\sqrt{m}} \geq \frac{n^{\xi/2}}{2}\right) \\ &\to 0 \end{aligned}$$

by (6) in Lemma 1. For part (II), Lemma 2 is applied and

$$\begin{aligned} \mathrm{P}\left(\bigcup_{k=m+1}^{n}\left\{S_k \geq k\delta + \frac{n^\alpha}{2}\right\}\right) &\leq \mathrm{P}\left(\bigcup_{k=m+1}^{n}\{S_k \geq k\delta\}\right) \\ &\leq \mathrm{P}\left(\bigcup_{k=m+1}^{\infty}\{S_k \geq k\delta\}\right) \\ &\to 0. \end{aligned}$$

Thus, $\mathrm{P}\left(\tilde{\mu}_1 \leq \mu(0) - \delta\right) \to 0$ as $n \to \infty$ if $\lambda \asymp n^\alpha$ for any $\alpha > 0$.

■

Theorem 2. *Following the conditions in Theorem 1 and letting x_o be any point in $(0, 1)$ at which $\mu(x)$ is continuous, denote $\tilde{\mu}(x_o)$ as the penalized estimator at x_o. Then, given any $\delta > 0$*

$$\lim_{n\to\infty} \mathrm{P}(|\tilde{\mu}(x_o) - \mu(x_o)| > \delta) = 0.$$

Proof. The idea of Hanson et al. (1973, Theorem 1) is used to show the consistency, with the addition of the penalty term λ. Let x_s be the closest observation point to x_o, by continuity these exists ξ such that $x_i - x_s \leq \xi$ implies $\mu(x_i) - \mu(x_s) < \delta/2$. Let $u = u_n = \max\{i :$

$x_i - x_s \leq \xi\}$, then

$$\begin{aligned}
\tilde{\mu}(x_s) &= \max_{a\leq s}\min_{s\leq b}\frac{\tilde{y}_a+\cdots+\tilde{y}_b}{b-a+1} \\
&\leq \max_{a\leq s}\frac{\tilde{y}_a+\cdots+\tilde{y}_u}{u-a+1} \\
&= \max_{a\leq s}\left\{\frac{\sum_{i=a}^{u}[\tilde{y}_i-\mu(x_i)]}{u-a+1}+\frac{\sum_{i=a}^{u}\mu(x_i)}{u-a+1}\right\} \\
&\leq \max_{a\leq s}\frac{\sum_{i=a}^{u}[\tilde{y}_i-\mu(x_i)]}{u-a+1}+\mu(x_u) \\
&\leq \max_{a\leq s}\frac{\sum_{i=a}^{u}[\tilde{y}_i-\mu(x_i)]}{u-a+1}+\mu(x_s)+\frac{\delta}{2} \\
&= \max\left\{\frac{\sum_{i=1}^{u}\varepsilon_i}{u}+\frac{\lambda}{2u},\frac{\sum_{i=2}^{u}\varepsilon_i}{u-1},\ldots,\frac{\sum_{i=s}^{u}\varepsilon_i}{u-s+1}\right\}+\mu(x_s)+\frac{\delta}{2}.
\end{aligned}$$

By Lemma 2, for any $\zeta > 0$ there is a positive integer M such that

$$\mathrm{P}\left(\max_{a\leq s}\frac{\sum_{i=a}^{u}\varepsilon_i}{u-a+1}>\zeta\right)\leq\zeta,$$

for all $u \geq M$. Noting that $u - s \sim \xi n$, $\lambda \asymp n^{\alpha}$ for $\alpha \in (0, 1)$, then $\lambda/(2u) \to 0$ and

$$\mathrm{P}\left(\max\left\{\frac{\sum_{i=1}^{u}\varepsilon_i}{u}+\frac{\lambda}{2u},\frac{\sum_{i=2}^{u}\varepsilon_i}{u-1},\ldots,\frac{\sum_{i=s}^{u}\varepsilon_i}{u-s+1}\right\}>\zeta\right)\leq\zeta$$

for $u \geq M$ Thus, as $n \to \infty$

$$\mathrm{P}(\tilde{\mu}(x_s)-\mu(x_s)<\delta)\to 1.$$

The proof of $\mathrm{P}(\tilde{\mu}(x_s) - \mu(x_s) > -\delta) \to 1$ is similar.

■

2.2.2 Optimal Penalty Parameter

We have shown that the penalized estimator is consistent everywhere at points where the regression function μ is continuous in a neighborhood, for a wide range of rates on the

penalty parameter. In this section we show that the optimal rate of growth of the penalty parameter depends on the behavior of the regression function at the endpoint. Because of this, we allow for different penalty parameters at either end of the x-value range. In particular, we show that if the slope is positive at zero and the second derivative is bounded, then $\alpha = 1/3$ provides the optimal rate. If the slope is zero, but the second derivative is positive and the third derivative is bounded, then α should be $2/5$. In fact, if $\mu^{(k)}(0) = 0$ for $k = 1, \ldots, K-1$ but $\mu^{(K)}(0) > 0$, and $\mu^{(K+1)}$ exists and is bounded, then the optimal rate for λ is $n^{K/(2K+1)}$. In the case where μ is constant on $[0, \delta)$ for $\delta \in (0, 1]$ but $\mu(1-d) > \mu(0)$ for some $d \in (0, 1)$, we find that the penalty parameter should be proportional to $n^{1/2}$.

A key result that will be used to obtain these optimal rates is Lemma 1. It is used in the following three Theorems 3, 4, and 5, which show the results of finding the optimal rate of the penalty parameter for different settings for the behavior of regression function at the left end point. The results are analogous for the right end point.

Theorem 3. *Suppose $\mu'(0) = \delta_0 > 0$ and $|\mu''(x)| \leq M$ in $[0, \delta_1)$, for some $\delta_1 \in (0, 1]$ and $M < \infty$. Then the optimal penalty parameter is of order $n^{1/3}$. Specifically, fix any $\xi > 0$. If $\lambda \asymp n^{1/3-\xi}$, then $P(\tilde{\mu}_1 > \mu(0)) \to 0$, and if $\lambda \asymp n^{1/3+\xi}$, then $P(\tilde{\mu}_1 > \mu(0)) \to 1$.*

Proof. First, suppose $\lambda \asymp n^{1/3-\xi}$, for $\xi > 0$, and let m be the smallest integer larger than $n^{2/3-\xi}$, then

$$\begin{aligned}
\mathrm{P}(\tilde{\mu}_1 > \mu(0)) &\leq \mathrm{P}\left(\bigcap_{k=1}^{m} \left\{ \bar{y}_k + \frac{\lambda}{2k} > \mu(0) \right\}\right) \\
&= \mathrm{P}\left(\bigcap_{k=1}^{m} \left\{ S_k < \sum_{i=1}^{k} [\mu(x_i) - \mu(0)] + \frac{\lambda}{2} \right\}\right) \\
&\leq \mathrm{P}\left(\bigcap_{k=1}^{m} \left\{ S_k < m[\mu(x_m) - \mu(0)] + \frac{\lambda}{2} \right\}\right) \\
&= \mathrm{P}\left(\max_{k=1,\ldots,m} S_k < m[\mu(x_m) - \mu(0)] + \frac{\lambda}{2}\right)
\end{aligned}$$

$$= \quad \mathrm{P}\left(\frac{\max_{k=1,\ldots,m} S_k}{\sqrt{m}} < \sqrt{m}[\mu(x_m) - \mu(0)] + \frac{\lambda}{2\sqrt{m}} \right).$$

By Taylor's theorem and since $x_m = m/n$,

$$\delta_0 \frac{m}{n} - \frac{M}{2}\frac{m^2}{n^2} \leq \mu(x_m) - \mu(0) \leq \delta_0 \frac{m}{n} + \frac{M}{2}\frac{m^2}{n^2}, \tag{10}$$

when n is large enough so that $x_m < \delta_1$. By the choice of m and λ,

$$\sqrt{m}[\mu(x_m) - \mu(0)] + \frac{\lambda}{2\sqrt{m}} = O(n^{-\xi/2}).$$

Using (5), we have $\mathrm{P}(\tilde{\mu}_1 > \mu(0)) \to 0$.

Second, let $\lambda = an^{1/3+\xi}$, for $\xi > 0$ and $a > 0$, and let m be the smallest integer larger than $n^{2/3+\xi}$. Then

$$\begin{aligned}
\mathrm{P}(\tilde{\mu}_1 > \mu(0)) &= 1 - \mathrm{P}\left(\bigcup_{k=1}^{n} \left\{ S_k > \sum_{i=1}^{k} [\mu(x_i) - \mu(0)] + \frac{\lambda}{2} \right\} \right) \\
&\geq 1 - \mathrm{P}\left(\bigcup_{k=1}^{m} \left\{ S_k > \frac{\lambda}{2} \right\} \right) - \mathrm{P}\left(\bigcup_{k=m+1}^{n} \left\{ S_k > \sum_{i=1}^{k} [\mu(x_i) - \mu(0)] \right\} \right) \\
&= 1 - (I) - (II).
\end{aligned}$$

For part (I),

$$\mathrm{P}\left(\bigcup_{k=1}^{m} \left\{ S_k > \frac{an^{1/3+\xi}}{2} \right\} \right) = \mathrm{P}\left(\max_{k=1,\ldots,m} \frac{S_k}{\sqrt{m}} > an^{\xi/2} \right) \to 0$$

after applying (5) of Lemma 1.

For part (II), since $\sqrt{m}(\mu(x_m) - \mu(0)) > \delta_0 n^{3\xi/2}$ by (10) and the choice of m, then

$$\mathrm{P}\left(\bigcup_{k=m+1}^{n}\left\{S_k > \sum_{i=1}^{k}[\mu(x_i) - \mu(0)]\right\}\right) \leq \mathrm{P}\left(\bigcup_{k=m+1}^{n}\left\{\frac{S_k}{\sqrt{k}} > \sqrt{m}[\mu(x_m) - \mu(0)]\right\}\right) \to 0$$

by Lemma 2. Thus, $\mathrm{P}(\tilde{\mu}_1 > \mu(0)) \to 1$ if $\lambda \asymp n^{1/3+\xi}$.

■

Theorem 4. *Suppose $\mu'(0) = \cdots = \mu^{(K-1)} = 0$, $\mu^{(K)} = \delta_0 > 0$ and $|\mu^{(K+1)}(x)| \leq M$ in $[0, \delta_1)$, for some $\delta_1 \in (0, 1]$ and $M < \infty$. Then the optimal penalty parameter is of order $n^{K/(2K+1)}$. Specifically, fix any $\xi > 0$. If $\lambda \asymp n^{K/(2K+1)-\xi}$, then $P(\tilde{\mu}_1 > \mu(0)) \to 0$, and if $\lambda \asymp n^{K/(2K+1)+\xi}$, then $P(\tilde{\mu}_1 > \mu(0)) \to 1$.*

The proof is similar to the $K = 1$ case. In place of (10) we have

$$\delta_0 \left(\frac{m}{n}\right)^K - \frac{M}{2}\left(\frac{m}{n}\right)^{K+1} \leq \mu(x_m) - \mu(0) \leq \delta_0 \left(\frac{m}{n}\right)^K + \frac{M}{2}\left(\frac{m}{n}\right)^{K+1},$$

and when $\lambda \asymp n^{K/(2K+1)-\xi}$, choose $m = n^{2K/(2K+1)-\xi}$. □

Theorem 5. *Suppose $\mu(x)$ is constant on $[0, \delta_0)$, for some $\delta_0 \in (0, 1)$, but $\mu(1-d) > \mu(0)$ for some $d \in (0, 1)$. Then the optimal penalty parameter is of order $n^{1/2}$. Specifically, fix any $\xi > 0$. If $\lambda \asymp n^{1/2-\xi}$, then $P(\tilde{\mu}_1 > \mu(0)) \to 0$, and if $\lambda \asymp n^{1/2+\xi}$, then $P(\tilde{\mu}_1 > \mu(0)) \to 1$.*

Proof. Define

$$A_k = \left\{\bar{y}_k + \frac{\lambda}{2k} > \mu(0)\right\} = \left\{S_k < k(\mu(x_k) - \mu(0)) + \frac{\lambda}{2}\right\}, \text{ for } k = 1, \ldots, n-1$$

and $A_n = \{\bar{y}_n > \mu(0)\}$, so that $\{\tilde{\mu}_1 > \mu(0)\} = \cap_{k=1}^{n} A_k$. Let $m = \max\{i : x_i \leq \delta_0\}$, for equally spaced x_i, namely m is the largest integer smaller than $\delta_0 n$, then

$$\mathrm{P}(\tilde{\mu}_1 > \mu(0)) \;\leq\; \mathrm{P}\left(\max_{k=1,\ldots,m} S_k < m[\mu(x_m) - \mu(0)] + \frac{\lambda}{2}\right)$$

$$= \quad \mathrm{P}\left(\frac{\max_{k=1,\ldots,m} S_k}{\sqrt{m}} < \frac{\lambda}{2\sqrt{m}} \right).$$

As n gets large, the probability becomes 0 when $\lambda \asymp n^{1/2-\xi}$ for $\xi > 0$.

Also, if $\lambda \asymp n^{1/2+\xi}$,

$$\begin{aligned} \mathrm{P}(\tilde{\mu}_1 > \mu(0)) &= \mathrm{P}\left(\bigcap_{k=1}^{n} A_k \right) \\ &\geq 1 - \mathrm{P}\left(\bigcup_{k=1}^{n-1} A_k^c \right) - \mathrm{P}(A_n^c). \end{aligned}$$

Clearly $\mathrm{P}(A_n^c) \to 0$ as $n \to \infty$, and

$$\mathrm{P}\left(\bigcup_{k=1}^{n-1} A_k^c \right) \leq \mathrm{P}\left(\bigcup_{k=1}^{n-1} \left\{ S_k > \frac{\lambda}{2} \right\} \right) = \mathrm{P}\left(\frac{\max_{i=1,\ldots,m} S_k}{\sqrt{n-1}} > \frac{n^{\xi}}{2\sqrt{\delta_0}} \right) \to 0.$$

■

2.3 Penalized Isotonic regression on a grid

We now extend the results to the isotonic regression model with two covariates on a $n_1 \times n_2$ grid. In the isotonic regression model $y_{ij} = \mu(x_{1i}, x_{2j}) + \varepsilon_{ij}$, the random errors ε_{ij} are further assumed to be normally distributed. We suppose that the covariates x_{1i} and x_{2j} are equally spaced on $[0, 1]$ where $i = 1, 2, \ldots, n_1$ and $j = 1, 2, \ldots, n_2$, $n = n_1 \times n_2$ with $n_1 \to \infty$, $n_2 \to \infty$ and $n_1 \asymp n_2$. Denote $\mu(x_{1i}, x_{2j}) = \mu_{ij}$ and let $\hat{\boldsymbol{\mu}} = (\hat{\mu}_{11}, \ldots, \hat{\mu}_{n_1 n_2})$ be the original least squared estimator. After adding a penalty on the range, the penalized estimator is $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_{11}, \ldots, \tilde{\mu}_{n_1 n_2})$, the projection of $\tilde{\boldsymbol{y}}$ onto $\mathcal{C}$, where $\lambda > 0, \mathbf{b} = (-1, 0, \cdots, 0, 1)^T$, and $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \lambda \mathbf{b}/2 = (y_{11} + \lambda/2, y_{12}, \ldots, y_{n_1,(n_2-1)}, y_{n_1 n_2} - \lambda/2)$.

Let $\mathcal{L}$ be the collection of all lower sets and let k_L be the number of elements in a lower set L. The penalized estimator of the first cell is

$$\tilde{\mu}_{11} = \min_{L:L\in\mathcal{L}} Av_y(L) = \min_{L:L\in\mathcal{L}} \left\{ \frac{\sum\sum_{i,j:(x_{1i},x_{2j})\in L} y_{ij}}{k_L} + \frac{\lambda}{2k_L} \right\}. \tag{11}$$

Similarly to the univariate case, consistency of $\tilde{\mu}_{11}$ requires the penalty parameter $\lambda \to \infty$ as n increases, and the following theorem shows that $\tilde{\mu}_{11}$ is consistent if $\lambda \asymp n^\alpha$ where $3/8 < \alpha < 1$.

Theorem 6. *Suppose $\mu(x_1, x_2)$ is continuous in a neighborhood of $(0,0)$ and isotonic with respect to x_1 and x_2, then for any $\delta > 0$,*
$\lim_{n\to\infty} \mathrm{P}\left(|\tilde{\mu}_{11} - \mu(0,0)| > \delta\right) = 0$ *if $\lambda \asymp n^\alpha$ where $3/8 < \alpha < 1$.*

Proof. First assume $\lambda = an^\alpha$ for $\alpha \in (0,1)$ and positive a, and choose any $\delta > 0$. By continuity of μ, there is a $d \in (0,1)$ such that $\mu(d,d) - \mu(0,0) < \delta/3$; define L_0 to be the largest lower set contained in $[0,d] \times [0,d]$, then $k_{L_0} \asymp d^2 n$. Define $\bar{\varepsilon}_{(L)} = \sum\sum_{i,j:(x_{1i},x_{2j})\in L} \varepsilon_{ij}/k_L$ and $\bar{\mu}_{(L)} = \sum\sum_{i,j:(x_{1i},x_{2j})\in L} \mu_{ij}/k_L$. For large enough n, $\lambda/(2k_{L_0}) < \delta/3$ and

$$\begin{aligned}
& \mathrm{P}(\tilde{\mu}_{11} \geq \mu(0,0) + \delta) = \mathrm{P}\left(\min_{L:L\in\mathcal{L}} \left\{ \frac{\sum\sum_{i,j:(x_{1i},x_{2j})\in L} y_{ij}}{k_L} + \frac{\lambda}{2k_L} \right\} \geq \mu(0,0) + \delta \right) \\
= \;& \mathrm{P}\left(\bar{\mu}_{(L)} + \bar{\varepsilon}_{(L)} + \frac{\lambda}{2k_L} \geq \mu(0,0) + \delta, \text{ for all } L \in \mathcal{L} \right) \\
= \;& \mathrm{P}\left(\bar{\varepsilon}_{(L)} \leq \bar{\mu}_{(L)} - \mu(0,0) - \delta + \frac{\lambda}{2k_L} \text{ for all } L \in \mathcal{L} \right) \\
\leq \;& \mathrm{P}\left(\bar{\varepsilon}_{(L_0)} \leq \bar{\mu}_{(L_0)} - \mu(0,0) - \delta + \frac{\lambda}{2k_{L_0}} \right) \\
\leq \;& \mathrm{P}\left(\bar{\varepsilon}_{(L_0)} \leq -\delta/3 \right) \\
\to \;& 0.
\end{aligned}$$

On the other hand, given $\delta > 0$,

$$\begin{aligned}
\mathrm{P}\left(\tilde{\mu}_{11} \leq \mu(0,0) - \delta\right) &= \mathrm{P}\left(\min_{L:L\in\mathcal{L}} \left\{ \frac{\sum\sum_{i,j:(x_{1i},x_{2j})\in L} y_{ij}}{k_L} + \frac{\lambda}{2k_L} \right\} \leq \mu(0,0) - \delta \right) \\
&= \mathrm{P}\left(\min_{L:L\in\mathcal{L}} \left\{ \bar{\mu}_{(L)} + \bar{\varepsilon}_{(L)} + \frac{\lambda}{2k_L} \right\} \leq \mu(0,0) - \delta \right) \\
&\leq \mathrm{P}\left(\bigcup_{L\in\mathcal{L}} B_L \right),
\end{aligned}$$

where

$$B_L = \left\{ \sqrt{k_L}\bar{\varepsilon}_{(L)} \geq \frac{\lambda}{2\sqrt{k_L}} + \sqrt{k_L}\delta \right\}.$$

First consider $\mathcal{L}_1$ as the collection of all lower sets that contain at most $n^{2\alpha-\xi}$ elements, namely, $\mathcal{L}_1 = \{L : k_L \leq n^{2\alpha-\xi}\}$ for some small value ξ, and define $\mathcal{L}_2 = \{L : k_L > n^{2\alpha-\xi}\}$. Sampson and Whitaker (1988) have connected representations of lower sets to partitions of natural numbers. Using the notation of Sampson and Whitaker (1988), any lower set on a $n_1 \times n_2$ grid can be denoted as a non-increasing sequence of n_1 integers $n_2 \geq m_1 \geq \cdots \geq m_{n_1} \geq 0$, where $m_k = \#\{(x_1, x_2) : x_1 = k\}$ for $k = 1, \ldots, n_1$. Here m_k is counting the number of elements in the kth column of the lower set and the non-increasing sequence can be viewed as a partition. To find the number of lower sets of size k_L, we need to find the number of possible partitions of k_L with constraints that each element in the partition can be no greater than n_2. For partition with no constraints, Hardy and Ramanujan (1918) showed that the asymptotic expression for the number of possible partitions of a natural number m is $exp\left(\pi\sqrt{2m/3}\right)/(4m\sqrt{3})$. Therefore, a conservative bound for the total number of lower sets in $\mathcal{L}_1$ is $exp(\pi\sqrt{2/3}n^{\alpha-\xi/2})/(4\sqrt{3})$.

The probability $\mathrm{P}(B_L)$ depends only on k_L and is maximized when $\lambda/(2\sqrt{k_L})$ and $\sqrt{k_L}\delta$ are balanced at $k_L = k^*$. Let L^* be a lower set of size k^*, then

$$\begin{aligned}
\mathrm{P}\left(\bigcup_{L\in\mathcal{L}_1} B_L\right) &\leq \frac{e^{\pi\sqrt{\frac{2}{3}}n^{\alpha-\xi/2}}}{4\sqrt{3}}\mathrm{P}(B_{L^*}) \\
&= \frac{e^{\pi\sqrt{\frac{2}{3}}n^{\alpha-\xi/2}}}{4\sqrt{3}}(1-\Phi(\sqrt{2\delta\lambda})) \\
&\leq \frac{e^{\pi\sqrt{\frac{2}{3}}n^{\alpha-\xi/2}}e^{-\delta n^{\alpha}}}{4\sqrt{3}\sqrt{2\pi}\sqrt{2\delta}n^{\alpha/2}} \\
&= \frac{1}{8\sqrt{3\pi\delta}}exp\left\{-\frac{\alpha}{2}\ln n + \pi\sqrt{\frac{2}{3}}n^{\alpha-\xi/2} - \delta n^{\alpha}\right\}
\end{aligned}$$

and the probability goes to 0 if $\alpha > 0$.

Now we focus on the collection of all lower sets in $\mathcal{L}_2$, then

$$\begin{aligned}
&\mathrm{P}\left(\bigcup_{L\in\mathcal{L}_2} B_L\right) = \mathrm{P}\left(\text{at least one } L\in\mathcal{L}_2\left\{k_L\bar{\varepsilon}_{(L)} \geq \frac{\lambda}{2} + k_L\delta\right\}\right) \\
\leq\ &\mathrm{P}\left(\text{at least one } L\in\mathcal{L}_2\left\{k_L\bar{\varepsilon}_{(L)} \geq n^{2\alpha-\xi}\delta\right\}\right) \\
\leq\ &\mathrm{P}\left(\text{at least one row } i\in\{1,2,\ldots,n_2\}\left\{\max_{k=1,2,\ldots,n_1}\sum_{j=1}^{k}\varepsilon_{ij} \geq \frac{\delta n^{2\alpha-\xi}}{n_2}\right\}\right) \\
\leq\ &n_2\cdot\mathrm{P}\left(\max_{k=1,2,\ldots,n_1}\sum_{j=1}^{k}\varepsilon_{1j} \geq \frac{\delta n^{2\alpha-\xi}}{n_2}\right) \\
=\ &n_2\cdot\mathrm{P}\left(\max_{k=1,2,\ldots,n_1}\sum_{j=1}^{k}\varepsilon_{1j}/\sqrt{n_1} \geq \frac{\delta n^{2\alpha-\xi}}{n_2\sqrt{n_1}}\right).
\end{aligned}$$

Using (5) and noting that $n_1 \asymp \sqrt{n}$ and $n_2 \asymp \sqrt{n}$, then

$$\mathrm{P}\left(\bigcup_{L\in\mathcal{L}_2} B_L\right) = 2\sqrt{n}\left[1-\Phi(\delta n^{2\alpha-\xi-3/4})\right] \to 0$$

if $2\alpha-\xi-3/4 > 0$, namely $\alpha > 3/8+\xi/2$. Since ξ is arbitrarily small, the weak consistency of $\tilde{\mu}_{11}$ holds when $\lambda \asymp n^{\alpha}$ where $3/8 < \alpha < 1$.

■

2.4 Simulations

The previous results provide rates for the penalty parameter but no explicit guidance on its actual value for a given n. Hence, we conducted a simulation study to determine an appropriate choice for a recommended proportionality constant between λ and the power of n given for its rate. In this simulation study, we first scaled λ based on the variability in the data, and then considered whether an additional proportionality adjustment was needed. A natural scaling adjustment is provided by σ, which can be estimated from the data. We refer to Meyer and Woodroofe (2000), who showed that

$$\frac{\mathbb{E}\|\mathbf{y}-\hat{\boldsymbol{\mu}}\|^2}{n-\mathbb{E}(D)} \leq \sigma^2 \leq \frac{\mathbb{E}\|\mathbf{y}-\hat{\boldsymbol{\mu}}\|^2}{n-2\mathbb{E}(D)}$$

where D, the number of distinct values among the estimator $\hat{\boldsymbol{\mu}}$, is used as the effective dimension of the model. They suggested that $\hat{\sigma}^2 = SSE/(n-1.5D)$ provides a good choice for isotonic regression, and we will adopt that estimator of σ^2 here. Another scaling adjustment is related to the derivative behavior at the boundary of underlying function, as shown in Pal (2008), $\mu'(0)$ is used as a scaling when the underlying function has a positive slope at zero. More generally, we suggest fitting a line based on the scatter plot of (x, y) first and denote the range of fitted line as r, then scaling the penalty parameter by r if the slope of the fitted line is positive. When the fitted line has non positive slope, the data tend to violate the assumption of monotonicity and the isotonic regression fit will be flat, thus $\sqrt{n}$ can be used as the penalty rate. A simulation was then conducted to determine the mean squared error (MSE) optimal value of m in $\lambda = m\hat{\sigma}rn^{\alpha}$ for a range of datasets following model (1). We found that $\lambda = m\hat{\sigma}rn^{\alpha}$ with m around 1 to 2 minimized the MSE for μ_1 and μ_n with moderate sample sizes as shown in Table 2.1-2.6. The multiplier $m = 1$ is suggested to be conservative without over penalization. For the two-dimensional predictor, a penalty parameter of $\lambda = mr\hat{\sigma}n^{1/2}$ is recommended with multiplier m around .5 to 1.5 and r being the range of fitted values of y regress on x_1 and x_2. The results of

the simulation study is shown in Table 2.7, 2.8, and 2.9, where 10000 datasets are generated under a linear plan $\mu(x_1, x_2) = x_1 + x_2$, a warped plan $\mu(x_1, x_2) = x_1 + x_2 + x_1x_2$, and $y = \max(\max(x_1, x_2) - 1/2, 0)$ where it is flat at the lower end and then increasing.

In the remainder of this section, we report on simulation results on inference for isotonic regression. We first discuss how to construct bootstrap pointwise confidence intervals, after which hypothesis testing in the setting of constant vs. increasing mean function is described.

2.4.1 Bootstrap Confidence Interval

Parametric bootstrap is used to construct 95% confidence intervals: first, generate $y_{bi} = \tilde{\mu}(x_i) + \varepsilon_i'$ for $i = 1, \ldots, n$ where independent errors ε_i' are generated from a normal distribution with mean 0 and estimated model variance $\hat{\sigma}^2$. Next, we fit penalized estimation based on $\boldsymbol{y_b}$ and repeat 5000 iterations. The point-wise 95% confidence intervals at each x_i are constructed from the sorted penalized estimators. An example data set is shown Figure 2.3, which also shows the spiking problem at both ends when calculating 95% confidence interval using unpenalized estimation.

The coverage rates computed based on 5000 data sets with $\mu(x) = x$ and $\sigma^2 = 1$ are shown in Figure 2.4; the confidence intervals constructed with the penalized version show significant improvements compared to unpenalized version, in both interval length and coverage probability. We also computed confidence intervals with the limit distribution results from Pal (2008). In that paper it is shown that if the penalty parameter $\lambda = \beta(\sigma^4/n^2a)^{1/3}$ where β is a centering constant and $a = \mu'(0)/2 > 0$, then $n^{1/3}(\tilde{\mu}_1 - \mu(0)) \Rightarrow (\mu'(0)\sigma^2/2)^{1/3}\mathbb{U}_1$ where $\mathbb{U}_1 = \inf(\mathbb{B}(t) + t^2 + \beta)/t$, $\mathbb{B}$ is a standard Brownian motion, and $t \geq 0$. We choose β so that the distribution of $\mathbb{U}_1$ is approximately centered at 0, and for simulated Brownian motions, we obtain $\beta = 1/2$. We did simulations with $\mu(x) = x$ on $[0, 1]$, and since $\mathbb{U}_1$ has heavy left tail, we chose the 95% probability interval with the shortest length. The resulting confidence intervals have coverage probabilities and lengths shown in Figure 2.4, where we used the known values of σ^2 and $\mu'(0)$. We find coverage probabilities are very similar to

our parametric bootstrap, but our lengths are substantially smaller, and unlike Pal (2008), we do not assume that σ^2 and $\mu'(0)$ are known. Besides the case of $\mu'(0) > 0$ discussed in Pal (2008), we also did simulations for models with $\mu'(0) = 0$ and $\mu(x)$ being constant at a boundary. The results can be found in Figure 2.6 and 2.7 and it showed substantial improvement of the coverage probabilities with smaller size of the confidence intervals.

The 95% confidence intervals for penalized and unpenalized isotonic regression, based on 3000 data sets simulated on a 10×10 grid, are compared in Figure 2.8, showing that the spiking problems are mostly corrected after adding the penalty. Coverage rates and confidence lengths are shown in Figure 2.9. The penalty improves the coverage rates from about 0.4 to above 0.95 at both lower and upper ends, while the confidence length is decreased from 2.3 to below 1.0.

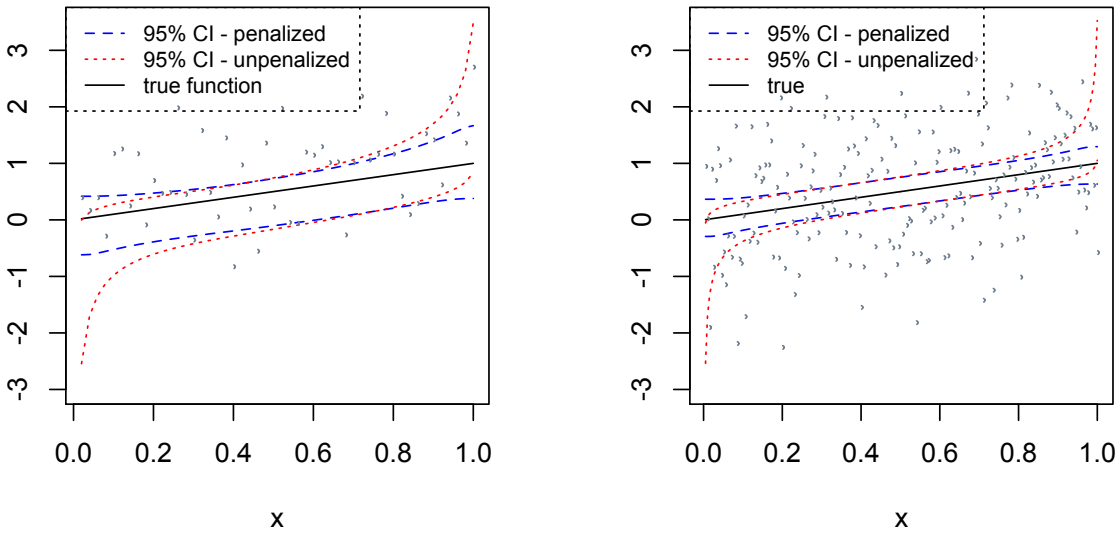

Figure 2.3: 95% confidence interval of $\mu(x) = x$ for sample size of 50(left) and 200(right) with variance $\sigma = 1$ and penalty $\alpha = 1/3$.

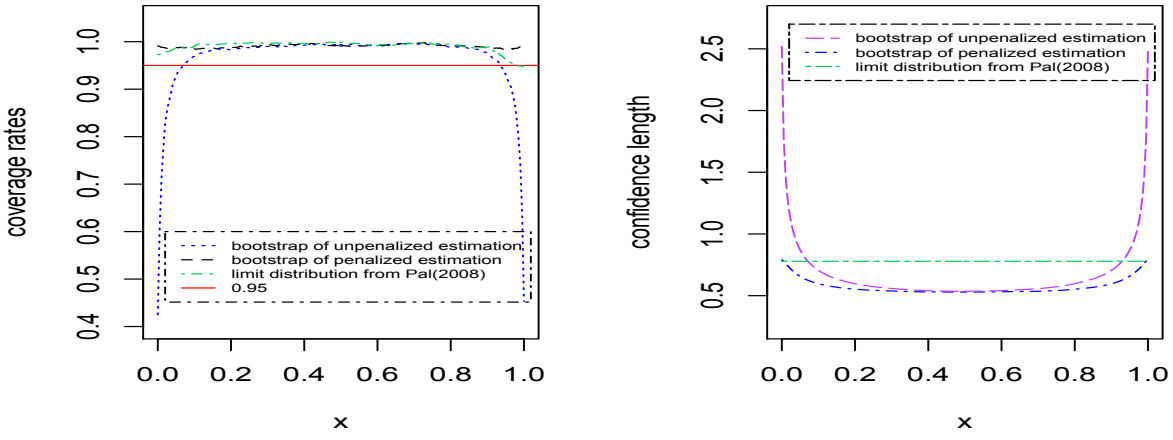

Figure 2.4: For $\mu(x) = x$ on $[0, 1]$ with $n = 200$, $\sigma = 1$ and $\alpha = 1/3$, coverage rates (left plot) and confidence length (right plot) of 95% confidence interval are calculated based on 5000 data sets.

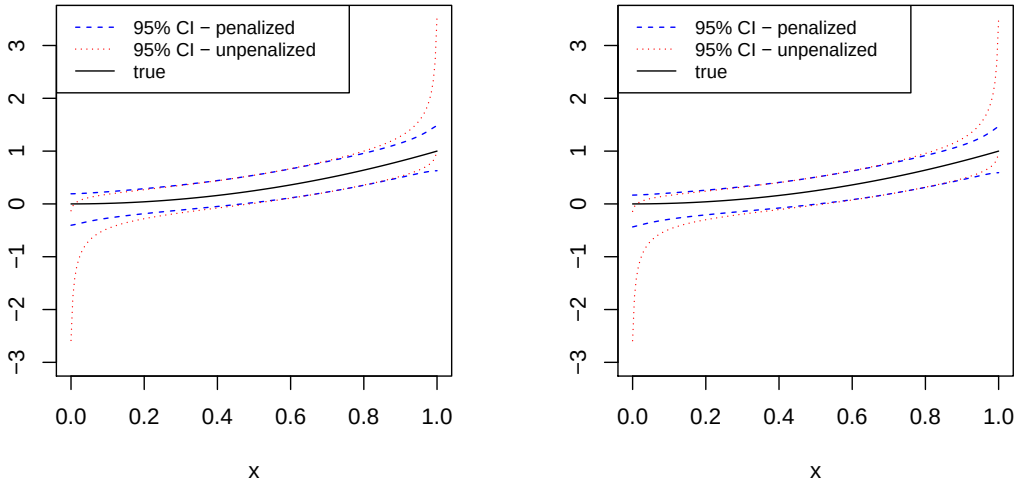

Figure 2.5: 95% confidence interval of $\mu(x) = x^2$ (left) and $\mu(x) = 0$ on $[0, 0.2)$ $\mu(x) = x^2 - .04$ on $[0.2, 1]$ (right) for sample size of 200 with variance $\sigma = 1$ and different penalty rates according to Section 2.2 .

2.4.2 Hypothesis Testing

The likelihood ratio test for constant versus monotone regression function is considered for the unpenalized alternative in Raubertas et al. (1986), and Robertson et al. (1988), Chapter 2. The test statistic is calculated as $(\|\tilde{\boldsymbol{y}} - \bar{\boldsymbol{y}}\|^2 - \|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{\mu}}\|^2)/\|\tilde{\boldsymbol{y}} - \bar{\boldsymbol{y}}\|^2$, which

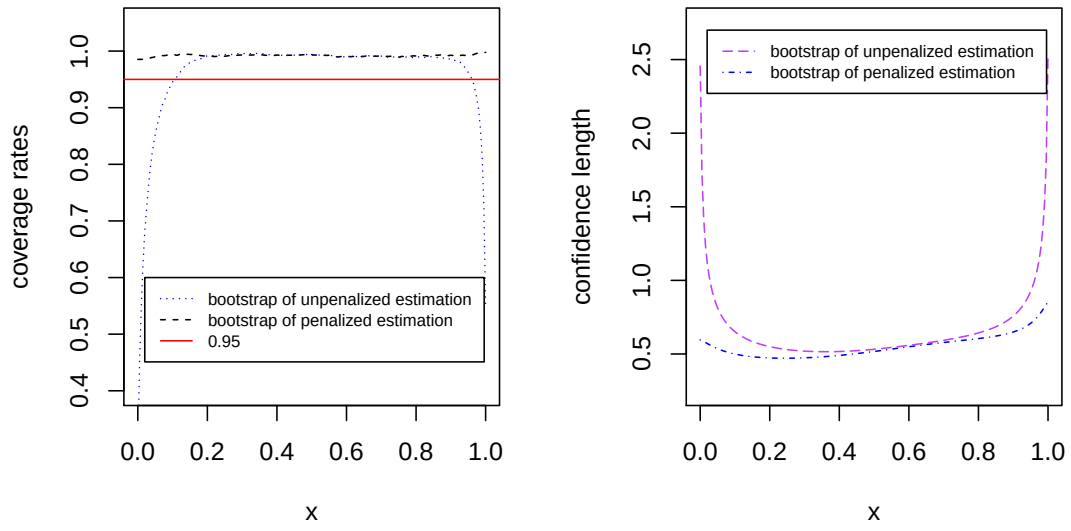

Figure 2.6: For $\mu(x) = x^2$ on $[0, 1]$ with $n = 200$, $\sigma = 1$, coverage rates (left plot) and confidence length (right plot) of 95% confidence interval are calculated based on 5000 data sets.

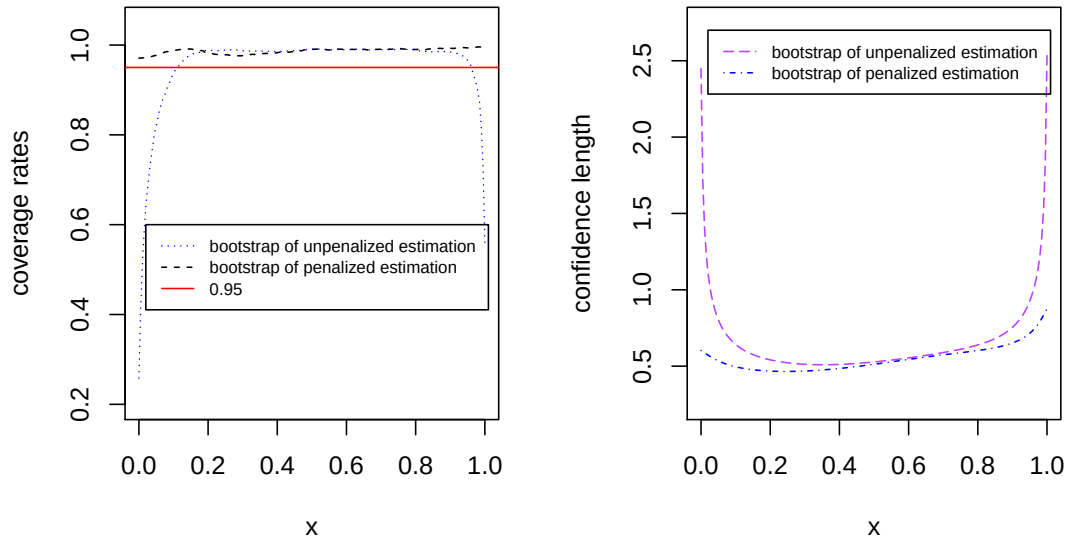

Figure 2.7: For $\mu(x) = 0$ on $[0, 0.2)$ and $\mu(x) = x^2 - .04$ on $[0.2, 1]$ with $n = 200$, $\sigma = 1$, coverage rates (left plot) and confidence length (right plot) of 95% confidence interval are calculated based on 5000 data sets.

is a mixture of beta random variables under the null distribution, for which the mixing parameters can be found through simulations. For the test with an isotonic regression model

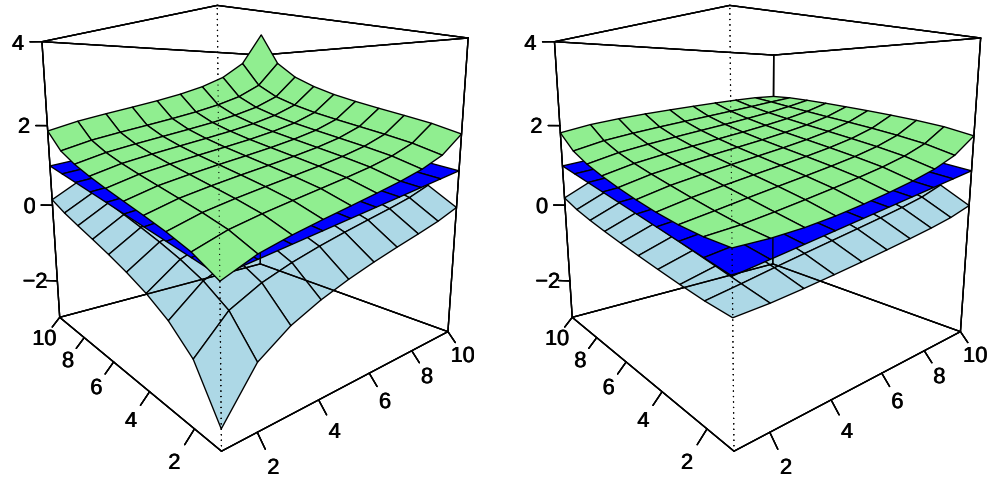

Figure 2.8: For $y = x_1 + x_2$ on a 10×10 grid with equally spaced covariates on $[0, 1]$, $\sigma = 1$ and $\alpha = 1/2$, 95% confidence interval are calculated based on 3000 data sets: unpenalized results are shown on the left plot and penalized results on the right.

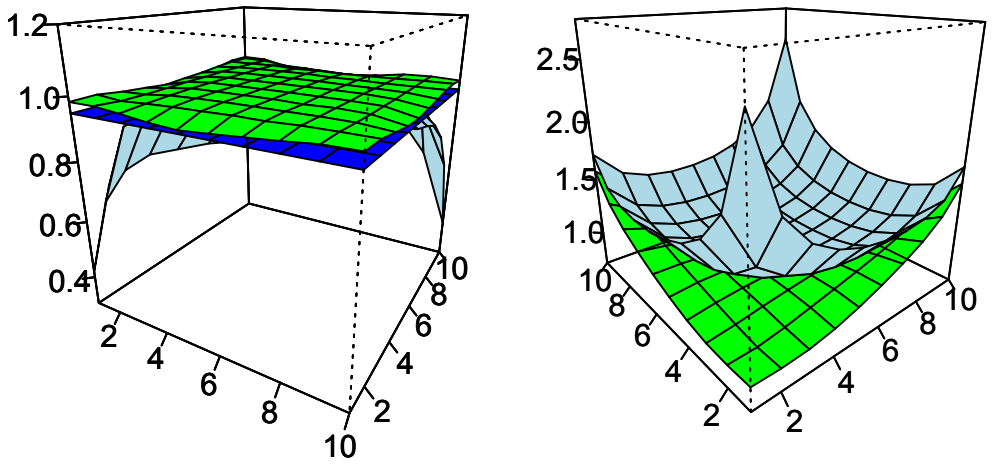

Figure 2.9: For $y = x_1 + x_2$ on a 10×10 grid with equally spaced covariates on $[0, 1]$, $\sigma = 1$ and $\alpha = 1/2$, coverage rates (left plot) and confidence length (right plot) of 95% confidence interval are calculated based on 3000 data sets: unpenalized results are shown in lighter surface, penalized results are in darker surface, and the middle surface in the left plot is the rate of 0.95.

as the alternative, we compare the power of the test using penalized isotonic regression with that using unpenalized regression, and also with the test using a parametric alternative.

For the hypothesis test of $H_0 : \mu$ is constant versus $H_a : \mu$ is monotone, the power of three cases are compared. First, data sets are generated under the alternative hypothesis, and the test statistic is calculated. For unpenalized isotonic regression, the test statistic is a mixture of beta random variable with simulated mixing parameters, and power is the proportion of small p-values. For penalized isotonic regression, critical values of the test statistic are found through simulations under the null hypothesis with penalty $\lambda = n^{1/3}$, and the power is the proportion of test statistics larger than the critical value. Also, the power of the one sided t-test is computed and compared with the isotonic regression results. Linear and sigmoidal regression models are used as alternative in the simulation and shown in Figure 2.10. In both cases, the power for the penalized test is higher than that of the unpenalized test. The one sided t-test beats the other two only when the linear regression model is correct as the alternative.

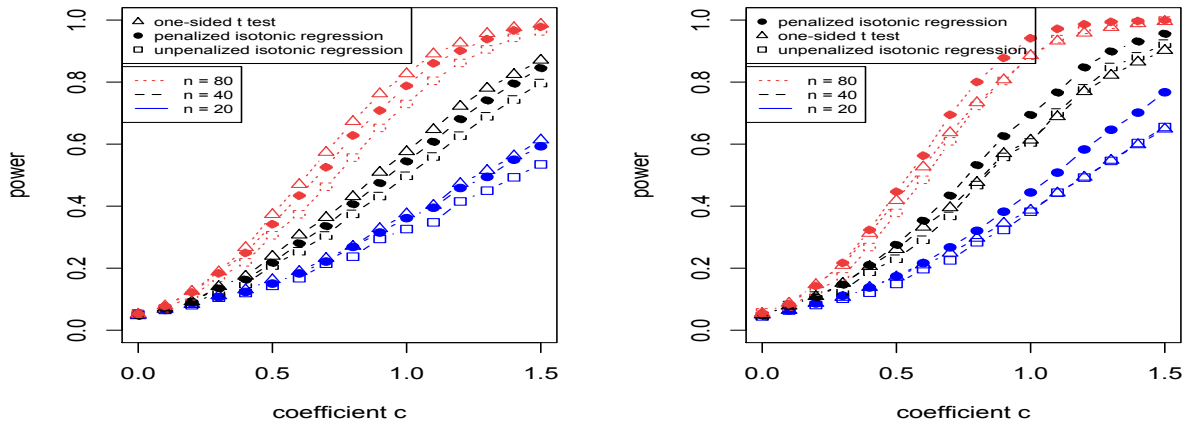

Figure 2.10: Power of hypothesis test of constant v.s. increasing: using $y = cx$ (left) and $y = c \cdot exp(40x - 30)/(1 + exp(40x - 30))$ (right) as alternative model with $\sigma = 1$.

For bivariate isotonic regression, we compare the power for the constant versus increasing test with penalized regression alternative with the standard test with unpenalized regression alternative, and also with the F-test for the significance of the slope parameters in the linear

model

$$\mu(x_1, x_2) = c(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2). \tag{12}$$

For the latter, the test against a constant function uses the $F(3, n-4)$ distribution. We also computed the constrained estimate for the parametric model (12), and used the test of Meyer and Wang (2012). Figure 2.11 shows the power of the four cases using the regression function (12) on 8×8 and 10×10 grids; in this setting, the F test did worst, as expected, but all the other three cases have very similar power. In this case, the parametric assumptions are met, but the power of both penalized and unpenalized nonparametric regression did as well as the constrained parametric regression.

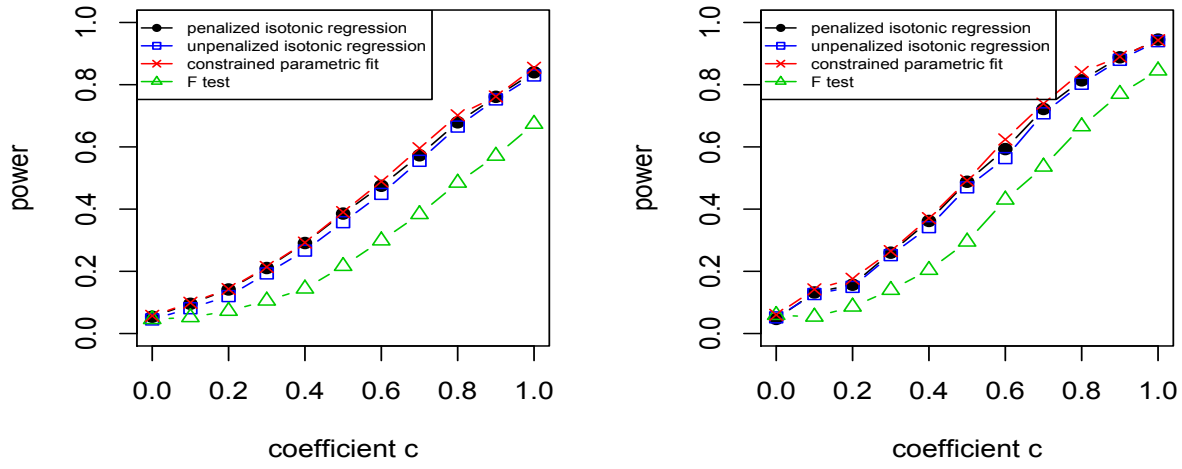

Figure 2.11: Power of hypothesis test of constant v.s. increasing: using $y = c(x_1 + x_2 + x_1 x_2)$ as alternative model on 5 by 5 grids (left) and 8 by 8 grids (right) with $\sigma = 0.5$.

In Figure 2.12, we show the results for the same four tests when the data come from model $y = c \max(\max(x_1, x_2) - 1/2, 0)$, so that the parametric model is misspecified. The nonparametric isotonic regression tests now dominate the parametric tests, and the penalized test is the best one overall.

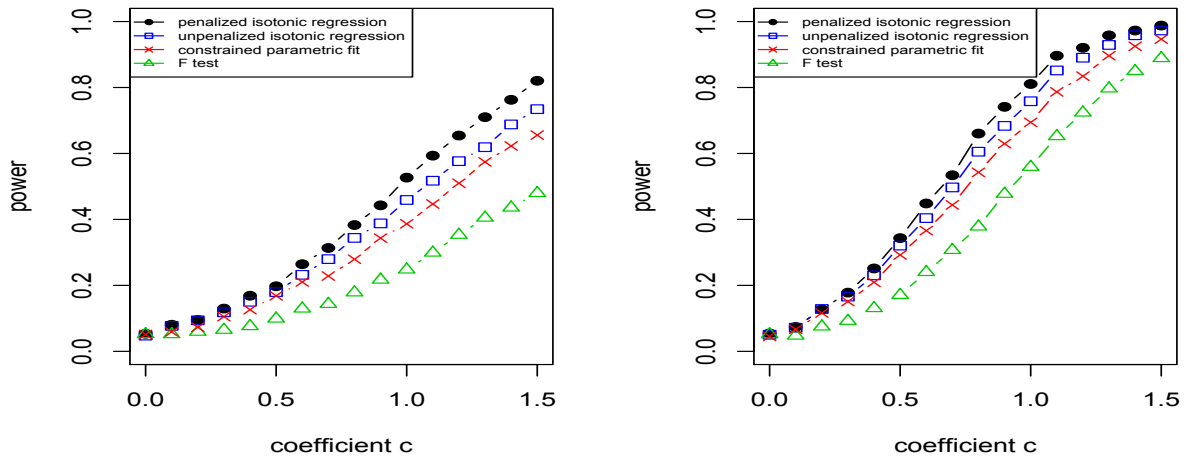

Figure 2.12: Power of hypothesis test of constant v.s. increasing: using $y = c \cdot \max(\max(x_1, x_2) - 1/2, 0)$as alternative model on 8 by 8 grids (left) and 10 by 10 grids (right) with $\sigma = 0.5$

2.4.3 Pollution Data Example

We fit a bivariate isotonic regression model on a pollution dataset described in Aldrin (2004). The data consist of 500 observations taken at a road in Oslo, Norway, between October 2001 and August 2003. The response variable is the values of logarithm of concentration of NO_2. Two covariates are the logarithm of number of cars per hour and the wind speed (meters/second). Since transportation is a large source of air pollution, the concentration of NO_2 is expected to be positively correlated with the cars number. On the other hand, higher wind speed contributes to the diffusion of air pollution and it will reduce the concentration of NO_2. Figure 2.13 compares the penalized fit and unpenalized fit, and it shows that the penalized estimation pulls up the lower end compared to the unpenalized estimation. It seems like that the surface at lower corner has been shifted up, however, as shown in Figure 2.15 the data points are very sparse when wind speed is high and cars number is low. There are 12 estimators being pulled up by the penalty, and the increment is more than 0.5 at some lower points. While in the other end with high cars number and low wind speed, data points are very dense, the penalty has a little effect on the estimators and pulls down the

unpenalized estimators by less than 0.1. The bootstrap confidence interval described in 4.1 is used to calculate the 95% confidence interval and the results are showed in Figure 2.14. The penalized version mostly correct the spiking problem at both ends and gives narrow confidence size.

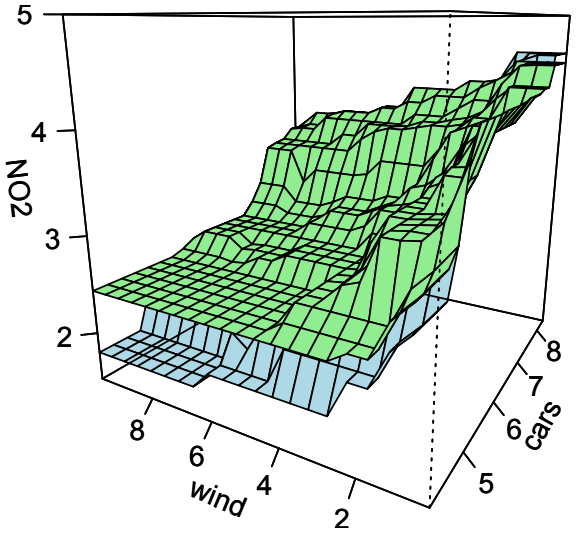

Figure 2.13: Pollution data example: penalized (upper) and unpenalized (lower) fit

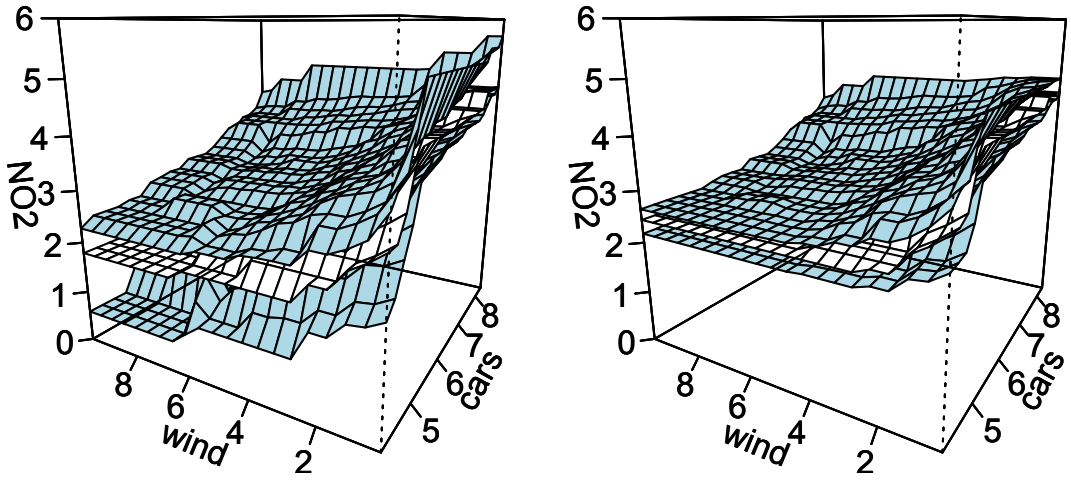

Figure 2.14: Pollution data example: unpenalized (left) and penalized (right) confidence interval

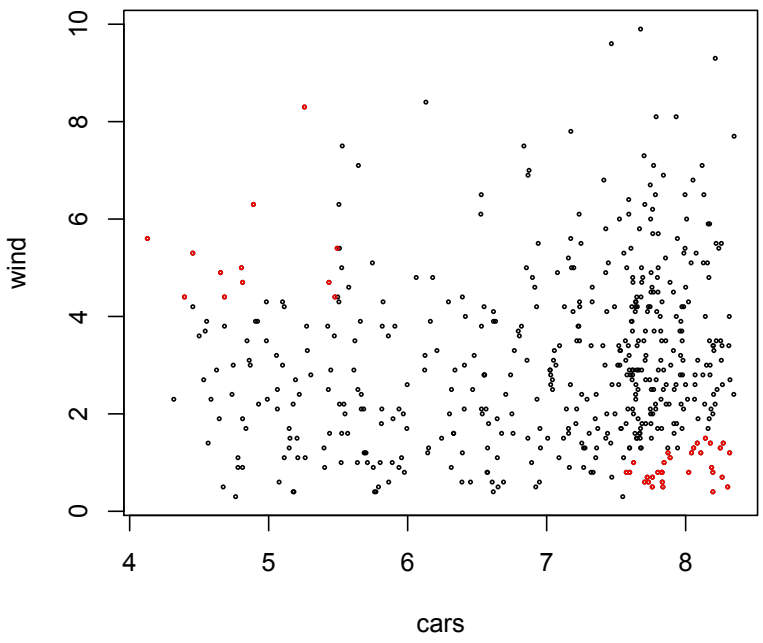

Figure 2.15: Pollution data example: the spread of data points are shown in the plan of wind and cars and the red points show the occurrence of different penalized and unpenalized estimators

2.5 Software

We have assumed so far that the random errors are iid with the same finite variance. If the random errors ε_i in model (1) have a covariance matrix $Cov(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{\Sigma}$ with known $\boldsymbol{\Sigma}$, then let $\boldsymbol{U}'\boldsymbol{U}$ be the Cholesky decomposition of $\boldsymbol{\Sigma}^{-1}$, and $\|\boldsymbol{y} - \boldsymbol{\mu}\|^2$ subject to $\boldsymbol{A}\boldsymbol{\mu} \geq 0$ will be transformed to $\|\boldsymbol{z} - \boldsymbol{\phi}\|^2$ subject to $\tilde{\boldsymbol{A}}\boldsymbol{\phi} \geq \boldsymbol{0}$, where $\boldsymbol{z} = \boldsymbol{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$, $\boldsymbol{\phi} = \boldsymbol{U}\boldsymbol{\mu}$, and $\tilde{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{U}^{-1}$. This transformation is important when time series data are observed and it is useful when we need a weighted least squared for non-distinct covariates, where $\boldsymbol{\Sigma}$ is just a diagonal matrix.

The code for the one- and two-dimensional penalized isotonic regression is available in the `R` package `isotonic.pen`, where the user has the option to choose a penalty parameter, or use the default choice. The code does not require a grid for the two-dimensional case,

although the default penalty parameter derived for the grid is used. Returned are the fit, the estimated standard deviation σ, and the point-wise confidence bounds at the x_i.

CHAPTER 3

SURVEY ESTIMATORS THAT RESPECT NATURAL ORDERINGS

3.1 Introduction

In large-scale government surveys, it is common that estimates are desired for numerous small domains. For example, the U.S. Bureau of Labor Statistics (BLS) uses the National Compensation Survey (http://www.bls.gov/ncs/) to produce wage estimates by location, job type and job level. The Forest Inventory and Analysis (FIA, http://www.fia.fs.fed.us/) of the U.S. Forest Service produces biomass estimates by region, ownership category, forest type, stand age and productivity class. Fine-scale domain estimates are produced by many other government, academic and corporate organizations.

In this survey context, qualitative assumptions often arise naturally. For instance, FIA estimates of average biomass can be expected to increase with increasing productivity classes, and similarly, BLS estimates of average wage should increase by job level within a job type and for a given location. When such qualitative constraints on domain estimates are not respected, concerns arise about the overall reliability of the survey, as these "inversions" are an indication of high variability of the domain estimates.

Isotonic regression, suitably extended to account for the survey design, provides a way to prevent violations of the constraints by adaptively collapsing neighboring domains. As will be shown in this article, it can be justified theoretically from a design-based perspective, which allows full access to asymptotic design-based inference tools, while also resulting in improvements in precision of the resulting estimators. Isotonic regression has never formally been applied in the survey context to our knowledge, but an extensive literature exists in model-based statistics. Early works on isotonic regression and the pooled adjacent violators

algorithm (PAVA), which will be applied here using survey weights, include Ayer et al. (1955) Brunk (1955), and VanEeden (1956). Robertson et al. (1988) give a thorough treatment of ordered estimation and inference.

Our goal is to construct design-based estimators of population domain means that respect an assumed ordering, and further to provide design-based variance estimates for construction of confidence intervals. The estimator itself has a natural interpretation, because as will be shown further below, it corresponds to the classical design-based domain mean estimator, but after the domains have been adaptively pooled to ensure monotonicity. Hence, it is readily implemented in the large-scale government survey context because the pooling can be done separately from the full-scale estimation and data processing steps. Crucially from a practical perspective, once the pooling is determined, the weighted domain estimation can be applied to all the survey variables (although the monotonicity will not necessarily be respected for other variables). The interpretation of the estimates as domain means is also maintained, greatly facilitating acceptance by data users. Similarly, asymptotically valid variance estimation can be done using classical design-based approaches once the domain pooling is determined, again greatly facilitating full-scale implementation.

The remainder of the article is as follows. In Section 3.2, we describe the estimator in more detail and obtain its design-based asymptotic properties. Section 3.3 demonstrates its practical behavior in simulation experiments, and includes a discussion of a replication-based variance estimator. Section 3.4 illustrates the proposed methods on data from the National Health and Nutrition Examination Survey (NHANES). Brief conclusions are stated in Section 3.5.

3.2 Estimator and Properties

Consider a finite population $U_N = \{1, \ldots, k, \ldots, N\}$ and let $U_{t,N}$ denote a domain with N_t elements for $t = 1, \ldots, T$. Assume the domains partition the population; that is, the $U_{t,N}$ are disjoint and $\sum_{t=1}^{T} N_t = N$. The indicator variable for domains is defined as $z_{tk} = 1$ if

$k \in U_t$ and $z_{tk} = 0$ otherwise. For a study variable y, we are interested in estimating the population domain means

$$\bar{y}_{U_t,N} = \frac{\sum_{k \in U_{t,N}} y_k}{N_t} = \frac{\sum_{k \in U_N} z_{tk} y_k}{\sum_{k \in U_N} z_{tk}}, \quad t = 1, \ldots, T. \tag{13}$$

Given a sampling design $p_N(\cdot)$, a probability sample s_N is drawn from U_N, where $p_N(s_N)$ is the probability of drawing the sample s_N. The sample size is denoted by n_N. For the sampling design $p_N(\cdot)$, we assume that we know the inclusion probabilities $\pi_{iN} = \mathrm{P}(i \in s_N) = \sum_{i \in s_N} p_N(s_N) > 0$ and $\pi_{ijN} = \mathrm{P}(i, j \in s_N) = \sum_{i,j \in s_N} p_N(s_N) > 0$ for all $i, j \in U_N$. In all these quantities, the subscript N denotes that fact that we will be working with a sequence of populations indexed by N, with associated sequences of sampling designs and samples, as is customary in design-based asymptotic settings. For simplicity of notation and also following standard practice, we will suppress the subscript N in what follows unless it is needed for clarity. Finally, we define the sample membership indicator $I_k = 1$ if $k \in s$ and $I_k = 0$ otherwise.

Suppose now that it is reasonable to assume that the $\bar{y}_{U_t}$ are non-decreasing over the T categories. If this qualitative information is not used in estimation, the T population domain means can be estimated by the usual weighted sample domain means, the Horvitz-Thompson estimator $\bar{y}_{s_t} = (\sum_{k \in s_t} y_k/\pi_k)/N_t$ or the often preferred Hájek estimator $\tilde{y}_{s_t} = (\sum_{k \in s_t} y_k/\pi_k)/(\sum_{k \in s_t} 1/\pi_k)$ where $\hat{N}_t = (\sum_{k \in s_t} 1/\pi_k)$. We will focus on the Hájek estimator in this article, because it does not require knowledge of the N_t and tends to be more efficient than the Horvitz-Thompson estimator in most practical survey applications.

When the $\tilde{y}_{s_t}$ are computed from small sample sizes, they are approximately unbiased but are highly variable and in particular, they can violate the monotonicity that was expected to hold. Therefore, it is of interest to replace them by domain estimators that do not violate monotonicity. The proposed isotonic estimator $(\hat{\theta}_1, \ldots, \hat{\theta}_T)^\top$ minimizes the weighted sum of

squared deviations from the sample domain means, over the set of ordered vectors; that is,

$$\min_{\theta_1,\ldots,\theta_T} \sum_{t=1}^{T} \hat{N}_t(\tilde{y}_{s_t} - \theta_t)^2, \text{ subject to } \theta_1 \leq \theta_2 \cdots \leq \theta_T.$$

Brunk (1955) presented a max-min formula for the solution of this constrained least squares problem, so that the estimator can be expressed as

$$\hat{\theta}_t = \max_{i \leq t} \min_{t \leq j} \tilde{y}_{s_{i:j}} \text{ ; where } \tilde{y}_{s_{i:j}} = \frac{\sum_{t=i}^{j} \hat{N}_t \tilde{y}_{s_t}}{\sum_{t=i}^{j} \hat{N}_t} = \frac{\sum_{k \in s_{i:j}} y_k/\pi_k}{\sum_{k \in s_{i:j}} 1/\pi_k}, \tag{14}$$

for $i \leq t \leq j$ and notation $s_{i:j}$ is used to denote $s_i \cup \cdots \cup s_j$. As (14) shows, the constrained estimator consists of ordered pooled weighted domain means, and a weighted Pooled Adjacent Violator Algorithm (PAVA) (Robertson et al., 1988) is an efficient way to simultaneously determine the pooling and compute $\hat{\theta}_t$.

A graphical interpretation of PAVA can be expressed through the greatest convex minorant (GCM) of a cumulative sum diagram. The left plot of Figure 3.1 shows that the T domains are collapsed into sets on which the (isotonic) pooled domain means are computed by PAVA, while the plot on the right shows the cumulative sums and the GCM. The left-hand slopes of the minorant are the $\hat{\theta}_t$ values. The pooling is determined by the "corner points" of the GCM, i.e the points where the GCM changes slope.

Although the isotonic estimator is meant to be used when the population domain means $\bar{y}_{U_t}$ are thought to be isotonic, our theoretical results will not rely on this assumption. As we will make more precise below, the pooled sample domain means estimate a pooled version of the population domain means, which will only be equal to the $\bar{y}_{U_t}$ if they are in fact monotone. As we will show in Theorem 7, the pooling of domains in the isotonic regression for the sample will be asymptotically the same as the pooling in the population.

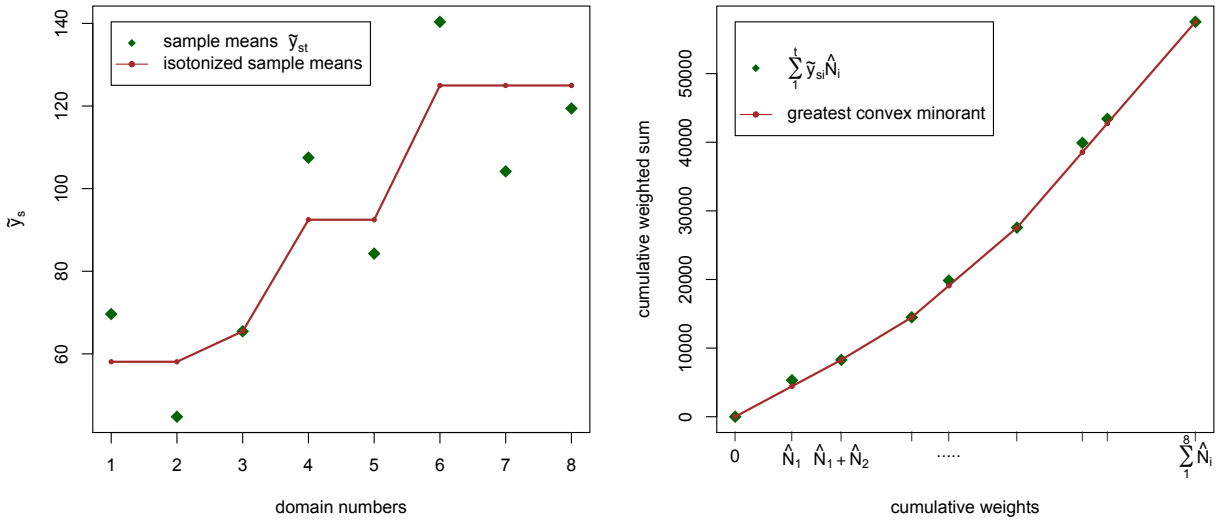

Figure 3.1: Isotonized sample domain means in the left plot are the left-hand slopes of the greatest convex minorant of the cumulative sum diagram shown in the right plot.

Let $U_{i:j}$ denote the pooled population domains $U_i \cup \cdots \cup U_j$, the population pooled domain means are

$$\bar{y}_{U_{i:j}} = \frac{\sum_{k \in U_{i:j}} y_k}{N_i + \cdots + N_j}$$

and for any fixed $i \leq j$, $\tilde{y}_{s_{i:j}}$ is an approximately unbiased estimator of $\bar{y}_{U_{i:j}}$. By straightforward linearization arguments and under mild assumptions (Särndal et al., 1992, Chapter 5), the variance of the asymptotic distribution of $\tilde{y}_{s_{i:j}}$ is given by

$$AV(\tilde{y}_{s_{i:j}}) = \frac{1}{(N_i + \cdots + N_j)^2} \sum_{k,l \in U_{i:j}} \sum \Delta_{kl} \left(\frac{y_k - \bar{y}_{U_{i:j}}}{\pi_k} \right) \left(\frac{y_l - \bar{y}_{U_{i:j}}}{\pi_l} \right) \tag{15}$$

with $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$.

A design consistent estimator of this quantity is

$$\hat{V}(\tilde{y}_{s_{i:j}}) = \frac{1}{(\hat{N}_i + \cdots + \hat{N}_j)^2} \sum_{k,l \in s_{i:j}} \sum \frac{\Delta_{kl}}{\pi_{kl}} \left(\frac{y_k - \tilde{y}_{s_{i:j}}}{\pi_k} \right) \left(\frac{y_l - \tilde{y}_{s_{i:j}}}{\pi_l} \right). \tag{16}$$

In Theorem 2, we will show that this classical variance estimator, applied after the pooling is obtained on the sample data, is also appropriate for inference on the $\hat{\theta}_t$.

Before stating these theoretical results more precisely, we list our required assumptions.

Assumption 1. *As $N \to \infty$, N_t/N is bounded away from 0 and 1, for $t = 1, 2, \ldots, T$.*

Assumption 2. *There exist constants $\mu_t, t = 1, 2, \ldots, T$ such that $|\bar{y}_{U_t} - \mu_t| = O(1/\sqrt{N})$.*

Assumption 3. *The limiting design covariances*

$$\Sigma_{tm} = \lim_{N \to \infty} \frac{1}{N_t N_m} \sum_{k,l \in U} \sum \Delta_{kl} \frac{z_{tk} y_k}{\pi_k} \frac{z_{ml} y_l}{\pi_l},$$

$t, m = 1, 2, \ldots, T$, satisfy $0 < n\,\Sigma_{tm} < C$ for some $C > 0$.

Assumption 4. *The limiting distribution of the Horvitz-Thompson domain estimators $(\bar{y}_{s_1}, \ldots, \bar{y}_{s_T})$ is multivariate normal,*

$$\sqrt{n} \begin{pmatrix} \frac{1}{N_1} \sum_U z_{1k} y_k (\frac{I_k}{\pi_k} - 1) \\ \vdots \\ \frac{1}{N_T} \sum_U z_{Tk} y_k (\frac{I_k}{\pi_k} - 1) \end{pmatrix} \to \mathcal{N}(0, \Sigma),$$

where Σ is the $T \times T$ matrix with elements $n\Sigma_{tm}$ and is invertible.

Assumption 5. *The Horvitz-Thompson domain mean covariance estimators,*

$$\hat{\Sigma}_{tt} = \frac{1}{{N_t}^2} \sum_{k,l \in s} \sum \frac{\Delta_{kl}}{\pi_{kl}} \frac{z_{tk} y_k}{\pi_k} \frac{z_{tl} y_l}{\pi_l},$$

are design consistent, i.e., $n(\hat{\Sigma}_{tt} - \Sigma_{tt}) = o_p(1)$.

Assumptions 1 and 2 are mild regularity conditions on the sequence of populations. More specifically, Assumption 1 bounds the sizes of domains in the population and ensures that there are no asymptotically vanishing domains. While we do not assume a model for the population data, Assumption 2 implies that the population domain means have well-defined limits, denoted μ_t, and we specify a rate at which these limits are approached as $N \to \infty$. Note that this rate would be the usual model-based rate for this convergence (in probability) if a typical probabilistic model for the y_k had been specified, given Assumption 1.

Assumptions 3–5 ensure that the usual design-based asymptotics hold for Horvitz-Thompson estimators of the population domain means, as is typically assumed when deriving design-based asymptotic results. This includes the fact that the variances and covariances have $O(n^{-1})$ rates, which holds for most common sampling designs (Assumption 3), asymptotic normality of the domain mean estimators (Assumption 4) and consistency of the unbiased covariance estimators (Assumption 5).

The following two Theorems contain the main theoretical results. Two technical Lemmas with proofs are in the Appendix.

Theorem 7. *Let $(\theta_1, \ldots, \theta_T)^\top$ be the isotonic regression of the population means $(\bar{y}_{U_1}, \ldots, \bar{y}_{U_T})^\top$, $(\hat{\theta}_1, \ldots, \hat{\theta}_T)^\top$ be the weighted isotonic regression of the sample domain means $(\tilde{y}_{s_1}, \ldots, \tilde{y}_{s_T})^\top$, and A be the event that the two isotonic regressions have the same pooling, i.e., the same set of corner points of the GCM. Then, under Assumptions 1–4, $\lim_{N\to\infty} \sqrt{n} P(A^c) = 0$, where A^c denotes the complement of A.*

Proof. For $t = 1, \ldots, T$, let $v_t = \sum_{i=1}^{t} N_i \bar{y}_{U_i}/N$ and $w_t = \sum_{i=1}^{t} N_i/N$, then

$$\frac{v_t - v_{t-1}}{w_t - w_{t-1}} = \bar{y}_{U_t} \qquad \text{and} \qquad \frac{v_j - v_i}{w_j - w_i} = \frac{N_{i+1}\bar{y}_{U_i} + \cdots + N_j \bar{y}_{U_j}}{N_{i+1} + \cdots + N_j}.$$

Let (w_t, g_t) be the points on the greatest convex minorant of (w_t, v_t) and θ_t be the left-derivative slope at (w_t, g_t), then θ_t furnishes the isotonic regression of $\bar{y}_{U_t}$. On the other

hand, use $(\hat{w}_t, \hat{v}_t)$ to represent the sample version of (w_t, v_t), where $\hat{v}_t = \sum_{i=1}^{t} \hat{N}_i \tilde{y}_{s_i}/N = \sum_{i=1}^{t} N_i \bar{y}_{s_i}/N$ and $\hat{w}_t = \sum_{i=1}^{t} \hat{N}_i/N$.

Let $\mathcal{T} = \{1, \ldots, t, \ldots, T\}$ and use $\mathcal{J} = \{j_1, \ldots, j_S\}$ to denote the S corner point indices of (w_t, g_t) where $0 < j_1 < j_2 < \cdots < j_S = T$. Denote $\delta_1 = \min_{s=1,\ldots,S-1}(\theta_{j_{s+1}} - \theta_{j_s})$, $\delta_2 = \min(w_{i+1} - w_i)$ for $i = 1, \ldots, T-1$, $\delta_3 = \min(v_i - g_i)$ for $i \in \mathcal{J}^c$, $R_w = \max w_t - \min w_t$ and $R_v = \max v_t - \min v_t$ for $t = 1, \ldots, T$. Notice that $R_w < 1$ by definition and R_v is bounded by Assumption 1 and 2. Let

$$\epsilon_1 = \min\left\{\frac{\delta_2}{2}, \frac{\delta_2\delta_3}{4R_v}, \frac{\delta_1\delta_2^2}{4R_v}\right\} \text{ and } \epsilon_2 = \min\left\{\frac{\delta_2\delta_3 - 4R_v\epsilon_1}{4R_w}, \frac{\delta_1\delta_2^2 - 4R_v\epsilon_1}{4R_w}\right\},$$

then event $A = \{|\hat{w}_t - w_t| < \epsilon_1 \text{ and } |\hat{v}_t - v_t| < \epsilon_2\}$ by Lemma 3 and $\mathrm{P}(A^c) \leq \mathrm{P}(|\hat{w}_t - w_t| \geq \epsilon_1) + \mathrm{P}(|\hat{v}_t - v_t| \geq \epsilon_2)$. Since

$$|\hat{v}_t - v_t| \;=\; \frac{1}{N}\left|\sum_{i=1}^{t} N_i \bar{y}_{s_i} - \sum_{i=1}^{t} N_i \bar{y}_{U_i}\right| \leq \frac{1}{N}\sum_{i=1}^{t} N_i |\bar{y}_{s_i} - \bar{y}_{U_i}|,$$

then

$$\mathrm{P}\left(|\hat{v}_t - v_t| < \epsilon_2\right) > \mathrm{P}\left(\frac{1}{N}\sum_{i=1}^{t} N_i |\bar{y}_{s_i} - \bar{y}_{U_i}| < \epsilon_2\right) > \mathrm{P}\left(|\bar{y}_{s_t} - \bar{y}_{U_t}| < \epsilon_2\right).$$

Assumption 4 implies $V(\bar{y}_{s_t}) = O(1/n)$ and by Chebyshev's inequality,

$$\mathrm{P}\left(|\hat{v}_t - v_t| \geq \epsilon_2\right) < \mathrm{P}\left(|\bar{y}_{s_t} - \bar{y}_{U_t}| \geq \epsilon_2\right) \leq \frac{V(\bar{y}_{s_t})}{\epsilon_2^2} = o\left(\frac{1}{\sqrt{n}}\right).$$

Similarly, $\mathrm{P}(|\hat{w}_t - w_t| \geq \epsilon_1) = o(1/\sqrt{n})$, so that $\mathrm{P}(A^c) = o(1/\sqrt{n})$. By Chebyshev's inequality, for every $\zeta > 0$,

$$\lim_{n\to 0} \mathrm{P}\left(|I_{A^c}| > \frac{\zeta}{\sqrt{n}}\right) \leq \lim_{n\to 0} \frac{\sqrt{n}E|A^c|}{\zeta} = \lim_{n\to 0} \frac{\sqrt{n}\mathrm{P}(A^c)}{\zeta} = 0.$$

■

Theorem 8. *For $t \in \{1, \ldots, T\}$, let θ_t, $\hat{\theta}_t$ be defined as in Theorem 1, and t_-, t_+ be the indices selected by isotonic regression so that θ_t is the pooled population mean from domain U_{t_-} through U_{t_+} with $t_- \leq t \leq t_+$. Under A1–A4,*

$$\hat{\theta}_t = \tilde{y}_{s_{t_-:t_+}} + o_p\left(\frac{1}{\sqrt{n}}\right)$$

and

$$\frac{\hat{\theta}_t - \theta_t}{\sqrt{AV(\hat{\theta}_t)}} \xrightarrow{\mathcal{L}} N(0,1),$$

with $AV(\hat{\theta}_t) = AV(\tilde{y}_{s_{t_-:t_+}})$ as defined in (15). Additionally, under A5,

$$\frac{\hat{\theta}_t - \theta_t}{\sqrt{\hat{V}(\hat{\theta}_t)}} \xrightarrow{\mathcal{L}} N(0,1),$$

where $\hat{V}(\hat{\theta}_t)$ is the variance estimator defined in (16) *with the pooling obtained by isotonic regression applied to the sample domain estimators $\tilde{y}_{s_1}, \ldots, \tilde{y}_{s_T}$.*

Proof. By definition, $\theta_t = (N_{t_-}\bar{y}_{U_{t_-}} + \cdots + N_{t^+}\bar{y}_{U_{t^+}})/(N_{t_-} + \cdots + N_{t^+})$ and $\tilde{y}_{s_{t_-}\cup\cdots\cup s_{t^+}}$ has the same pooling as θ_t. By Theorem 1, $I_{\{A^c\}} = o_p(1/\sqrt{n})$, so that

$$\begin{aligned}
\hat{\theta}_t - \theta_t &= (\tilde{y}_{s_{t_-:t^+}} - \theta_t)I_{\{A\}} + (\tilde{y}_{s_C} - \theta_t)I_{\{A^c\}} \\
&= (\tilde{y}_{s_{t_-:t^+}} - \theta_t)(1 - o_p(1/\sqrt{n})) + (\tilde{y}_{s_C} - \theta_t)o_p(1/\sqrt{n})
\end{aligned}$$

$$= \quad \tilde{y}_{s_{t\_:t^+}} - \theta_t + o_p(1/\sqrt{n}),$$

where $\tilde{y}_{s_C}$ represents the domain mean estimator with different pooling than θ_t. By Assumption 2 - 3, $\tilde{y}_{s_C}$ will converge to the population domain mean with the same pooling as $\tilde{y}_C$ and $\tilde{y}_{s_C} = O_p(1)$. Assumption 3 - 4 and Lemma 2 immediately lead to $(\hat{\theta}_t - \theta_t)/\sqrt{AV(\hat{\theta}_t)} \to N(0,1)$. Further,

$$n(\hat{V}(\hat{\theta}_t) - AV(\hat{\theta}_t)) = n[\hat{V}(\tilde{y}_{s_{t\_:t^+}}) - AV(\hat{\theta}_t)]I_{\{A\}} + [\hat{V}(\tilde{y}_{s_C}) - \hat{V}(\tilde{y}_{s_{t\_:t^+}})]I_{\{A^c\}}$$

where $\hat{V}(\tilde{y}_{s_C}) - \hat{V}(\tilde{y}_{s_{t\_:t^+}}) = O_p(1/n)$ and $n[\hat{V}(\tilde{y}_{s_{t\_:t^+}}) - AV(\hat{\theta}_t)] = o_p(1)$ by Lemma 4, so that $n[\hat{V}(\hat{\theta}_t) - AV(\hat{\theta}_t)] = o_p(1)$ and by Slutsky's theorem,

$$\frac{\hat{\theta}_t - \theta_t}{\sqrt{\hat{V}(\hat{\theta}_t)}} \xrightarrow{\mathcal{L}} N(0,1).$$

■

Theorem 7 states that, with design probability approaching 1 as $N \to \infty$, the collapsing of domains produced by the PAVA estimator is the same in both sample and population, and that the probability of a different pooling goes to zero at faster than $n^{1/2}$ rate. This result is the key step allowing for the design-based asymptotic results obtained in Theorem 8, which include the design consistency, the asymptotic distribution with respect to the sequence of sampling designs and the fact that the variance estimator computed after the pooling is obtained for the sample results in asymptotically correct inference.

Consider now another survey variable denoted x, which also satisfies Assumptions 2-5 but which is not assumed to have monotone domain means. While we do not derive it formally here, it is readily seen that an analogue of Theorem 8 applies for the domain means of x computed on the pooled domains selected by the weighted PAVA for variable y, because of the asymptotic equivalence between the sample and population pooling based on y obtained

in Theorem 7. In other words, the pooled domain estimators for unconstrained variables remain asymptotically normal with valid inference. This makes it possible to use a survey variable y to select a domain pooling, and then apply that pooling to the other variables in that same survey. This in turn ensures consistency of estimates for different variables released in tabular form, as is often done in large-scale government survey programs.

3.3 Simulations

To demonstrate the practical properties of constrained survey domain estimation, we simulate populations and samples using four limiting domain mean values

$$\begin{aligned}
\text{sigmoid:} \quad & \mu_{1t} = exp(20bt/T - b10)/(1 + exp(20bt/T - b10)) \\
\text{linear:} \quad & \mu_{2t} = 1 + b(t/T - 0.5) \\
\text{quadratic:} \quad & \mu_{3t} = 1 + b(t/T)^2 \\
\text{bump:} \quad & \mu_{4t} = 1 + b(t/T)I_{\{t/T<0.5 \text{ or } t/T>0.7\}} + I_{\{0.5<t/T\leq 0.7\}}
\end{aligned}$$

where b is a coefficient, T is the total number of domains, and $t = 1, 2, \ldots, T$. For μ_1, μ_2, and μ_3, the mean are increasing with respect to t, while μ_4 has a 'bump' in the middle of a linear trend. The population size is set as $N = 10000$, with $N_t = N/T$, and the population values y_k are generated by adding independent and identically distributed $N(0, \sigma^2)$ errors to the μ_t values. We use two possible values for the coefficient b and the standard deviation of the errors σ respectively: $b = 0.5$, $b = 1$, $\sigma = 0.5$, and $\sigma = 1$. Samples are generated from a stratified sampling design with simple random sampling without replacement in all strata, with $H = 4$ strata that cut across the T domains. The strata are determined by a variable z which is correlated with y, generated here by adding $N(0, 1)$ errors to $(t/T)\sigma$. Stratum membership is determined by sorting the population based on z and assigning N/H elements to each stratum based on their ranked z. Finally, the total sample size $n = 200$ is divided among the strata as $(25, 50, 50, 75)$, resulting in an informative design. Explicit expressions

for the domain mean estimators and corresponding variance estimators for this design are Särndal et al. (1992, Chapter 10.3).

For each combination of mean function, coefficient, and standard deviation, 50000 replicate samples are selected for $T = 5$, $T = 10$ and $T = 20$. Our target is to estimate the population domain means $\bar{y}_{U_t}$, which are possibly not strictly increasing due to the randomly generated model errors. For each sample, both the unconstrained estimate $(\tilde{y}_{s_1}, \ldots, \tilde{y}_{s_T})^\top$, the constrained estimate $(\hat{\theta}_1, \ldots, \hat{\theta}_T)^\top$, and their corresponding standard errors are computed, and then a 95% confidence interval is constructed. The design-averaged performance of these two types of estimators can be compared since the population is kept fixed for the replicated samples. When the total number of domains is $T = 20$, there are about 30 to 40 samples out of the 50000 replicate samples that have no observations in at least one domain. Although the isotonized estimator has a natural interpolation for empty cells, there is no corresponding non-isotonized estimator, so we exclude these samples from calculation of the average performance.

Results are displayed in Figure 3.2 to 3.13. The first column of every plot shows examples of fitting one sample data, the second column and third column present the average of estimators with the 95% confidence intervals, as well as the average length of confidence intervals and the coverage rates. In the case of sigmoid model with $T = 5$, shown in Figure 3.2 to 3.4,the sample domain means are already "close to" isotonic, especially for the smaller variance, so that the improvements are small, but for the $T = 10$ and $T = 20$ cases, the confidence interval lengths are substantially smaller for the isotonic estimator, with improved coverage. The population domain means are not strictly monotone in the simulation, but the constrained estimators are still close to the population domain means with smaller size in confidence interval and high coverage rates. The "price" for the improvements in the performance of the confidence intervals is that the constrained estimator has some bias at boundaries.

Simulations with linear and quadratic mean structures show similar results in Figure 3.5 to 3.10: the length of confidence intervals of the constrained estimator are smaller while the coverage rates are competitive with the unconstrained estimator. The bump model μ_4 represent the case where the monotone assumption is violated. Figures 3.11 to 3.13 give the outcomes when the constrained estimation is applied and it is shown that the estimation starting from the bump becomes worse. The outcomes show that the performance of the constrained estimators degrades with increasing size of the monotonicity violation, with bias leading to incorrect confidence intervals with very low coverage rates. In situations of a modest bump, the constrained estimation may still cover the true population domain means, but coverage rates are dropping significantly. Overall, these simulation results suggest that the constrained estimation will help improve the inference of confidence intervals when the monotone assumption is at least close to correct, especially when the number of domains is large. However, if the qualitative constraints are not met, the constrained estimation should be applied with care.

As part of the exploration of the practical uses of constrained domain estimation, we investigated jackknife variance estimation. Replication methods are commonly used for large-scale surveys, so it is of interest to see whether this method would perform satisfactorily for our proposed estimation approach. One issue, however, is that the usual jackknife variance estimators are asymptotically consistent only under sufficiently smooth conditions for the estimators (Shao and Wu, 1989). In our case, the estimator is not smooth, because of the sample-dependent pooling of the domains. A modified jackknife method (Shao and Wu, 1989) is usually suggested to obtain consistency with less stringent smoothness conditions. However, both the usual delete-1 and the modified jackknife are rarely used on large-scale surveys due to the computational burden, so that we will investigate the more commonly implemented "delete-a-group" (DAG) jackknife (Kott, 2001).

To compare the jackknife method with the linearization-based method, we continue the stratified sampling design and choose the case of $\mu_1(t) = exp(20bt/T - b10)/(1+exp(20bt/T -$

$b10)$) with $T = 20$ domains as an example. First, denote d as the number to be deleted from the sample, and divide each stratum into n/d groups. Second, for $r = 1, 2, \ldots, n/d$, delete the rth group from each stratum, adjust the sample weights as $w_k^{(r)} = w_k/(1 - d/n)$ in each stratum, and calculate the constrained estimator $\hat{\theta}_t^{(r)}$ using the remaining sample and the adjusted weights. Third, the jackknife variance estimator is

$$\hat{V}(\hat{\theta}_t) = \left(\frac{n/d - 1}{n/d}\right) \sum_{r=1}^{n/d} (\hat{\theta}_t^{(r)} - \hat{\theta}_t)^2.$$

We investigate the behavior of the DAG jackknife variance estimator with $d = 1$, 10, and 20 as well as the linearization-based variance estimator, based on 50000 replicate samples. For our sample size of $n = 200$, $d = 10$ corresponds to a variance estimator with 20 replicates, while $d = 20$ has only 10 replicates.

In top plots of Figure 3.14 to 3.17, we compare the averaged variance estimate in the constrained and unconstrained cases, each containing four scenarios of variance estimation: the linearization based, delete-1 jackknife, delete-10 jackknife, and delete-20 jackknife. In the unconstrained case, the linearization-based estimator is underestimating the variance while the jackknife method is overestimating, which corresponds to the behavior generally observed in practice for these estimators. The results for the constrained estimator are markedly different, however, with all 4 variance estimators exhibiting very similar behavior, and also performing significantly better than for the unconstrained estimator. The bottom plots in Figure 3.14 to 3.17 compare the coverage rates and the sizes of the confidence intervals, and show that the four variance estimation methods are very similar for the constrained estimator, and achieve better coverage and smaller confidence intervals than for the unconstrained estimator.

3.4 Example using NHANES cholesterol data

As an illustration of the usefulness of the proposed estimation method, we consider data from the 2011-2012 National Health and Nutrition Examination Survey (NHANES), downloaded from the Centers for Disease Control and Prevention website. There are $N = 3250$ observations with complete records for cholesterol levels and waist size; for the purposes of illustration, we treat these observations as a finite population and we sample it using stratified simple random sampling without replacement. The sample size n is set to be 60, 210, and 450. We sample $n/3$ subjects in each of three waist-size categories that represent the lower tenth percentile of the population, the upper tenth, and the middle 80%. Waist size is related to cholesterol level, so that the sampling is informative with respect to that variable. We are interested in estimating the average log cholesterol level in ten domains, defined by successive age groups. If the researchers believe that the population average cholesterol level is increasing with age category, then a constrained method can provide more precise estimates.

A representative sample of size 210 is shown in the last plot of Figure 3.19. The unconstrained domain estimates are weighted sample averages and shown as red dashed line. Although the population means are not quite monotone, the constrained estimates in blue dots show less variation and are considerably closer to the population averages, despite the monotonicity violation. Sampling and estimation are then repeated 50000 times, and the average estimates with confidence intervals are shown in the first three plots in Figures 3.19. The constrained estimator gives narrower confidence intervals while the coverage rates are close to the unconstrained case. A drop of the coverage rate is shown at the largest age group, due to the violation of the monotonicity in population means. The box plots in Figures 3.19 summarize the estimates in each of the ten age categories. The unconstrained estimators of domain means are unbiased but have larger variation, while the monotone fits have smaller variance, but are slightly biased for several of the age groups.

When the sample size is small as $n = 60$, as shown in Figure 3.18, some of the age categories may contain no observations in the sample and the coverage rates are below 0.80 in the unconstrained case, the constrained fit improves the coverage rates with smaller confidence length. For $n = 450$ in Figure 3.20, it is similar to the case of $n = 210$, the constrained estimate gives narrower confidence intervals while the coverage rates are close to the unconstrained case. For all three cases of sample sizes, the unconstrained estimates of domain means are unbiased but have larger variation, while the monotone fits have smaller variance, but are slightly biased larger at the largest age group and biased small at the smallest age group. If the researchers believe that the population average cholesterol level is increasing with age category, then a constrained method can provide better estimates.

3.5 Discussion

Isotonic regression is readily implemented in the estimation of domain means from survey data, and can lead to dramatic improvements in precision when the monotonicity constraint is at least approximately valid. This chapter focuses only on the constrained estimation in a case of simple ordering, but this approach is applicable to surveys with more complicated domain structures including multiple covariates and combinations of constrained and unconstrained relationships.

The method described in this chapter focused on the estimation of domain means and did not require that the N_t be known. If they are known, however, then a straightforward extension of this method is to develop a post-stratified estimator with adaptively selected post-stratification. The approach would be based on using isotonic regression to determine the pooling of the post-strata based on a monotonicity constraint on the direct estimators, followed by computing post-stratified weights as the original survey weights times the ratio of the true post-stratum sizes over the estimated post-stratum sizes. We conjecture that this method will lead to an adaptively post-stratified estimator that will be asymptotically equivalent to one with the same fixed post-strata.

Outside of the survey context, constrained estimation from isotonic regression is known to be biased at boundaries (Sampson et al., 2003), which is often referred to as the "spiking" problem. This could be seen in the simulations in Section 3, with a drop of the confidence interval's coverage rates at both ends of domain range. The penalty on the range from Chapter 2 can be used here as well, to correct the spiking problem and to improve the behaviors of both constrained estimation and the variance estimates at boundaries. The idea of penalized isotonic regression could be embedded in survey context and further improve the adaptive domain estimation.

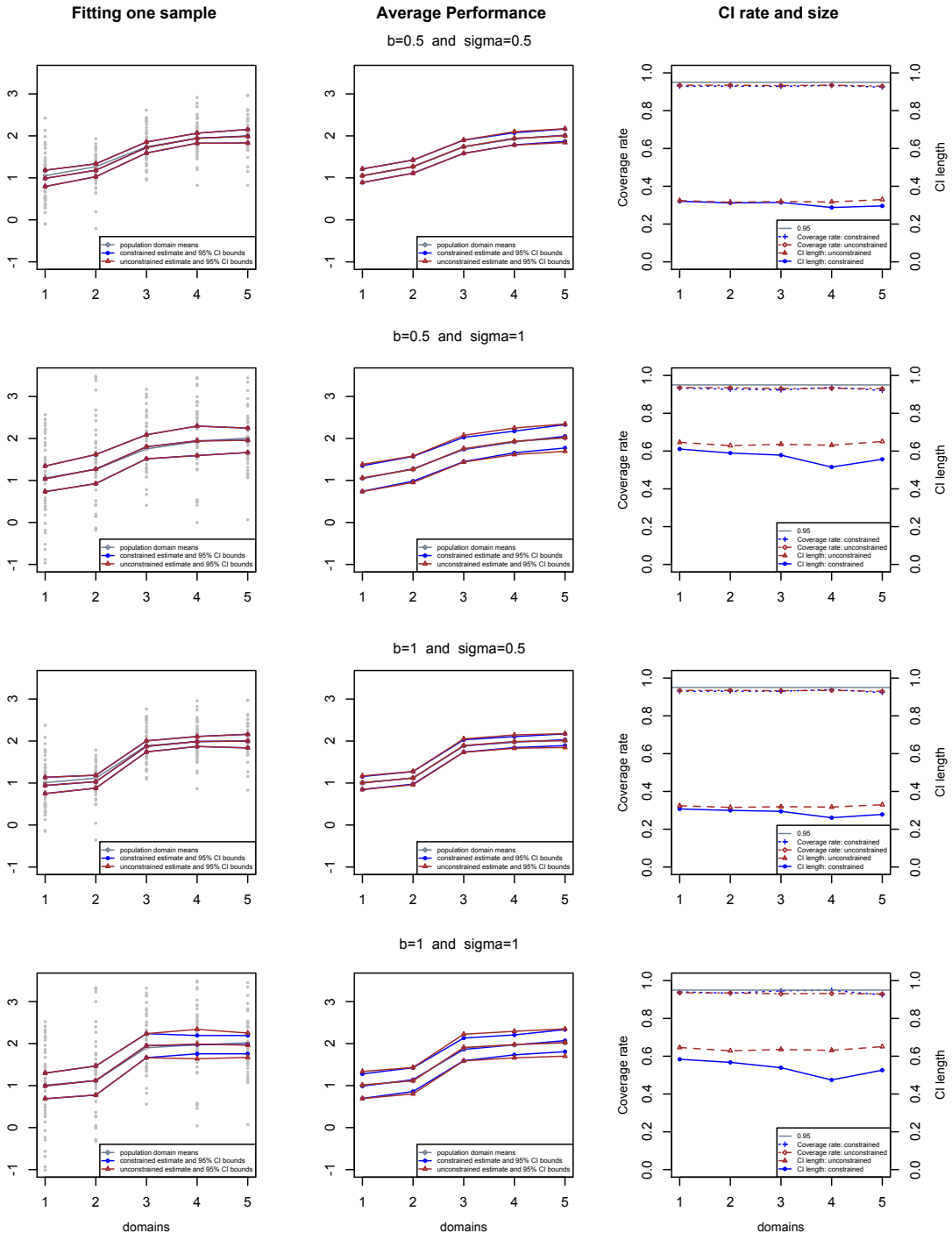

Figure 3.2: Comparisons of the constrained and unconstrained fit for a typical sample, as well as the average performance over 50000 replicate samples for $\mu_1(t) = \exp(20bt/T - 10b)/(1 + \exp(20bt/T - 10b))$ with $T = 5$ and $n = 200$.

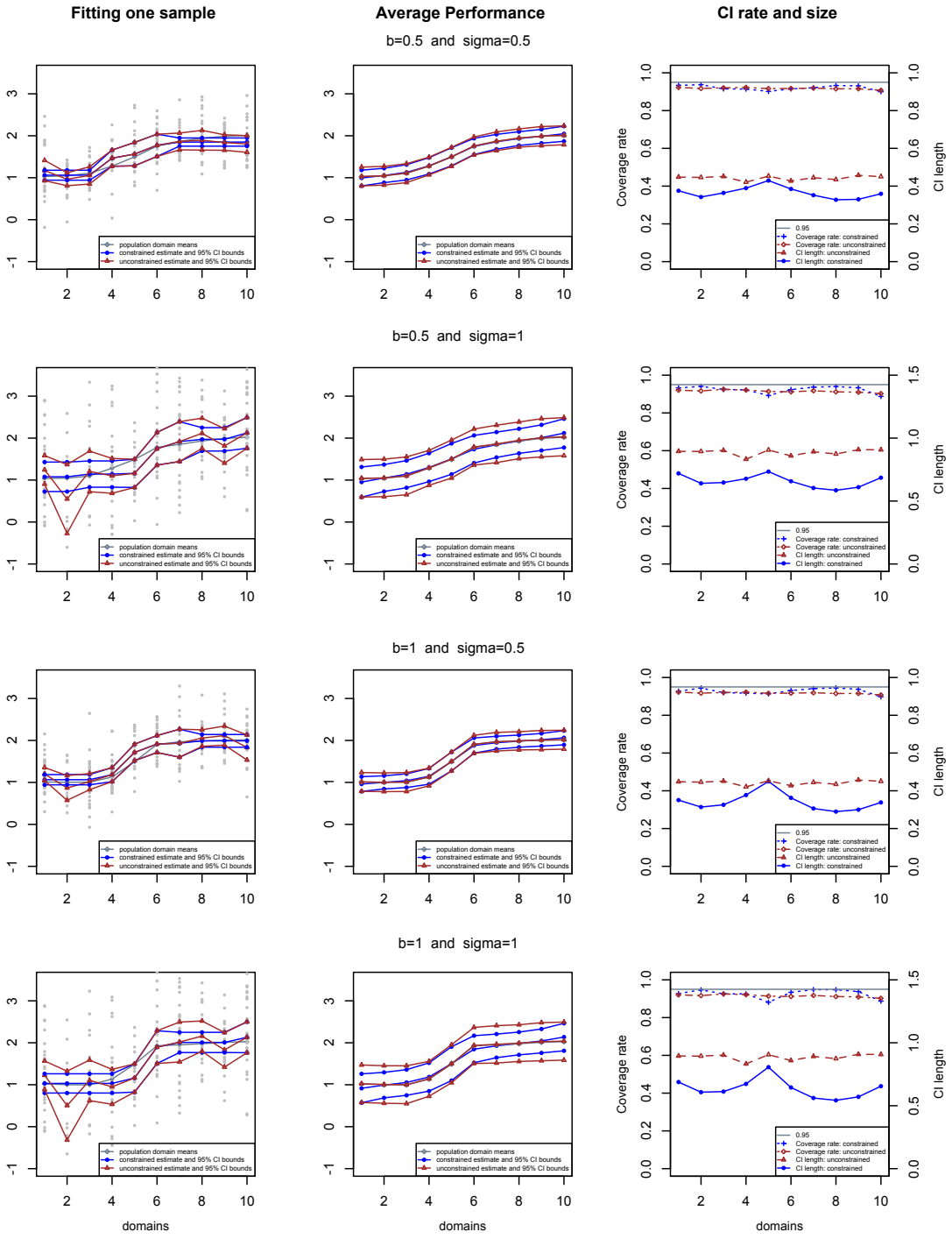

Figure 3.3: Comparisons of the constrained and unconstrained fit for a typical sample, as well as the average performance over 50000 replicate samples for $\mu_1(t) = \exp(20bt/T - 10b)/(1 + \exp(20bt/T - 10b))$ with $T = 10$ and $n = 200$.

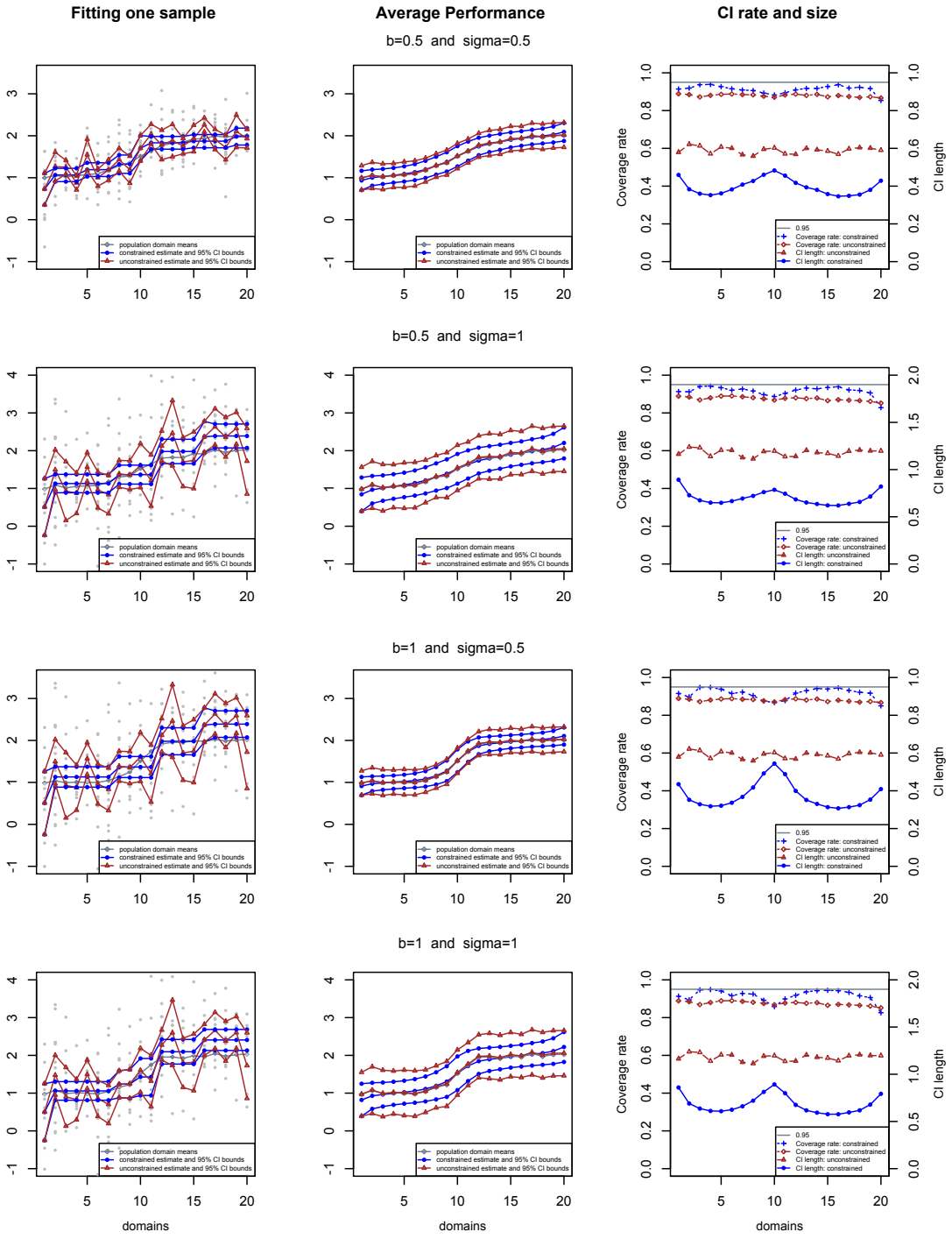

Figure 3.4: Comparisons of the constrained and unconstrained fit for a typical sample, as well as the average performance over 50000 replicate samples for $\mu_1(t) = \exp(20bt/T - 10b)/(1 + \exp(20bt/T - 10b))$ with $T = 20$ and $n = 200$.

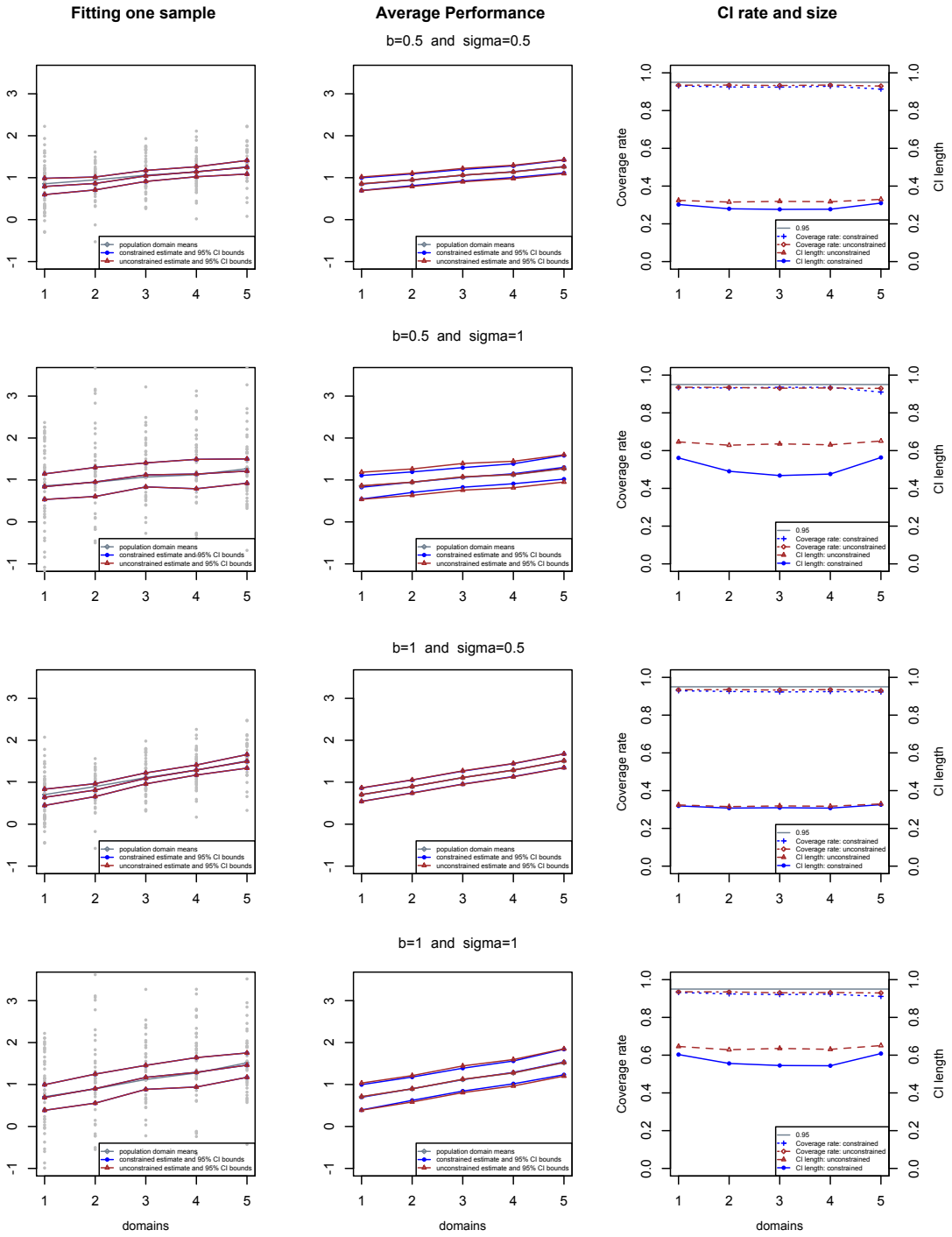

Figure 3.5: Comparisons of the constrained and unconstrained fit for a typical sample, as well as the average performance over 50000 replicate samples for $\mu_2(t) = 1 + b(t/T - 0.5)$ with $T = 5$ and $n = 200$.

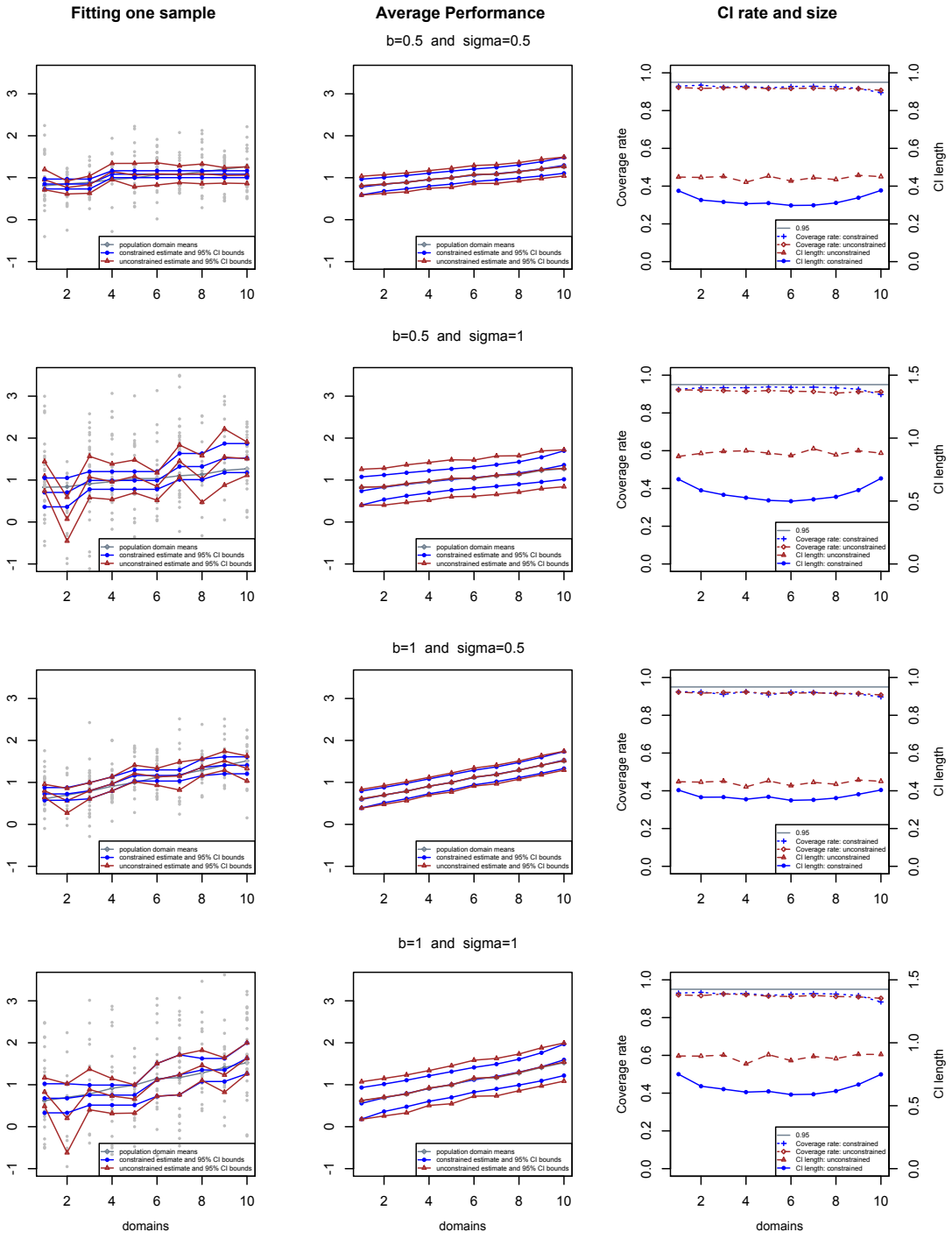

Figure 3.6: Comparisons of the constrained and unconstrained fit for a typical sample, as well as the average performance over 50000 replicate samples for $\mu_2(t) = 1 + b(t/T - 0.5)$ with $T = 10$ and $n = 200$.

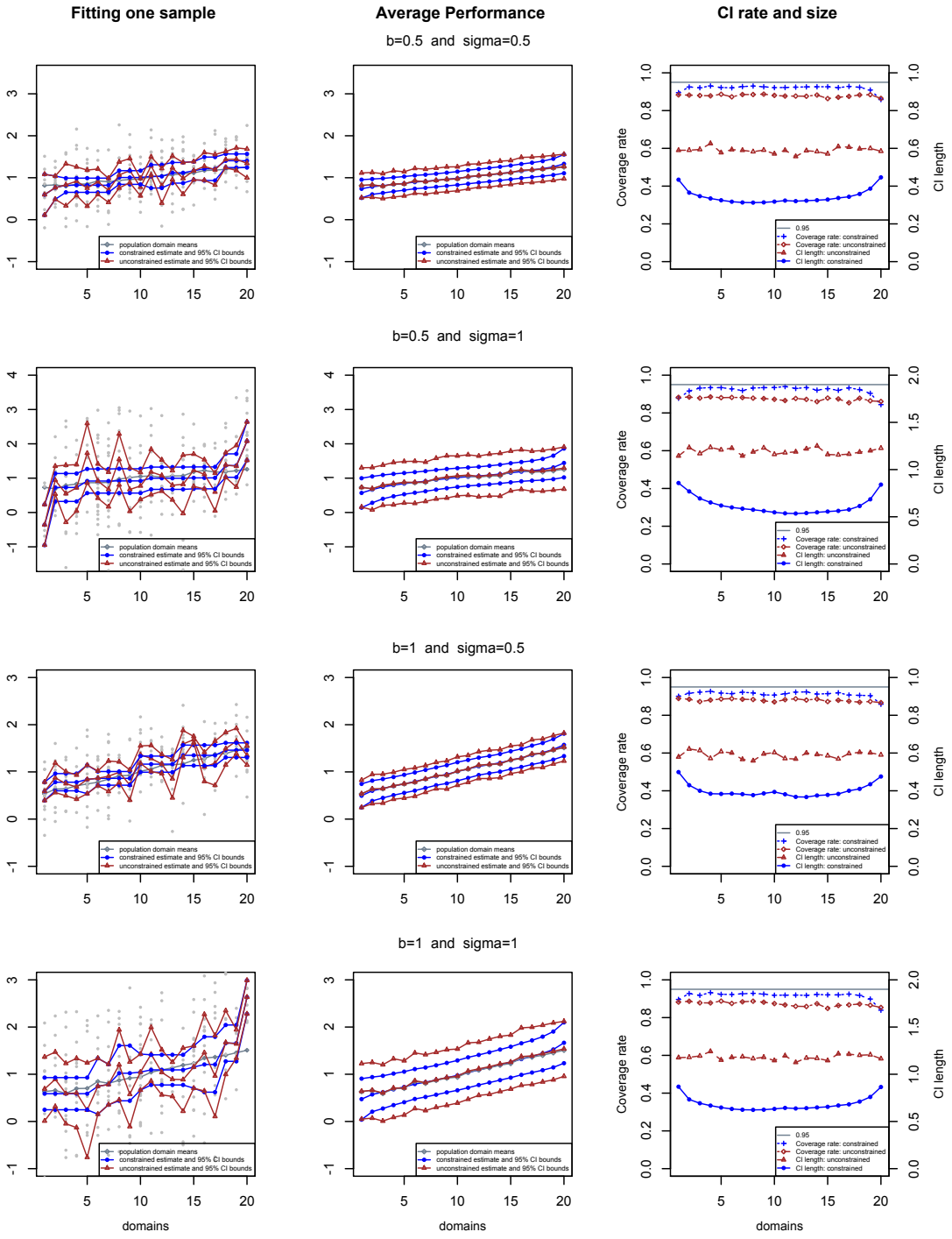

Figure 3.7: Comparisons of the constrained and unconstrained fit for a typical sample, as well as the average performance over 50000 replicate samples for $\mu_2(t) = 1 + b(t/T - 0.5)$ with $T = 20$ and $n = 200$.

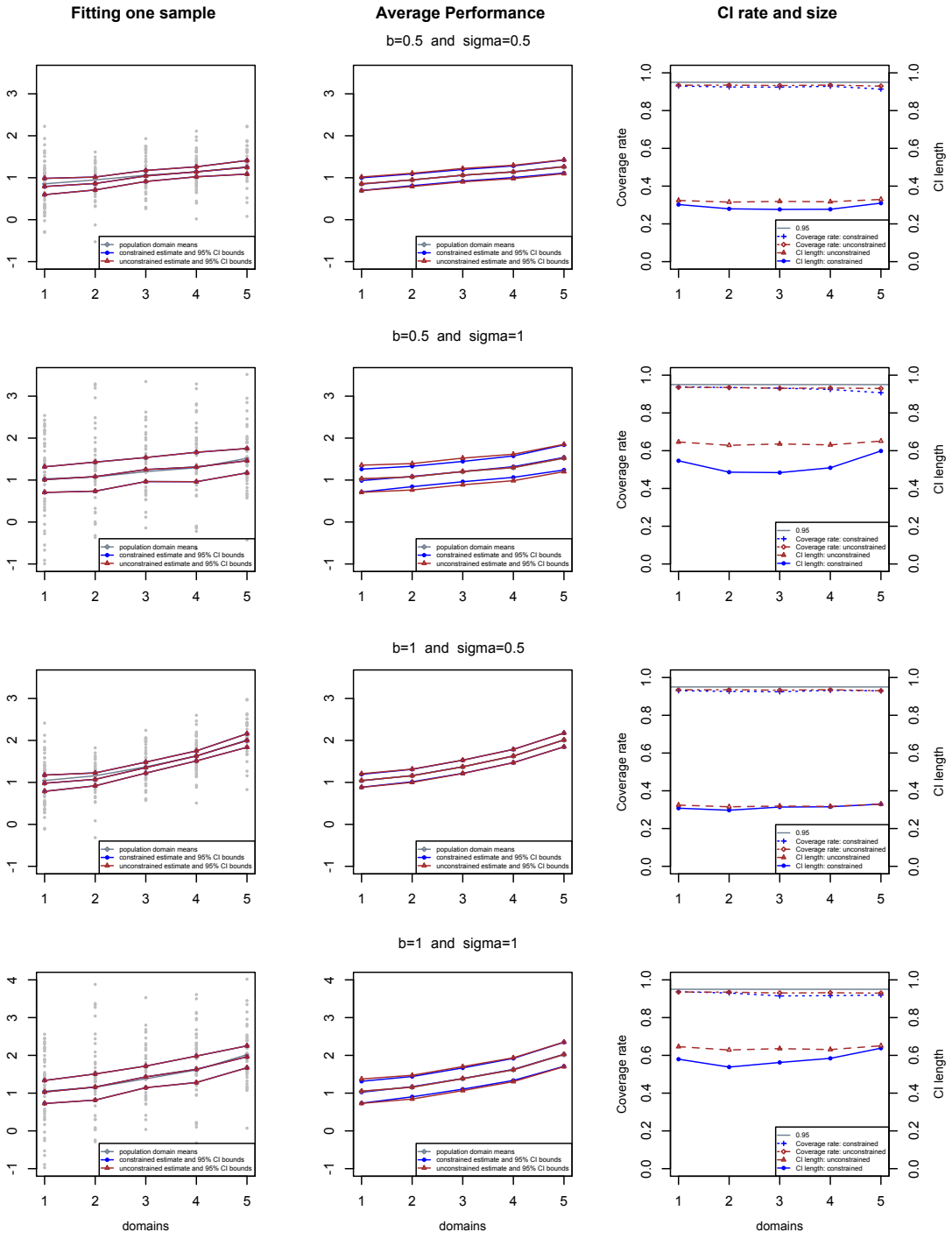

Figure 3.8: Comparisons of the constrained and unconstrained fit for a typical sample, as well as the average performance over 50000 replicate samples for $\mu_3(t) = 1 + b(t/T)^2$ with $T = 5$ and $n = 200$.

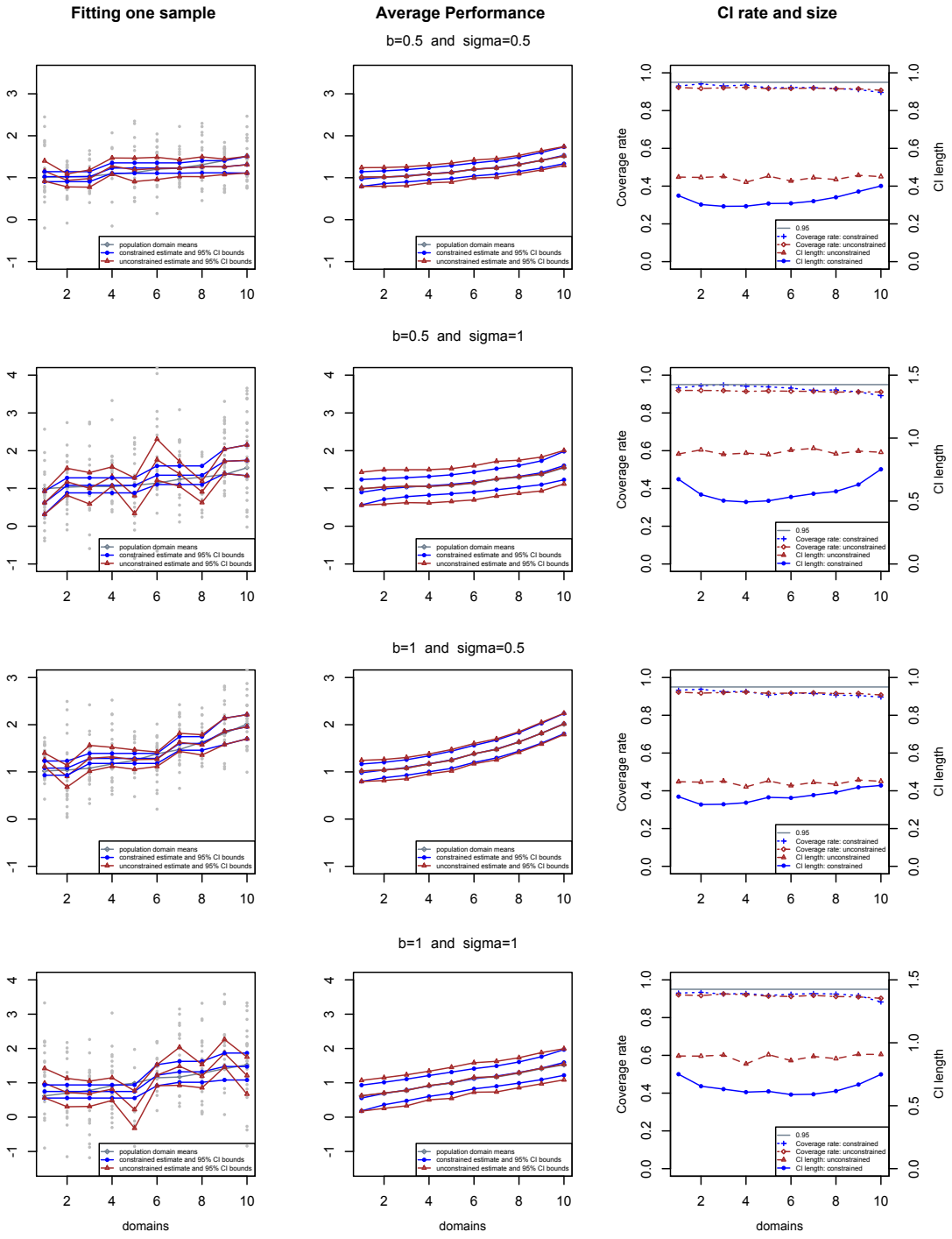

Figure 3.9: Comparisons of the constrained and unconstrained fit for a typical sample, as well as the average performance over 50000 replicate samples for $\mu_3(t) = 1 + b(t/T)^2$ with $T = 10$ and $n = 200$.

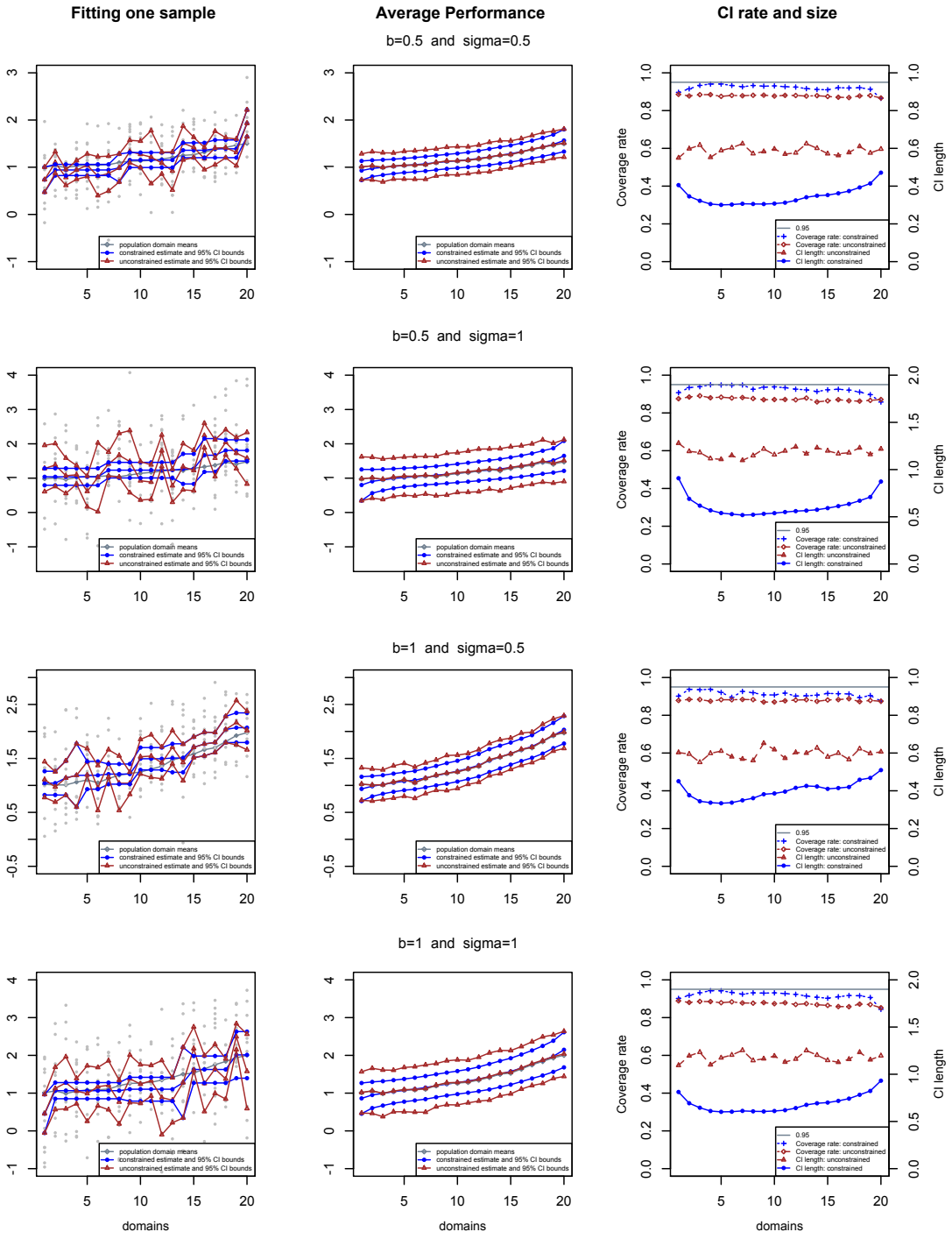

Figure 3.10: Comparisons of the constrained and unconstrained fit for a typical sample, as well as the average performance over 50000 replicate samples for $\mu_3(t) = 1 + b(t/T)^2$ with $T = 20$ and $n = 200$.

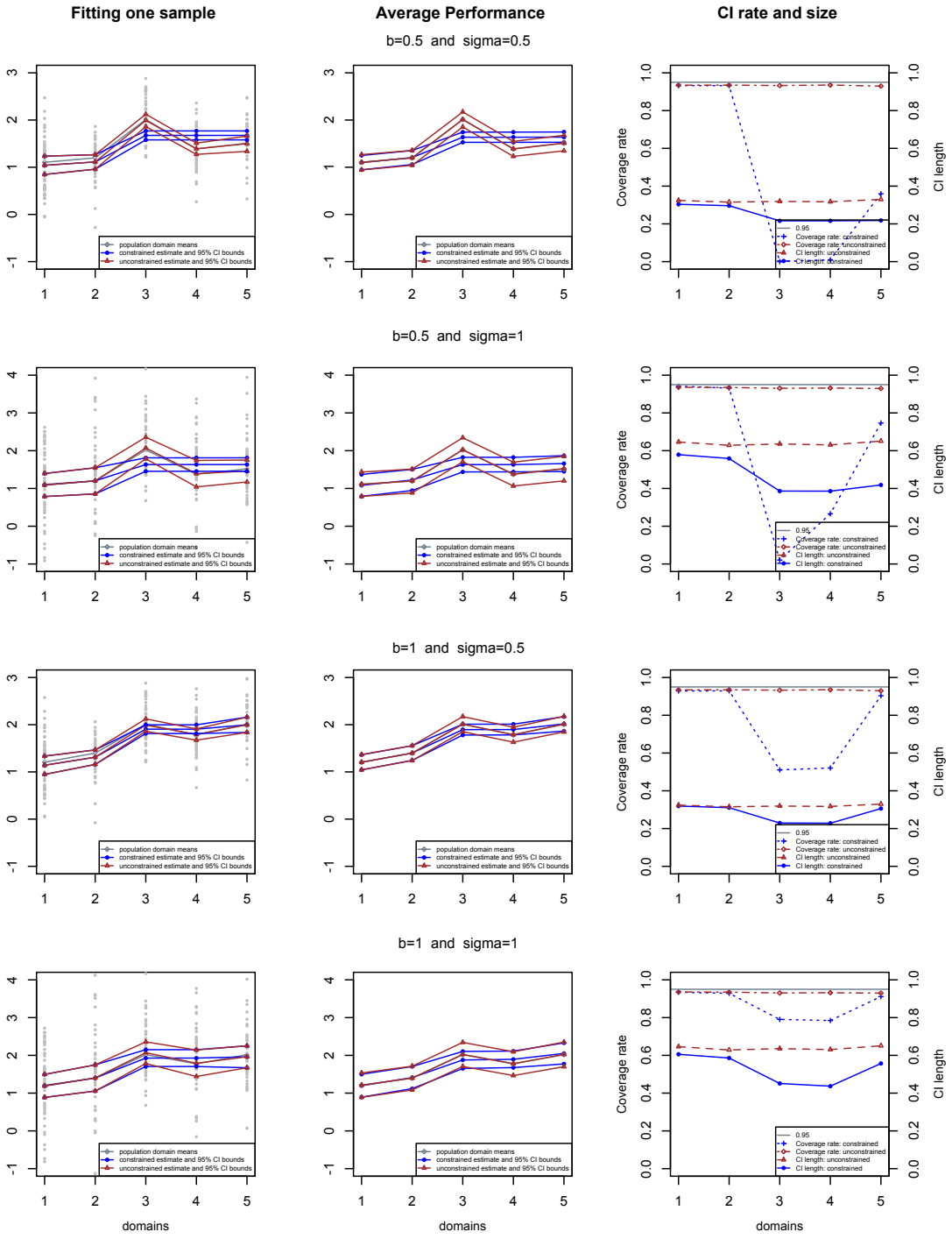

Figure 3.11: Comparisons of the constrained and unconstrained fit for a typical sample, as well as the average performance over 50000 replicate samples for $\mu_4(t) = 1 + b(t/T)I_{\{t/T<0.5 \text{ or } t/T>0.7\}} + I_{\{0.5<t/T\leq 0.7\}}$ with $T = 5$ and $n = 200$.

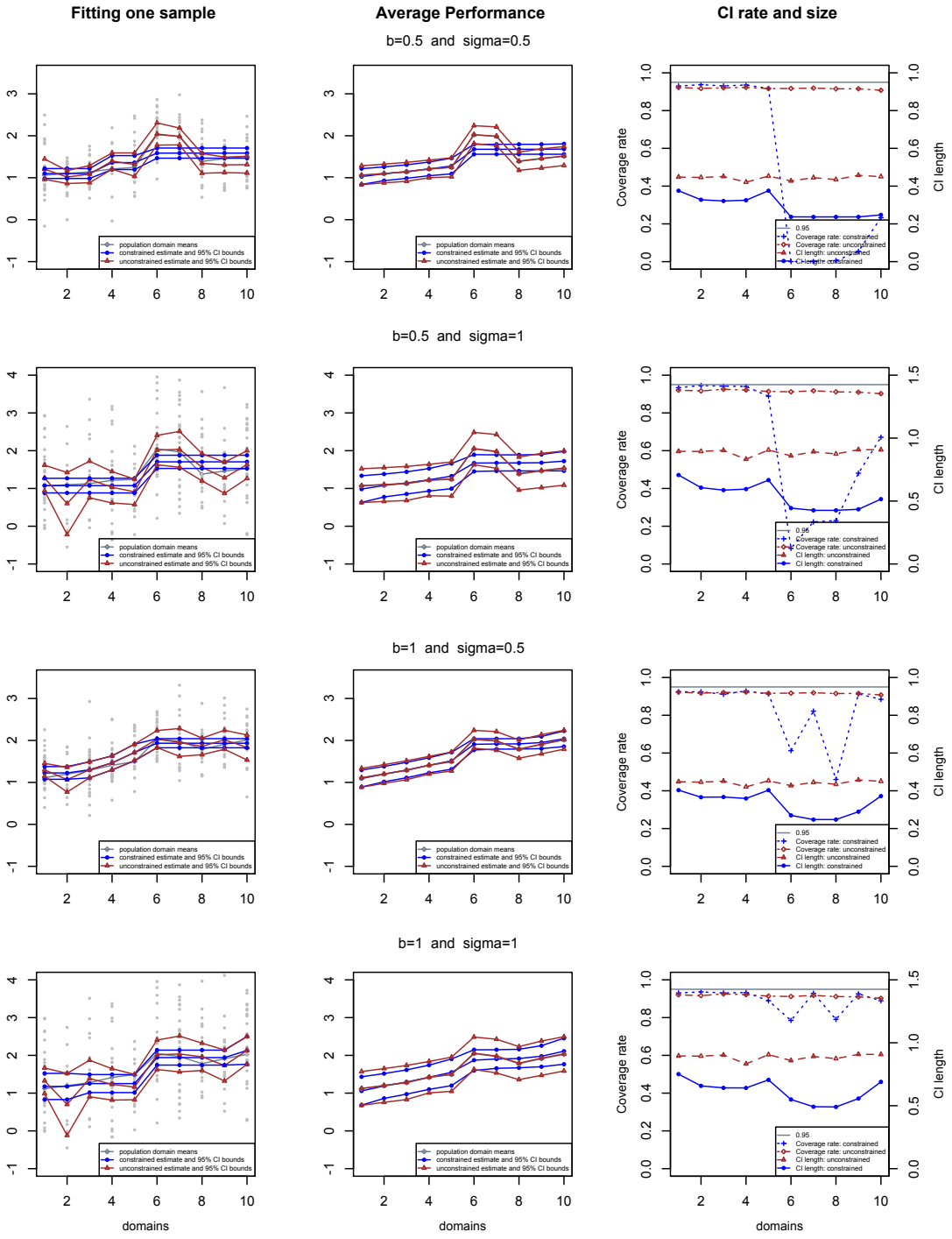

Figure 3.12: Comparisons of the constrained and unconstrained fit for a typical sample, as well as the average performance over 50000 replicate samples for $\mu_4(t) = 1 + b(t/T)I_{\{t/T<0.5 \text{ or } t/T>0.7\}} + I_{\{0.5<t/T\leq 0.7\}}$ with $T = 10$ and $n = 200$.

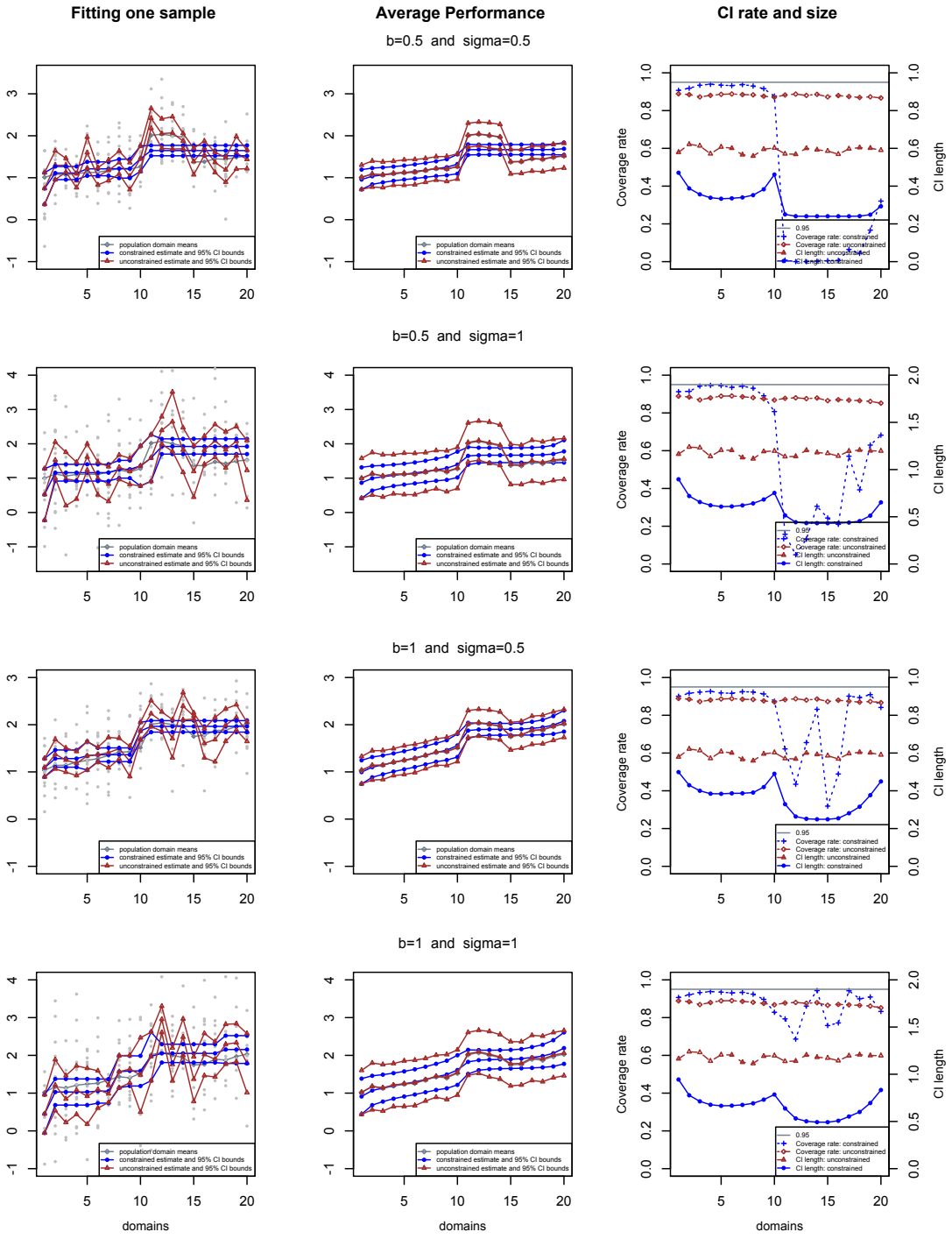

Figure 3.13: Comparisons of the constrained and unconstrained fit for a typical sample, as well as the average performance over 50000 replicate samples for $\mu_4(t) = 1 + b(t/T)I_{\{t/T<0.5 \text{ or } t/T>0.7\}} + I_{\{0.5<t/T\leq 0.7\}}$ with $T = 20$ and $n = 200$.

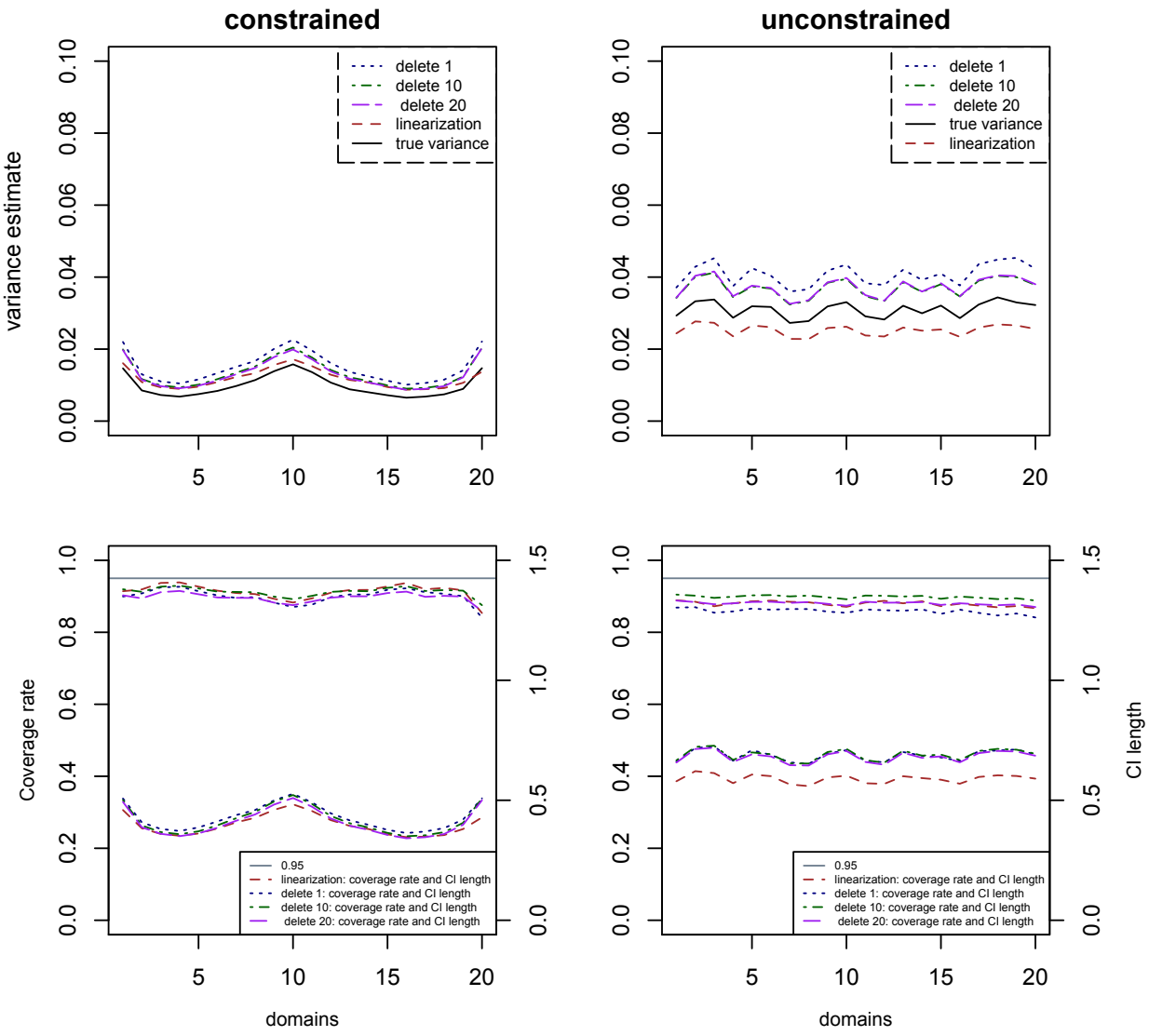

Figure 3.14: Comparison of variance estimate, coverage rates, and confidence interval length for $\mu_1(t) = \exp(20bt/T - 10b)/(1 + \exp(20bt/T - 10b))$ with $T = 20$, $n = 200$, $b = 0.5$ and $\sigma = 0.5$.

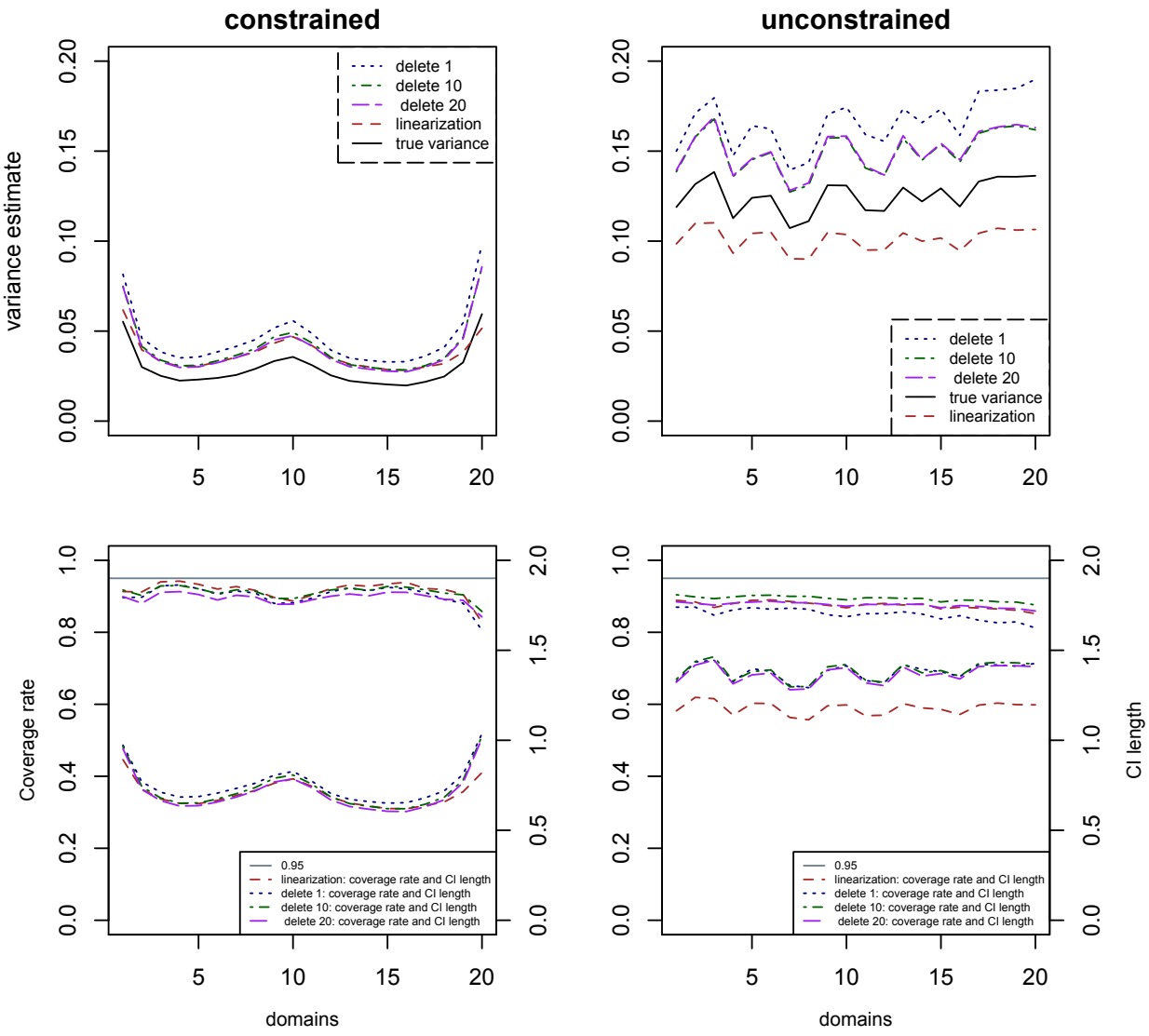

Figure 3.15: Comparison of variance estimate, coverage rates, and confidence interval length for $\mu_1(t) = \exp(20bt/T - 10b)/(1 + \exp(20bt/T - 10b))$ with $T = 20$, $n = 200$, $b = 0.5$ and $\sigma = 1$.

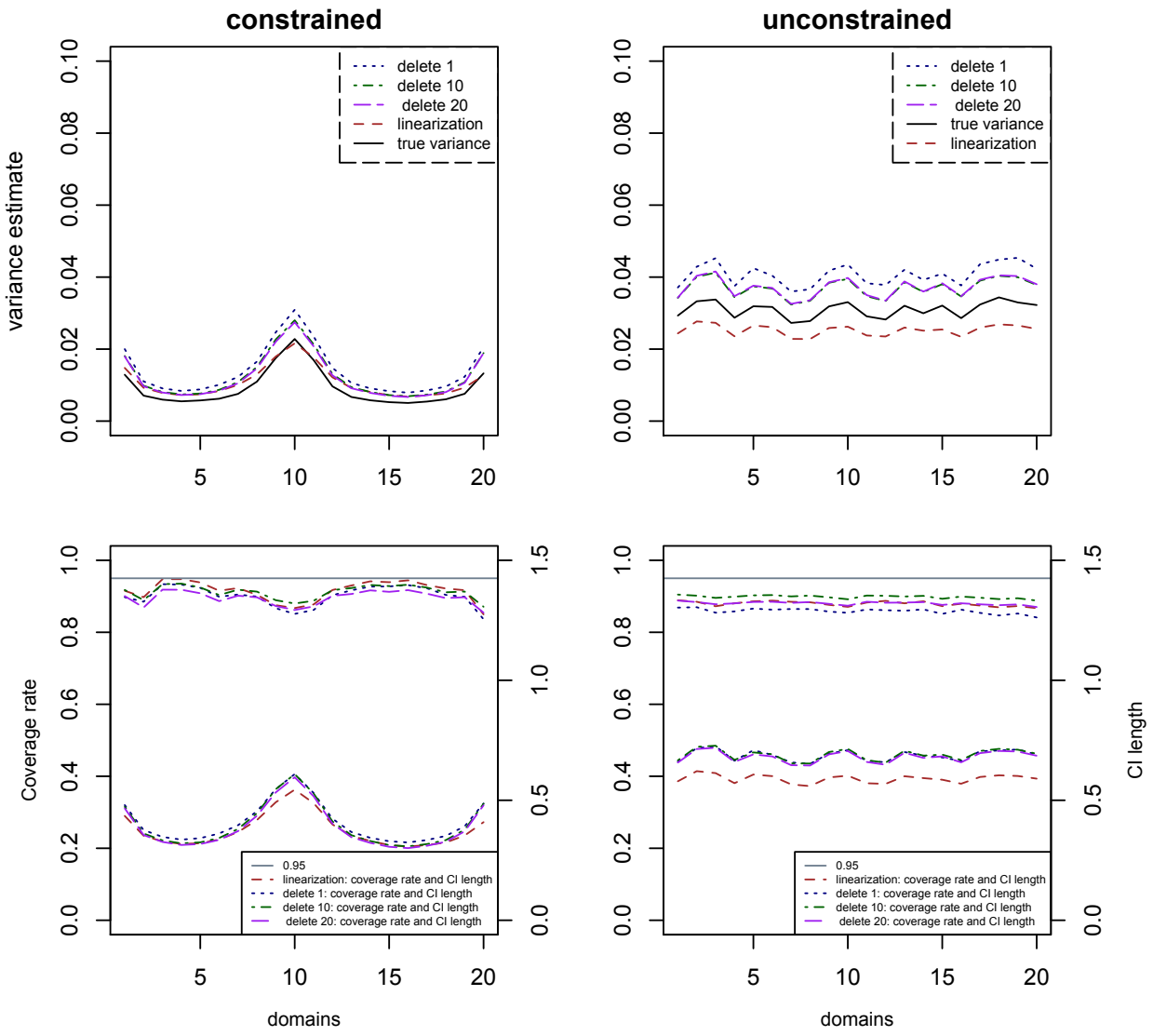

Figure 3.16: Comparison of variance estimate, coverage rates, and confidence interval length for $\mu_1(t) = \exp(20bt/T - 10b)/(1 + \exp(20bt/T - 10b))$ with $T = 20$, $n = 200$, $b = 1$ and $\sigma = 0.5$.

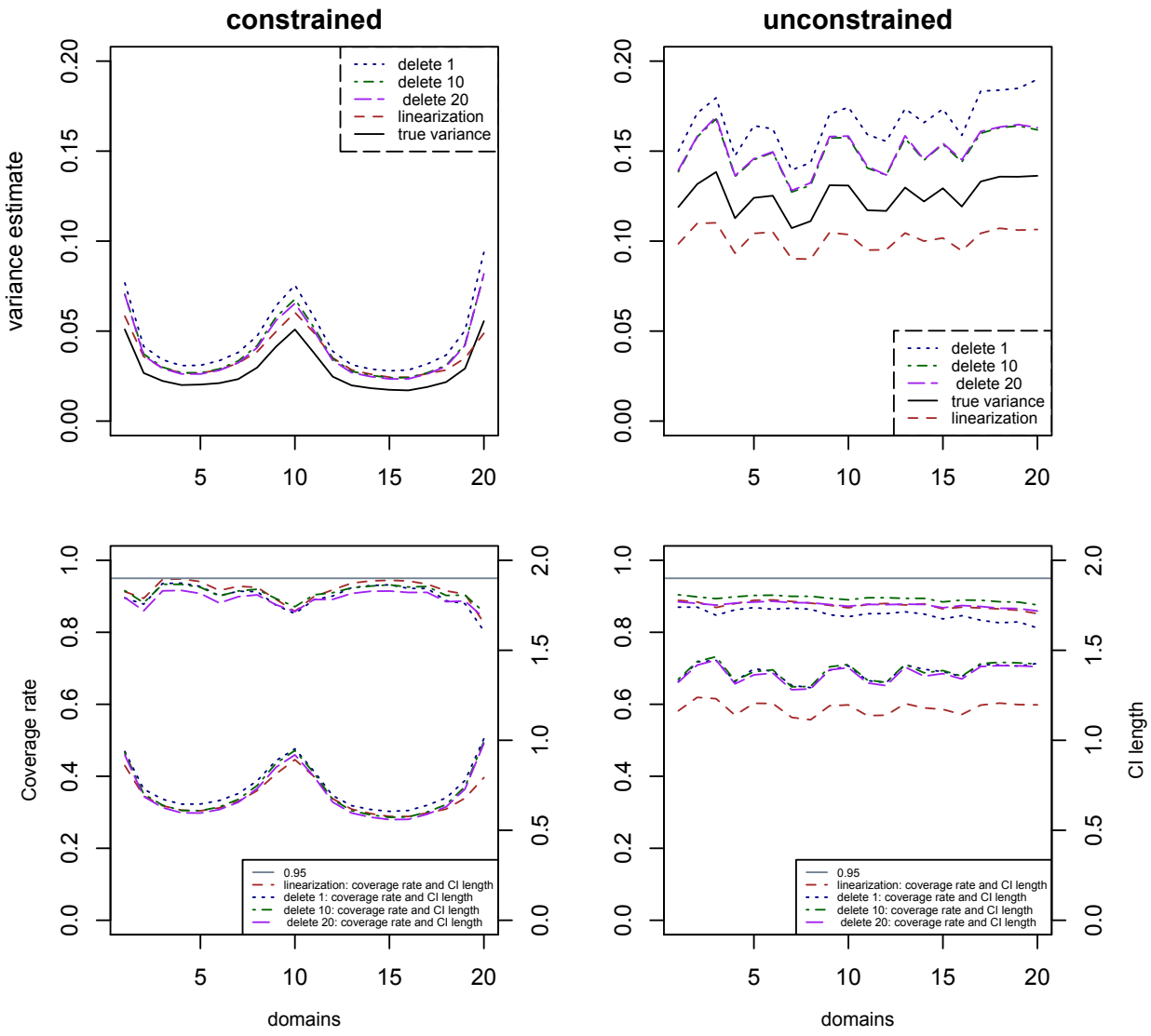

Figure 3.17: Comparison of variance estimate, coverage rates, and confidence interval length for $\mu_1(t) = \exp(20bt/T - 10b)/(1 + \exp(20bt/T - 10b))$ with $T = 20$, $n = 200$, $b = 1$ and $\sigma = 1$.

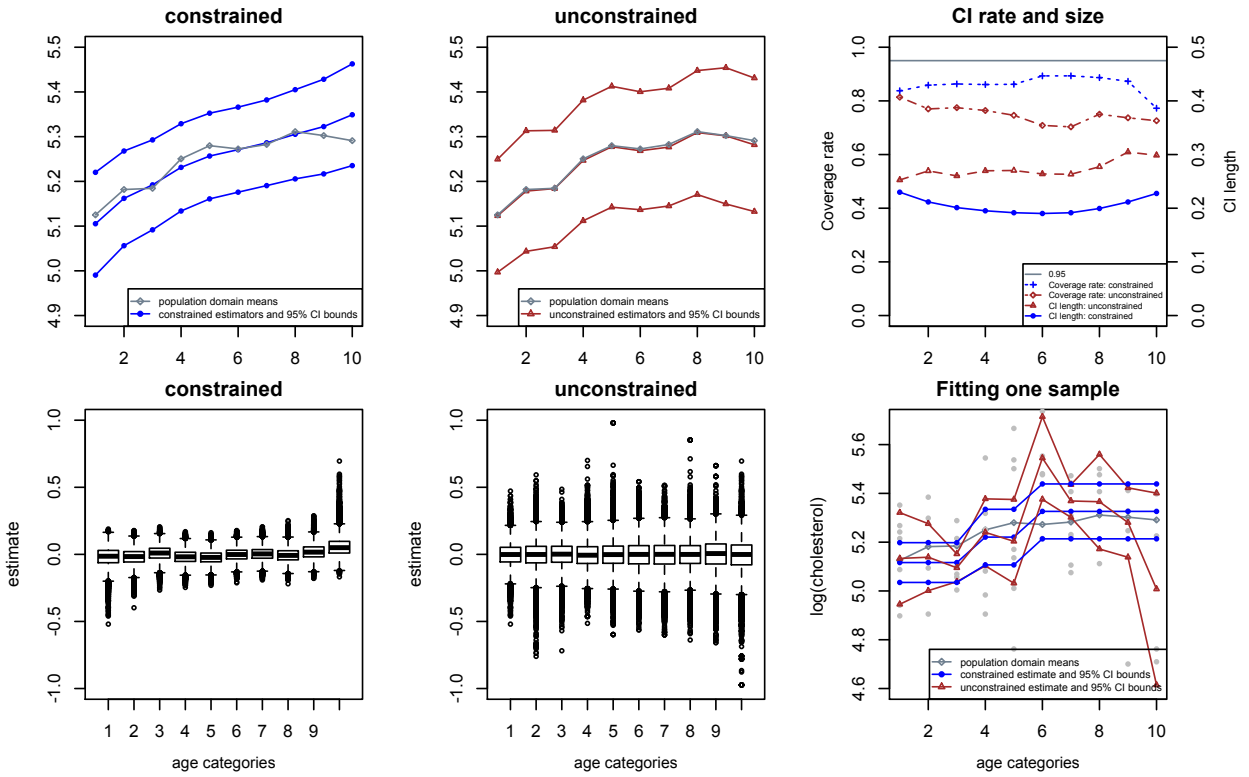

Figure 3.18: NHANES data example: comparison of the average performance of constrained and unconstrained estimation with $n = 60$ based on 50000 replicate samples and fitting constrained and unconstrained estimation on one sample in the last plot.

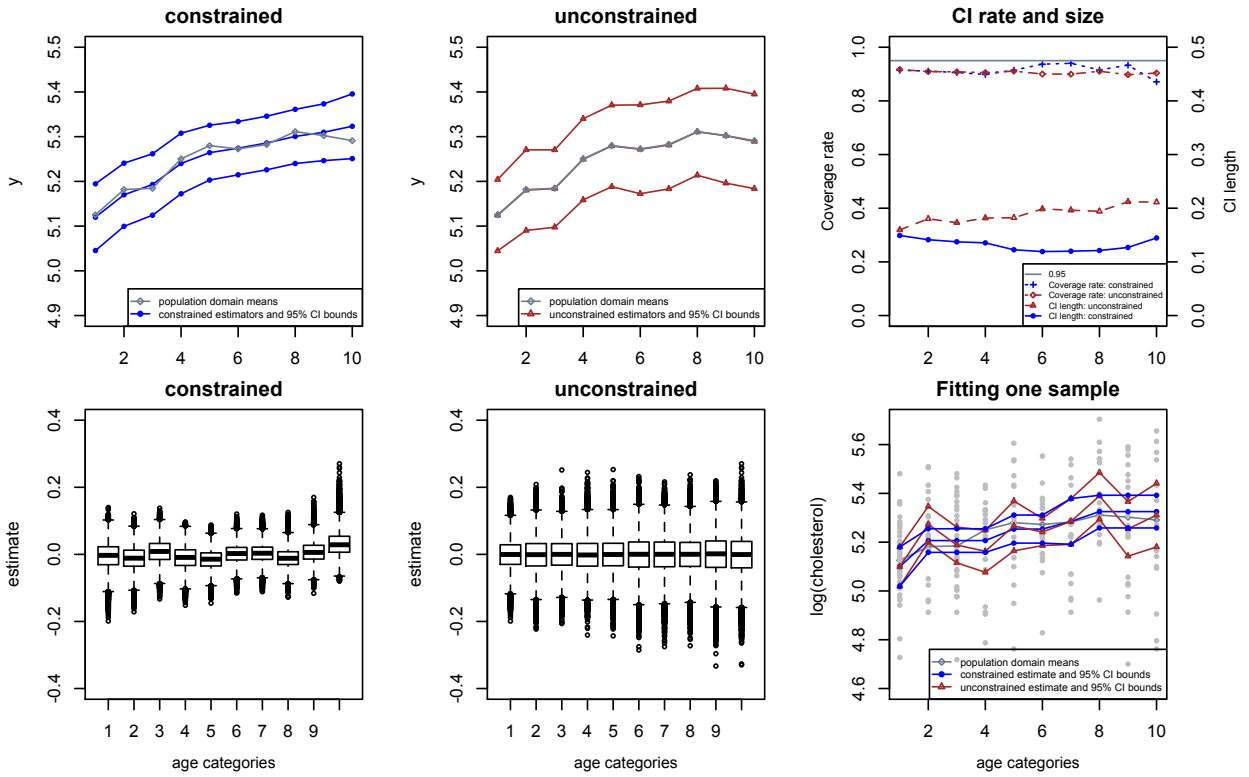

Figure 3.19: NHANES data example: comparison of the average performance of constrained and unconstrained estimation with $n = 210$ based on 50000 replicate samples and fitting constrained and unconstrained estimation on one sample in the last plot.

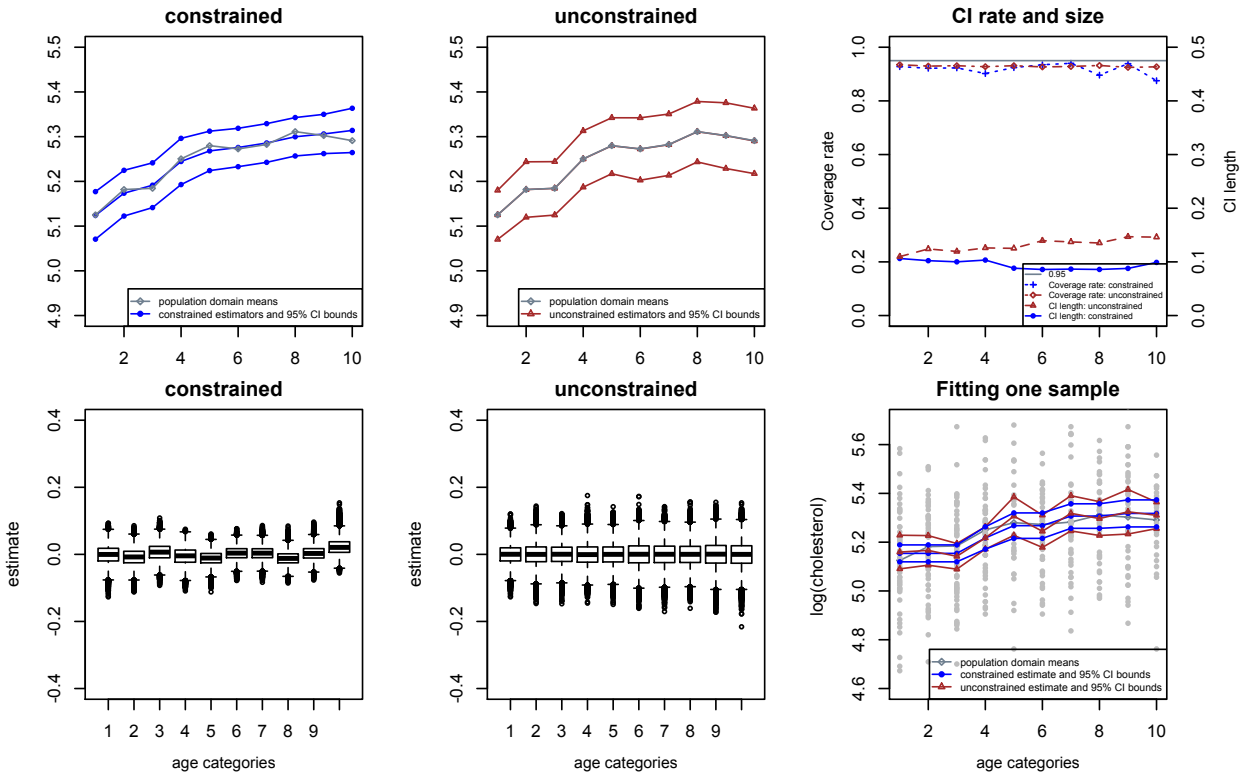

Figure 3.20: NHANES data example: comparison of the average performance of constrained and unconstrained estimation with $n = 450$ based on 50000 replicate samples and fitting constrained and unconstrained estimation on one sample in the last plot.

CHAPTER 4

ISOTONIC REGRESSION IN SURVEY SAMPLING

4.1 Introduction

Consider a finite population $U_N = \{1, \ldots, k, \ldots, N\}$ with a study variable y and two categorical covariates x_1 and x_2. Suppose it is reasonable to assume that the means of y values are non-decreasing in both covariates. This implies a partial ordering. For two points $\boldsymbol{a} = (a_1, a_2)$ and $\boldsymbol{b} = (b_1, b_2)$, a partial ordering is defined as $\boldsymbol{a} \preceq \boldsymbol{b}$ if $a_1 \leq b_1$ and $a_2 \leq b_2$. It is called a partial ordering since not all pairs of points are comparable. Suppose the covariates x_1 and x_2 have t_1 and t_2 levels respectively, then it is assuming that y has a partial ordering on a $t_1 \times t_2$ grids, and each combination of the levels of x_1 and x_2 can be considered as a 'cell' or, as more commonly referred to in the survey context, a domain. While one could maintain the double indices in what follows, we will instead switch to a single index $t = 1, \ldots, T$ to denote the domains. Partial ordering over domains defined by two ordinal covariates, and indeed for any other monotonicity arrangement across domains, will be presented by a constraint matrix, so that the indexing is not necessary to describe the relationships between domains. Let $U_{t,N}$ to denote a domain with N_t elements. Define the indicator variable for domains as $z_{tk} = 1$ if $k \in U_{t,N}$ and $z_{tk} = 0$ otherwise. Our goal is to estimate the population domain means

$$\bar{y}_{U_t,N} = \frac{\sum_{k \in U_{t,N}} y_k}{N_t} = \frac{\sum_{k \in U_N} z_{tk} y_k}{\sum_{k \in U_N} z_{tk}}, \quad t = 1, \ldots, T \tag{17}$$

with the assumed partial ordering.

While we will be interested in sample based estimation, we first introduce the basic concepts of isotonic regression at the level of the population, i.e. as if the full population

were observed. Sample-based estimation will be discussed in section 4.2. The isotonized population means minimize

$$\sum_{t=1}^{T} \frac{N_t}{N}(\bar{y}_{U_t} - \theta_t)^2, \text{ subject to } \boldsymbol{A\theta} \geq 0 \tag{18}$$

where $\boldsymbol{A}$ is the constraint matrix. Each row of $\boldsymbol{A}$ represent a constraint ordering between two comparable points, and the constraint matrix $\boldsymbol{A}$ is set to be irreducible such that none of its rows is a positive linear combination of two of more other rows, and the zero vector is not positive linear combination of two of more rows (Meyer, 1999).

As noted above, the constraint matrix $\boldsymbol{A}$ can not only represent a partial ordering on grids, but any other ordering between domains, including a complete ordering on one dimension, a constrained ordering on one out of two dimensions as shown in the following examples.

Example 1: Given a study variable y and one covariate x with T levels, the constraint assumption implies a complete ordering $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_T$. The vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_T)^\top$ and the constraint matrix $\boldsymbol{A}$ is a $(T-1) \times T$ matrix with entries $A_{t,t} = -1$ and $A_{t,t+1}$ for $t = 1, \ldots, T-1$. (This special case was the topic of Chapter 3.)

Example 2: For a non-decreasing constraint on two covariates, we can describe the vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_T)^\top = (\theta_{11}, \theta_{12}, \ldots, \theta_{1t_2}, \ldots, \theta_{t_1 1}, \ldots, \theta_{t_1 t_2})^\top$ and $\boldsymbol{A}$ is a $m \times T$ matrix with $m = 2t_1t_2 - t_1 - t_2$ and $T = t_1t_2$. For a partial ordering on a 2×3 grids, the constraint matrix is

$$\boldsymbol{A} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \\ -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 \end{pmatrix}$$

Example 3: Still on a 2×3 grids and assume that the mean of y values are non decreasing in the covariate with 3 levels, then the constraint matrix becomes

$$\boldsymbol{A} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix}$$

Finally, we will also consider 'relaxed monotone' constraints such that for $t_- \leq t_+$

$$\frac{\sum_{t=1}^{T} K\left(\frac{t-t_-}{h}\right)\theta_t}{\sum_{t=1}^{T} K\left(\frac{t-t_-}{h}\right)} \leq \frac{\sum_{t=1}^{T} K\left(\frac{t-t_+}{h}\right)\theta_t}{\sum_{t=1}^{T} K\left(\frac{t-t_+}{h}\right)} \tag{19}$$

where K can be viewed as a weight function and h is a bandwidth. If K is such that $K(0) = 1$ and $K(t) = 0$ for $t \neq 0$, then it reverts to the complete ordering constraints $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_T$. We recommend $K(x) = exp\{-|x|\}$ such that closer points have higher weights, but other strictly decreasing kernels can readily be used here. The bandwidth h will also affect the constraint: with smaller bandwidth h, the constrained estimate will be closer to monotonicity. In the case of relaxed ordering, the constraint matrix no longer contains

only -1, 1 and 0. For one dimension, $\boldsymbol{A}$ is a $(T-1)\times T$ matrix with entries

$$A_{ij} = \frac{K\left(\frac{j+1-i}{h}\right)}{\sum_{t=1}^{T} K\left(\frac{j+1-t}{h}\right)} - \frac{K\left(\frac{j-i}{h}\right)}{\sum_{t=1}^{T} K\left(\frac{j-t}{h}\right)}. \tag{20}$$

Furthermore, when there are two covariates and the domains are arranged in a grid, the relaxed monotone constraints can be imposed to either or both dimensions.

Example 4: For one dimension with $T = 10$ and set the bandwidth $h = 0.5$, then the constraint matrix is

$$\boldsymbol{A} = \begin{pmatrix}
-0.76 & 0.66 & 0.09 & 0.01 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
-0.09 & -0.67 & 0.66 & 0.09 & 0.01 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
-0.01 & -0.09 & -0.66 & 0.66 & 0.09 & 0.01 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & -0.01 & -0.09 & -0.66 & 0.66 & 0.09 & 0.01 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & -0.01 & -0.09 & -0.66 & 0.66 & 0.09 & 0.01 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & -0.01 & -0.09 & -0.66 & 0.66 & 0.09 & 0.01 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & -0.01 & -0.09 & -0.66 & 0.66 & 0.09 & 0.01 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & -0.01 & -0.09 & -0.66 & 0.67 & 0.09 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & -0.01 & -0.09 & -0.66 & 0.76
\end{pmatrix}$$

where small values are rounded as 0.00.

Therefore, (18) can be used to describe a general case of shape regression problem, where the constraint matrix $\boldsymbol{A}$ represents the shape restriction on θ_t.

4.2 Method and Theory

Let $\boldsymbol{W} = diag\{N_t/N\}$; then minimizing (18) is the same as minimizing $\|\boldsymbol{z_U} - \boldsymbol{\phi}\|^2$ subject to $\tilde{\boldsymbol{A}}\boldsymbol{\phi} \geq 0$ where $\boldsymbol{z_U} = \boldsymbol{W}^{1/2}\bar{\boldsymbol{y}}_{\boldsymbol{U}}$, $\tilde{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{W}^{-1/2}$, and $\boldsymbol{\phi} = \boldsymbol{W}^{1/2}\boldsymbol{\theta}$. The constraint set $\mathcal{C} = \{\boldsymbol{\phi} \in \mathbb{R}^T : \tilde{\boldsymbol{A}}\boldsymbol{\phi} \geq 0\}$ is a closed convex cone, and the minimizer $\hat{\boldsymbol{\phi}}$ is called the projection of $\boldsymbol{z}$ onto the constraint set $\mathcal{C}$.

We first review the results of the constraint cone and its polar cone that will be used through the paper; for detailed discussion please see Meyer (1999). The polar cone of $\mathcal{C}$ is defined as $\boldsymbol{\Omega}^0 = \{\boldsymbol{\rho} \in \mathbb{R}^T : \langle \boldsymbol{\phi}, \boldsymbol{\rho} \rangle \leq 0, \text{ for all } \boldsymbol{\phi} \in \mathcal{C}\}$. A vector $\boldsymbol{\gamma}$ is called an edge of the cone $\boldsymbol{\Omega}^0$ if it can not be written as a sum of two linearly independent vectors contained in the cone. The constraint matrix $\tilde{\boldsymbol{A}}$ is irreducible if none of its rows is a positive linear combination of two of more other rows, and the zero vector is not a positive linear combination of two of more rows. Let $\boldsymbol{\gamma_j}$ be rows of $-\tilde{\boldsymbol{A}}$, then $\boldsymbol{\gamma_j}$ are edges of the polar cone $\boldsymbol{\Omega}^0$ when $\tilde{\boldsymbol{A}}$ is irreducible and they generate the cone so that $\boldsymbol{\Omega}^0 = \{\boldsymbol{\rho} \in \mathbb{R}^T : \boldsymbol{\rho} = \sum_{j=1}^m b_j \gamma_j, b_j \geq 0, j = 1, \ldots, m\}$.

Let m_1 be the row rank of $\tilde{\boldsymbol{A}}$, then $m \geq m_1$ in the case of a partial ordering. Let $\boldsymbol{V}$ be the null space of $\tilde{\boldsymbol{A}}$ and $\boldsymbol{\Omega} = \mathcal{C} \cap \boldsymbol{V}^\perp$, then $\boldsymbol{\Omega}$ is a closed convex cone that does not contain a linear space of dimension one or greater. Let $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{m_2}$ be edges of $\boldsymbol{\Omega}$ where m_2 can be large when $\tilde{\boldsymbol{A}}$ is not full row-rank, then the edges of the constraint cone $\mathcal{C}$ can be found and $\mathcal{C}$ can be written as $\mathcal{C} = \{\boldsymbol{\phi} \in \mathbb{R}^T : \boldsymbol{\phi} = \boldsymbol{v} + \sum_{j=1}^{m_2} b_j \delta_j, \text{ where } b_j \geq 0, j = 1, \ldots, m_2, \text{ and } \boldsymbol{v} \in \boldsymbol{V}\}$.

Each set $J \subseteq \{1, \ldots, m\}$ determines a sector $\mathcal{C}_J = \{\boldsymbol{v} + \sum_{j \in J} a_j \boldsymbol{\gamma}_j + \sum_{i \in I} b_i \boldsymbol{\delta}_i : a_j > 0, j \in J, \text{ and } i \in I\}$ where $I = I(J) = \{i : \langle \boldsymbol{\gamma}_j, \boldsymbol{\delta}_i \rangle = 0, \text{ for all } j \in J\}$. If the constraint matrix is full row-rank, then these sectors are disjoint and they cover $\mathbb{R}^T$; for not full row-rank case, they can overlap. For all $\boldsymbol{y} \in \mathbb{R}^T$, there is a $J \subseteq \{1, \ldots, m\}$ such that $\boldsymbol{y} = \boldsymbol{v} + \sum_{j \in J} a_j \boldsymbol{\gamma}_j + \sum_{i \in I} b_i \boldsymbol{\delta}_i$ with $a_j > 0, j \in J, \text{ and } i \in I$. By Proposition 4 of Meyer (1999), the projection of $\boldsymbol{y}$ onto $\boldsymbol{\Omega}^0$ is the projection of $\boldsymbol{y}$ on the linear space spanned by $\{\boldsymbol{\gamma}_j : j \in J\}$. Since the projection of $\boldsymbol{z}$ onto the constraint cone $\mathcal{C}$ is the residual of the projection of $\boldsymbol{z}$ onto the polar cone $\boldsymbol{\Omega}^0$ (Meyer, 1999), for computational efficiency, we will look at the projection of $\boldsymbol{z}$ onto the polar cone.

A set $J \subseteq \{1, \ldots, m\}$ determines the linear space spanned by $\boldsymbol{\gamma}_j$, $j \in J$ on which $\boldsymbol{z_U}$ lands, and let $\boldsymbol{\eta}$ be the projection of $\boldsymbol{z_U}$ onto this linear space. By Karush-Kuhn-Tucker (KKT) conditions,

$$\langle \boldsymbol{z_U} - \boldsymbol{\eta}, \boldsymbol{\gamma_j} \rangle = 0 \quad \text{for } j \in J$$

$$\langle \boldsymbol{z_U} - \boldsymbol{\eta}, \boldsymbol{\gamma_j} \rangle < 0 \quad \text{for } j \notin J. \tag{21}$$

An assumption will be made on (21) to guarantee that $\boldsymbol{z}$ is in the interior of a sector so that the projection onto the cone is unique.

We now introduce the sampling design and the sample-based estimators. A probability sample s_N is drawn from U_N via a sampling design $p_N(\cdot)$, where $p_N(s_N)$ is the probability of drawing the sample s_N. Let n_N be the sample size, and assume that we know the inclusion probabilities $\pi_{iN} = \mathrm{P}(i \in s_N) = \sum_{i \in s_N} p_N(s_N) > 0$ and $\pi_{ijN} = \mathrm{P}(i, j \in s_N) = \sum_{i,j \in s_N} p_N(s_N) > 0$ for all $i, j \in U_N$. For simplicity of notation, the subscript N will be suppressed. Define the sample membership indicator $I_k = 1$ if $k \in s$ and $I_k = 0$ otherwise. The Horvitz-Thompson estimator $\bar{y}_{s_t} = (\sum_{k \in s_t} y_k/\pi_k)/N_t$ or the often preferred Hájek estimator $\tilde{y}_{s_t} = (\sum_{k \in s_t} y_k/\pi_k)/\sum_{k \in s_t} 1/\pi_k$ can be used to estimate the population domain means without the qualitative constrains.

We first consider using the Horvitz-Thompson estimator $\bar{y}_{s_t}$. To accommodate the assumption of partial ordering, the constrained estimator $(\hat{\theta}_1, \ldots, \hat{\theta}_T)^\top$ is proposed to minimize

$$\sum_{t=1}^{T} \frac{N_t}{N} (\bar{y}_{s_t} - \theta_t)^2, \text{ subject to } \boldsymbol{A\theta} \geq 0. \tag{22}$$

We will assume that the sampling design is such that there is at least one observation in each domain. We rewrite (22) as projecting $\boldsymbol{z_{sHT}} = \boldsymbol{W}^{1/2}\bar{\boldsymbol{y}}_{\boldsymbol{s}}$ on to the cone $\mathcal{C} = \{\boldsymbol{\phi} \in \mathbb{R}^T : \tilde{\boldsymbol{A}}\boldsymbol{\phi} \geq 0\}$ where $\tilde{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{W}^{-1/2}$, $\boldsymbol{\phi} = \boldsymbol{W}^{1/2}\boldsymbol{\theta}$, and $\boldsymbol{W} = diag\{N_t/N\}$. Notice that the cone is the same as the cone in the case of population.

Under assumptions 1-5 from Section 3.2, theoretical results are presented in the following theorems.

Theorem 9. *When minimizing (17), let $\boldsymbol{\gamma_j}$ be rows of $-\tilde{\boldsymbol{A}} = -\boldsymbol{A}\boldsymbol{W}^{-1/2}$ and a set $J \subseteq \{1, \ldots, m\}$ determines the linear space spanned by $\boldsymbol{\gamma}_j$, $j \in J$ on which $\boldsymbol{z_U} = \boldsymbol{W}^{1/2}\bar{\boldsymbol{y}}_{\boldsymbol{U}}$ is landing and let $\boldsymbol{\eta}$ be the projection of $\boldsymbol{z_U}$ onto this linear space. Assume that there exist*

$\epsilon > 0$ *and* $N^* > 0$ *such that for* $N > N^*$

$$\langle \boldsymbol{z_U} - \boldsymbol{\eta}, \boldsymbol{\gamma_j} \rangle < -\epsilon \quad \textit{for } j \notin J. \tag{23}$$

Let A *be the event that the same set* J *determines the linear space spanned by* $\boldsymbol{\gamma_j}$, $j \in J$ *on which* $\boldsymbol{z_{sHT}}$ *lands, then* $I_{\{A^c\}} = o_p(1/\sqrt{n})$.

Proof. Let $\boldsymbol{\eta_{sHT}}$ be the projection of $\boldsymbol{z_{sHT}}$ onto the the linear space spanned by $\boldsymbol{\gamma_j}$, $j \in J$, by KKT conditions, event A is equivalent to

$$\langle \boldsymbol{z_{sHT}} - \boldsymbol{\eta_{sHT}}, \boldsymbol{\gamma_j} \rangle = 0 \quad \text{for } j \in J \tag{24}$$

$$\langle \boldsymbol{z_{sHT}} - \boldsymbol{\eta_{sHT}}, \boldsymbol{\gamma_j} \rangle < 0 \quad \text{for } j \notin J. \tag{25}$$

By orthogonality of projection, (24) holds and we will continue to prove (25).

Let $\boldsymbol{P}$ be the projection matrix of the linear space spanned by $\boldsymbol{\gamma_j}$ for $j \in J$, then $\boldsymbol{z_U}^\top(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\gamma_j} < -\epsilon$ by (23) and

$$\begin{aligned}
\langle \boldsymbol{z_{sHT}} - \boldsymbol{\eta_{sHT}}, \boldsymbol{\gamma_j} \rangle &= \boldsymbol{z_{sHT}}^\top(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\gamma_j} \\
&= (\boldsymbol{z_{sHT}} - \boldsymbol{z_U} + \boldsymbol{z_U})^\top(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\gamma_j} \\
&= \boldsymbol{z_U}^\top(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\gamma_j} + (\boldsymbol{z_{sHT}} - \boldsymbol{z_U})^\top(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\gamma_j} \\
&= \boldsymbol{z_U}^\top(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\gamma_j} + R_J.
\end{aligned}$$

Thus, $\langle \boldsymbol{z_{sHT}} - \boldsymbol{\eta_{sHT}}, \boldsymbol{\gamma_j} \rangle = \boldsymbol{z_U}\top(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\gamma_j} + R_J$ and $\mathrm{P}(A^c) < \mathrm{P}(R_J \geq \epsilon)$.

Since $V(\bar{y}_{s_t}) = O(1/n)$ and by Chebyshev's inequality,

$$\mathrm{P}\left(|\bar{y}_{s_t} - \bar{y}_{U_t}| \geq \epsilon\right) \leq \frac{V(\bar{y}_{s_t})}{\epsilon^2} = o\left(\frac{1}{\sqrt{n}}\right).$$

then $\mathrm{P}(A^c) < \mathrm{P}(R_J \geq \epsilon) = o(1/\sqrt{n})$.

By Chebyshev's inequality again, for every $\zeta > 0$,

$$\lim_{n\to 0} \mathrm{P}\left(|I_{A^c}| > \frac{\zeta}{\sqrt{n}}\right) \leq \lim_{n\to 0} \frac{\sqrt{n}E|A^c|}{\zeta} = \lim_{n\to 0} \frac{\sqrt{n}\mathrm{P}(A^c)}{\zeta} = 0.$$

■

Theorem 10. *Let* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_T)^\top$ *be the isotonic regression of the population means* $\bar{\boldsymbol{y}}_{\boldsymbol{U}} = (\bar{y}_{U_1}, \ldots, \bar{y}_{U_T})^\top$, $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_T)^\top$ *be the weighted isotonic regression of the sample domain means* $\bar{\boldsymbol{y}}_{\boldsymbol{s}} = (\tilde{y}_{s_1}, \ldots, \tilde{y}_{s_T})^\top$, *and* A *be the event defined as in Theorem 9. Under Assumptions 1-4 from Section 3.2,*

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} + o_p\left(\frac{1}{\sqrt{n}}\right)\mathbf{1}_{T\times 1}$$

and

$$\frac{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}}{\sqrt{AV(\hat{\boldsymbol{\theta}})}} \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{1})$$

where

$$AV(\hat{\boldsymbol{\theta}}) = \boldsymbol{W}^{-1/2}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{W}^{1/2}\boldsymbol{\Sigma}\boldsymbol{W}^{-1/2}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{W}^{1/2}$$

and $\boldsymbol{\Sigma}$ *is the variance-covariance matrix of* $\bar{\boldsymbol{y}}_{\boldsymbol{s}}$ *with elements*

$$\Sigma_{tm} = \frac{1}{N_t N_m}\sum_{k,l\in U}\sum \Delta_{kl}\frac{z_{tk}y_k}{\pi_k}\frac{z_{ml}y_l}{\pi_l}.$$

Additionally, Under Assumption 5 from Section 3.2,

$$\frac{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}}{\sqrt{\hat{V}(\hat{\boldsymbol{\theta}})}} \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{1})$$

where

$$\hat{V}(\hat{\boldsymbol{\theta}}) = \boldsymbol{W}^{-1/2}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{W}^{1/2}\hat{\boldsymbol{\Sigma}}\boldsymbol{W}^{-1/2}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{W}^{1/2}$$

and $\hat{\boldsymbol{\Sigma}}$ *has elements*

$$\hat{\Sigma}_{tm} = \frac{1}{N_t N_m} \sum_{k,l \in s} \sum \frac{\Delta_{kl}}{\pi_{kl}} \frac{z_{tk} y_k}{\pi_k} \frac{z_{ml} y_l}{\pi_l}.$$

Proof. By Theorem 9, $I_{\{A^c\}} = o_p(1/\sqrt{n})$, so that

$$\begin{aligned} \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} &= \boldsymbol{W}^{-1/2}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{W}^{1/2}(\bar{\boldsymbol{y}}_{\boldsymbol{s}} - \bar{\boldsymbol{y}}_{\boldsymbol{U}})I_{\{A\}} + (\boldsymbol{C} - \boldsymbol{\theta})I_{\{A^c\}} \\ &= \boldsymbol{W}^{-1/2}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{W}^{1/2}(\bar{\boldsymbol{y}}_{\boldsymbol{s}} - \bar{\boldsymbol{y}}_{\boldsymbol{U}}) + o_p(1/\sqrt{n})\mathbf{1}_{T\times 1}, \end{aligned}$$

where $(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{W}^{1/2}\bar{\boldsymbol{y}}_{\boldsymbol{s}}$ is the projection of $\boldsymbol{W}^{1/2}\bar{\boldsymbol{y}}_{\boldsymbol{s}}$ onto the linear space spanned by γ_j with $j \in J$ and $\boldsymbol{C}$ denotes other cases of the estimate. Assumptions 3 and 4 lead to

$$\frac{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}}{\sqrt{AV(\hat{\boldsymbol{\theta}})}} \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{1}).$$

Furthermore, by Assumption 5,

$$\frac{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}}{\sqrt{\hat{V}(\hat{\boldsymbol{\theta}})}} \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{1}).$$

■

4.3 Simulation

In simulation study, three limiting domain mean values on a $t_1 \times t_2$ grid are used to demonstrate the practical properties of the constrained estimation

$$\mu_1 = i/t_1 + j/t_2$$
$$\mu_2 = i/t_1 + j/t_2 + (i/t_1) * (j/t_2)$$
$$\mu_3 = 4 \max\{\max\{i/t_1, j/t_2\} - 0.6, 0\}$$

where $i = 1, \ldots, t_1$ and $j = 1, \ldots, t_2$. These three limiting domain means have a partial ordering on the $t_1 \times t_2$ grid and let $T = t_1 \times t_2$. We first set the population size as $N = 10000$ with N/T in each domain, then the population values y_k are generated by adding independent and identically distributed $N(0, \sigma^2)$ errors to the μ values. Samples are generated from a stratified sampling design with simple random sampling without replacement in all strata, with $H = 4$ strata that cut across the T domains. The strata are determined by a variable z which is generated by adding $N(0, 1)$ errors to $(i/t_1)\sigma$. For each limiting domain mean function, 50000 replicate samples are selected for $t_1 = t_2 = 5$ and $t_1 = t_2 = 10$. For each sample, both the unconstrained and constrained estimate are calculated as well as their corresponding standard errors. A 95% confidence interval is constructed and compared.

Figure 4.1 show the average performance of the coverage rate and confidence length for $\mu_1 = i/t_1 + j/t_2$ with $t_1 = t_2 = 5$, the sample size is $n = 500$ in the first row and $n = 1000$ in the second row. In both cases, the constrained estimate increases the coverage rates with smaller associated confidence size. The coverage rates are all below 0.95 and it is due to smaller sample size. In a 5×5 grid, sample size 500 will only assign 20 observations in each domain on average. When the sample size is increasing, the coverage rates of the unconstrained estimate are getting closer to 0.95, as shown in Table 4.1. Similar results are shown in Figure 4.2 in a 10×10 grid. When the sample size is 1000, around 130 samples have no observations in at least one domain and they are excluded from the calculation of average

performance. Figure 4.3 - 4.4 present the simulation results for $\mu_2 = i/t_1 + j/t_2 + (i/t_1) * (j/t_2)$, and the constrained estimate provide higher coverage rates with smaller size of confidence interval as well.

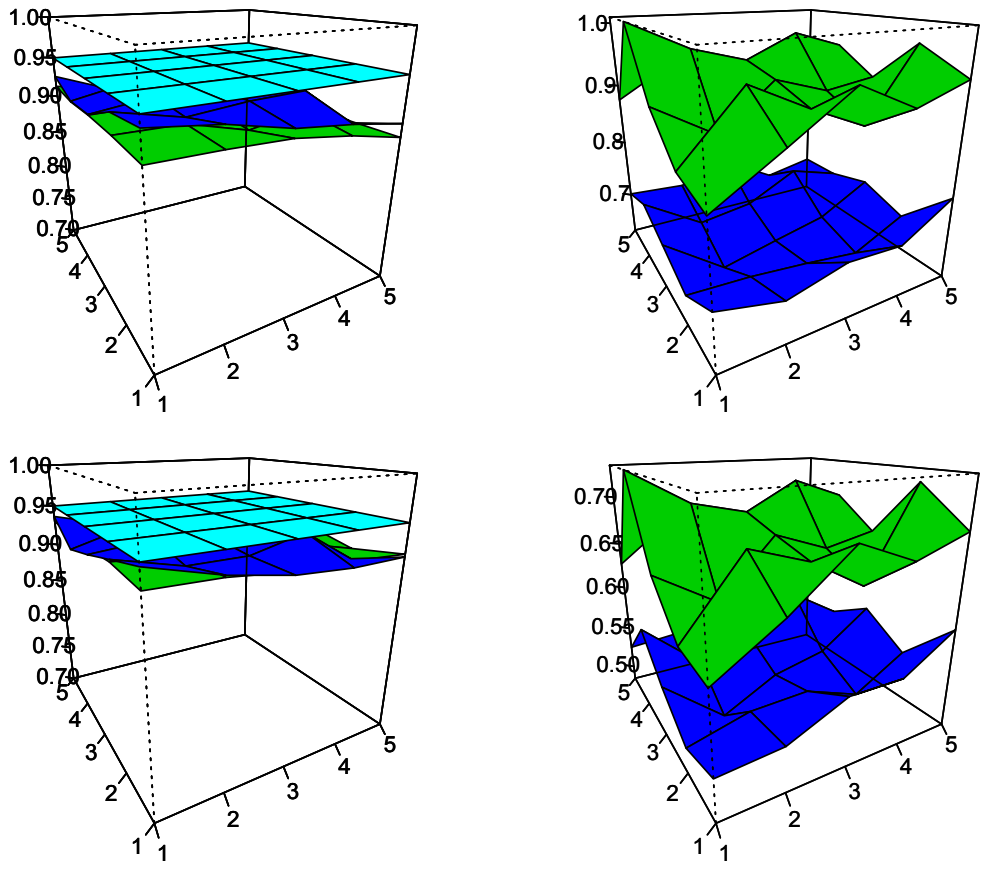

Figure 4.1: Comparisons of the coverage rate (left plot) and confidence length (right plot) for constrained (blue) and unconstrained (green) fit based on 50000 replicate samples for $\mu_1 = i/t_1 + j/t_2$ with $t_1 = t_2 = 5$, $\sigma = 1$, $N = 10000$, $n = 500$ (first row) and $n = 1000$ (second row).

In the case of $\mu_3 = 4 \max\{\max\{i/t_1, j/t_2\} - 0.6, 0\}$, the plane is flat at the lower end and then linearly increase to the upper end. Figure 4.5 shows that the coverage rates of constrained estimate are dropping at end points and the junction point of flat and increasing planes. There are two reasons to cause the decreased coverage rates: first, the constrained estimate has spiking problem and the estimate tend to be biased small at the smallest cell and biased large at the biggest cell; second, the population size is not large and the

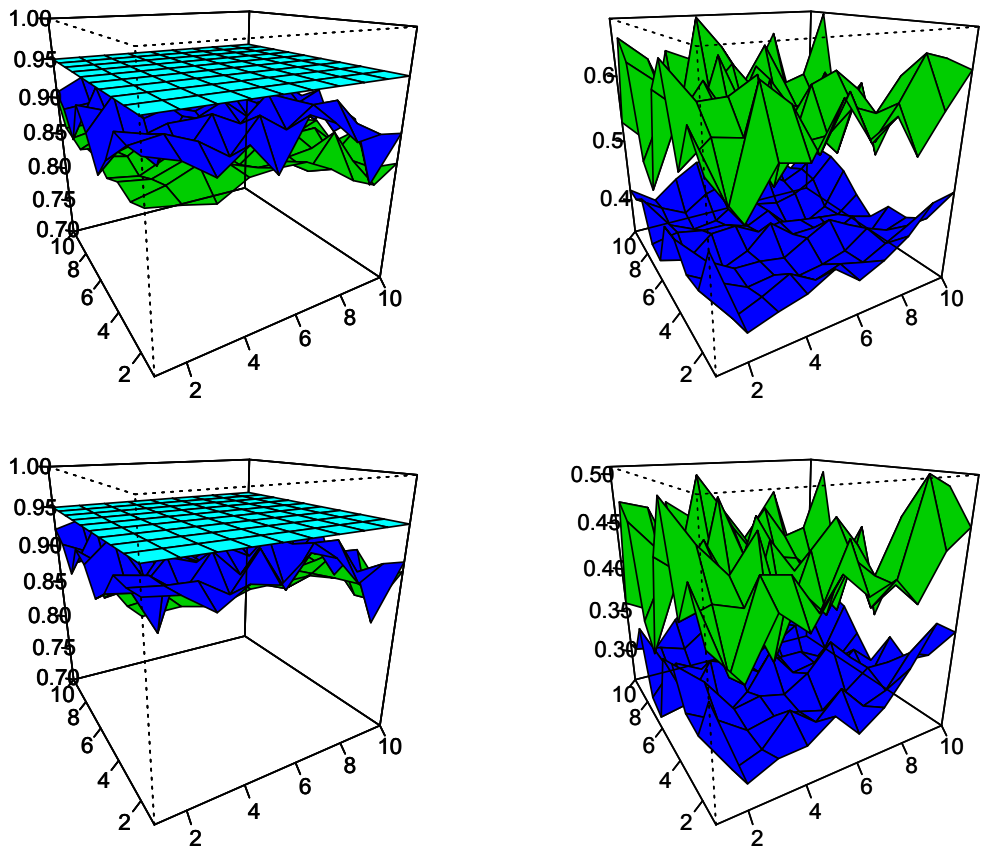

Figure 4.2: Comparisons of the coverage rate (left plot) and confidence length (right plot) for constrained (blue) and unconstrained (green) fit based on 50000 replicate samples for $\mu_1 = i/t_1 + j/t_2$ with $t_1 = t_2 = 10$, $\sigma = 0.5$, $N = 10000$, $n = 1000$ (first row) and $n = 2000$ (second row).

monotonicity constraints tend to be violated in the grid. When the population size increased to $N = 100000$, Figure 4.6 compare the performance of confidence interval estimate of different constrained estimate with the unconstrained estimate. The first row of Figure 4.6 just presents the usual constrained fit and the unconstrained fit, after increasing the sample size, the coverage rates are improved in some of the cells but behave bad at end points and the junction point. A penalty on the range can fix the spiking problem (Wu et al., 2015) and improve the confidence interval estimate at boundaries. As shown in the second row of Figure 4.6, the penalized constrained estimate increases the coverage rates at both lower and upped end points, but the dropping of coverage rate in the middle point is not affected by the

penalty. At the junction point of flat and increasing parts, constrained estimate inclines to be pooled down towards the flat plane at lower end cells. The relaxed constrained estimate can be helpful in this situation as presented in the third row of Figure 4.6, it improves the coverage rates at the junction point in the middle of the grid. The last row of Figure 4.6 shows the results of penalized constrained estimate with relax ordering constraints and it provides the best results of the coverage rates. In general, all constrained estimate give shorter length of confidence interval with competitive coverage rates. A penalized version of constrained fit can fix the spiking problem at boundary points and improve the coverage rates. A relax constrained estimate will improve the results at the junction point of flat to increasing. Combining the penalized constrained fit with a relax ordering assumption may give a good fit overall.

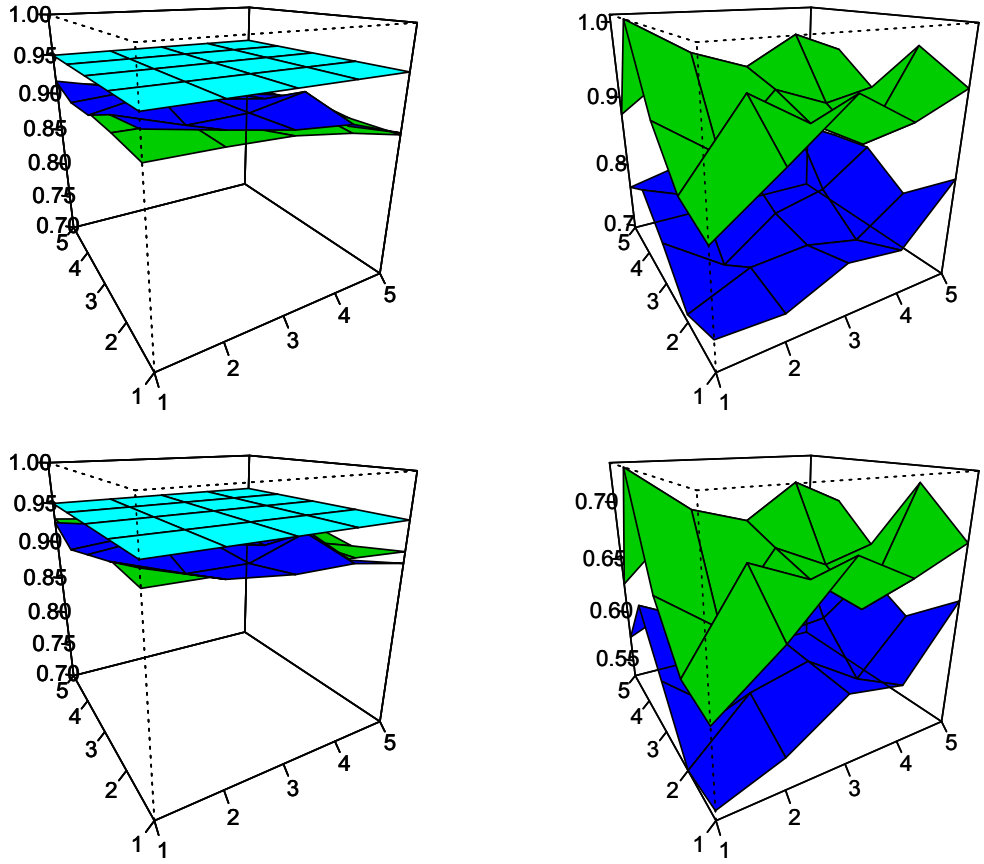

Figure 4.3: Comparisons of the coverage rate (left plot) and confidence length (right plot) for constrained (blue) and unconstrained (green) fit based on 50000 replicate samples for $\mu_2 = i/t_1 + j/t_2 + (i/t_1) * (j/t_2)$ with $t_1 = t_2 = 5$, $\sigma = 1$, $N = 10000$, $n = 500$ (first row) and $n = 1000$ (second row).

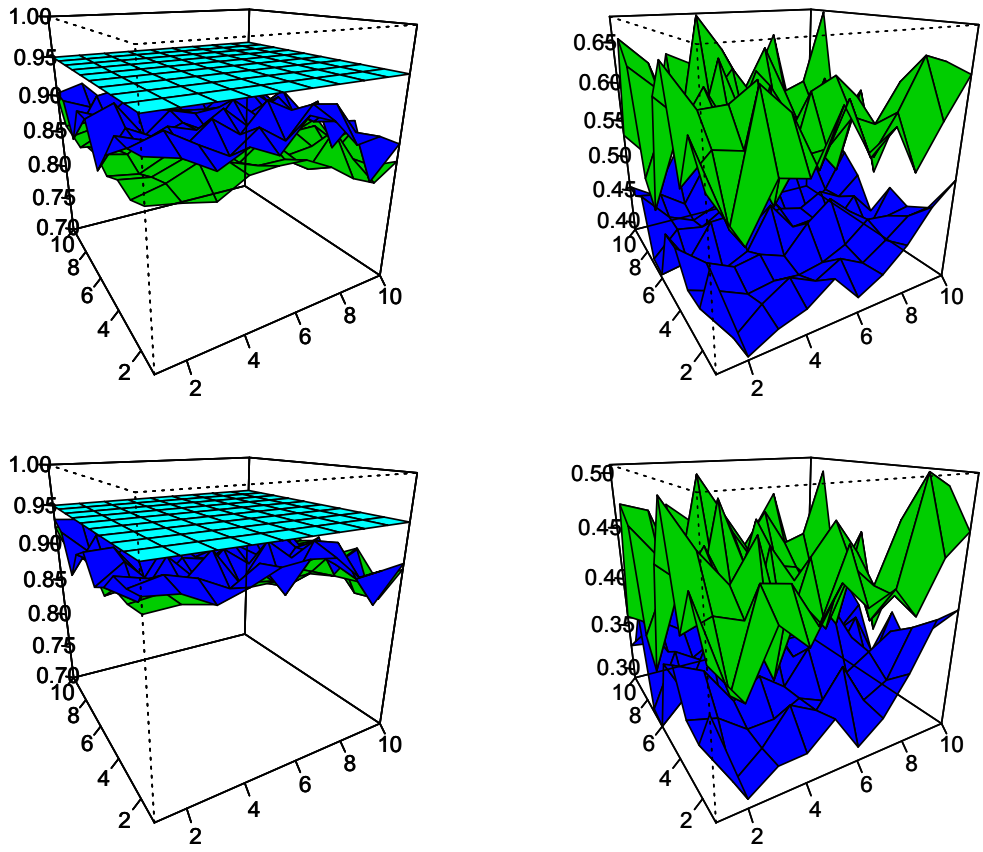

Figure 4.4: Comparisons of the coverage rate (left plot) and confidence length (right plot) for constrained (blue) and unconstrained (green) fit based on 50000 replicate samples for $\mu_2 = i/t_1 + j/t_2 + (i/t_1) * (j/t_2)$ with $t_1 = t_2 = 10$, $\sigma = 0.5$, $N = 10000$, $n = 1000$ (first row) and $n = 2000$ (second row).

Table 4.1: The coverage rates of unconstrained estimate for $\mu_1 = i/t_1 + j/t_2$ with $t_1 = t_2 = 5$ and $n = 500, 1000, 1500, 2000$.

500	1000	1500	2000	5000
0.909	0.927	0.936	0.939	0.945
0.912	0.932	0.935	0.939	0.943
0.912	0.928	0.936	0.941	0.943
0.909	0.928	0.935	0.939	0.944
0.913	0.930	0.937	0.938	0.944
0.908	0.926	0.933	0.938	0.942
0.911	0.928	0.935	0.938	0.942
0.914	0.928	0.936	0.940	0.941
0.904	0.925	0.932	0.935	0.944
0.904	0.923	0.932	0.936	0.942
0.905	0.920	0.928	0.932	0.941
0.892	0.918	0.928	0.930	0.941
0.899	0.922	0.928	0.931	0.942
0.897	0.922	0.930	0.935	0.940
0.905	0.922	0.928	0.933	0.940
0.895	0.920	0.931	0.934	0.940
0.904	0.920	0.929	0.933	0.939
0.890	0.916	0.929	0.934	0.942
0.885	0.914	0.926	0.932	0.943
0.897	0.919	0.929	0.934	0.941
0.881	0.916	0.926	0.931	0.943
0.880	0.906	0.920	0.927	0.938
0.888	0.919	0.929	0.936	0.942
0.878	0.912	0.925	0.929	0.940
0.882	0.915	0.928	0.931	0.942

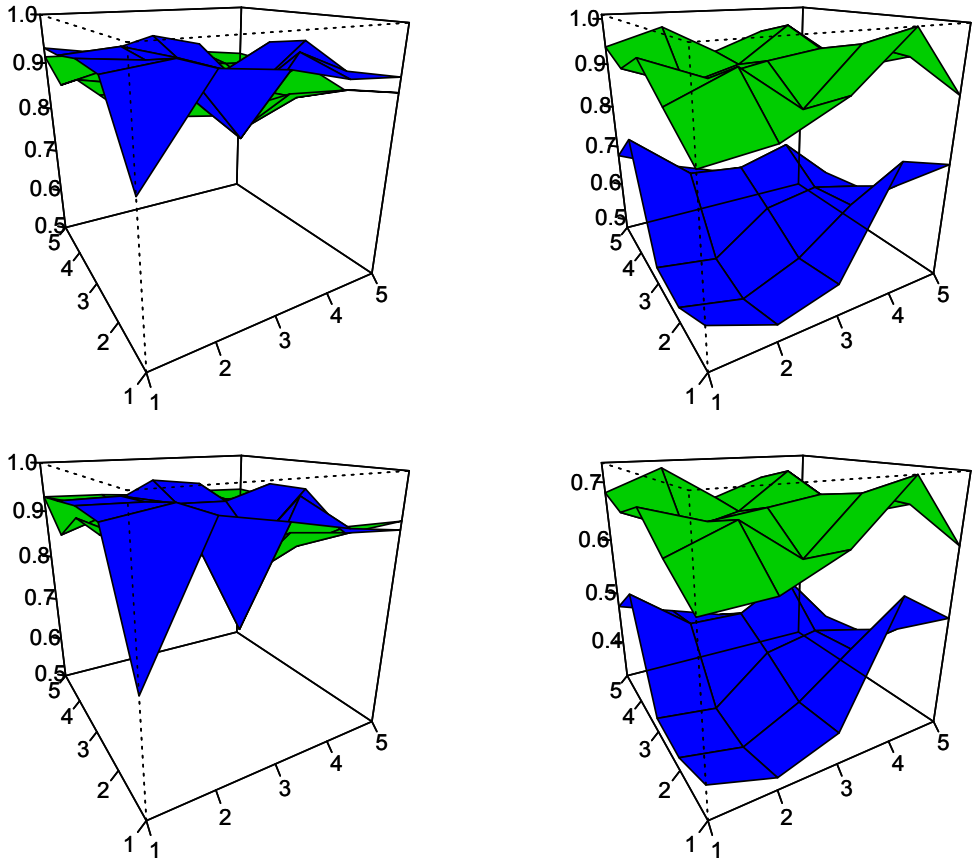

Figure 4.5: Comparisons of the coverage rate (left plot) and confidence length (right plot) for constrained (blue) and unconstrained (green) fit based on 50000 replicate samples for $\mu_3 = 4\max\{\max\{i/t_1, j/t_2\} - 0.6, 0\}$ with $t_1 = t_2 = 5$, $\sigma = 1$, $N = 10000$, $n = 500$ (first row) and $n = 1000$ (second row).

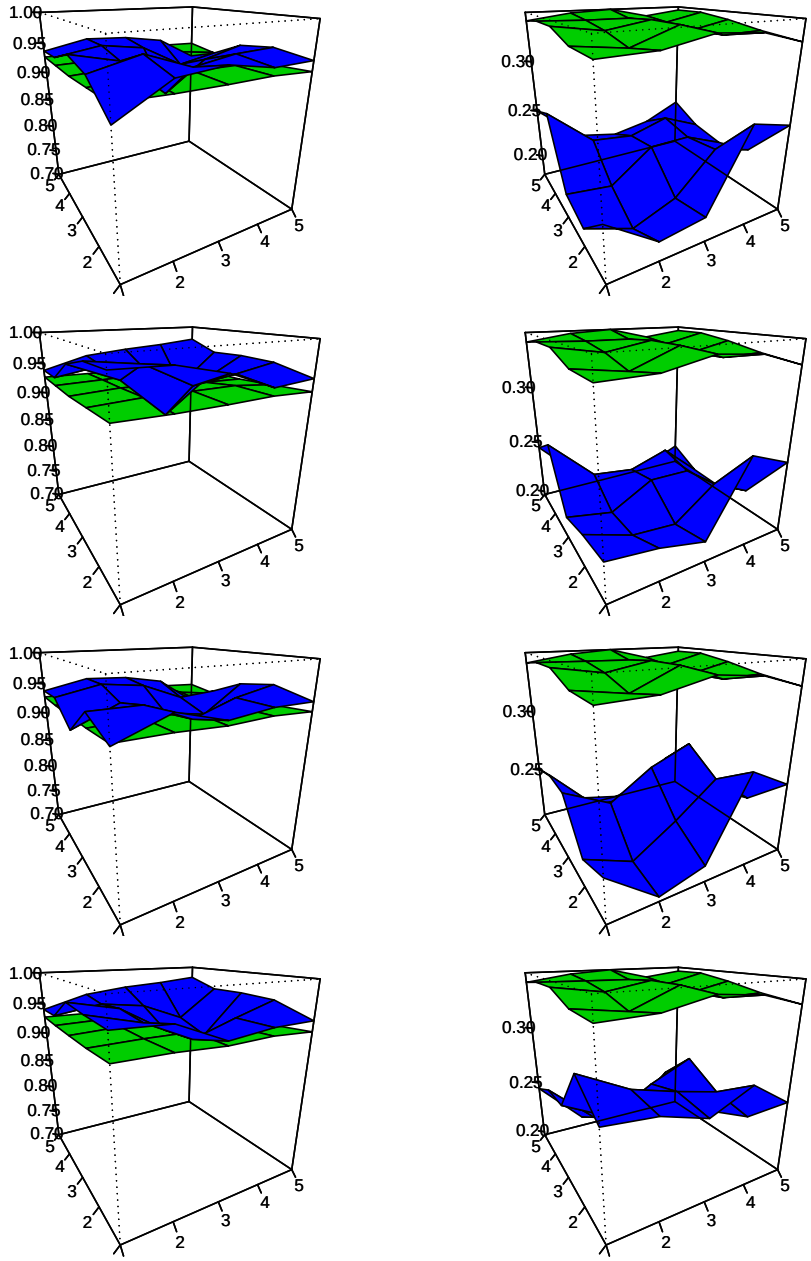

Figure 4.6: Comparisons of the coverage rate (left column) and confidence length (right column) for constrained estimate (blue in first row), penalized constrained estimate (blue in second row), constrained estimate of relax ordering (blue in third row), penalized constrained estimate of relax ordering (blue in fourth row), and unconstrained fit (green in each row) based on 50000 replicate samples for $\mu_3 = 4\max\{\max\{i/t_1, j/t_2\} - 0.6, 0\}$ with $t_1 = t_2 = 5$, $\sigma = 1$, $N = 100000$ and $n = 1000$.

CHAPTER 5

CONCLUSION AND FUTURE WORK

In this dissertation, we first propose a new method to correct the 'spiking' problem common in isotonic regression, by adding a penalty on the range of the estimates. The optimal penalty is investigated and is shown to depend on the derivatives of the true regression function at the boundaries. For univariate and bivariate isotonic regression, confidence intervals are constructed using the penalized estimate and bootstrapping method. Simulation studies show the improvement of the coverage rates at end points with decreased confidence size. The power of the hypothesis test of constant versus increasing regression function under different model is also studied and the results support the penalized estimate.

Next, we apply isotonic regression in survey sampling when natural orderings arise in survey data. For univariate case, the PAVA-based constrained estimate is constructed by adaptively collapsing neighboring domains. For bivariate case, a cone projection algorithm gives constrained domain estimate with more general qualitative assumptions. The theoretical properties of the constrained domain estimate are studied and a relaxed monotone constraint is proposed and implemented through cone projection algorithm. Simulation studies are used to compare the confidence interval using usual constrained estimate, penalized constrained estimate, relax constrained estimate, and penalized constrained estimate with relax constraints.

In the future work, the theoretical properties of penalized constrained estimate in survey context need to be studied and the relaxed monotone constraints also need more work on selection of the bandwidth. To make this work more applicable, the codes for constrained survey estimate will be summarized and put into an R package. Additionally, isotonic regression can be incorporated into the model-assisted framework for improving the survey

estimates when some auxiliary variables are related to the study variable in a shape-restricted way. More general constrained regression model can be used to handle more covariates and different constraint settings.

REFERENCES

Aldrin, M. (2004). No_2 data. http://lib.stat.cmu.edu/datasets/.

Ayer, M., Brunk, H., Ewing, G., Reid, W., and Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 26(4):641–647.

Barlow, R., Bartholomew, D., Bremner, J., and Brunk, H. (1972). *Statistical inference under order restrictions: the theory and application of isotonic regression*. J. Wiley.

Block, H., Qian, S., and Sampson, A. (1994). Structure algorithms for partially ordered isotonic regression. *Journal of Computational and Graphical Statistics*, 3(3):285–300.

Brunk, H. (1955). Maximum likelihood estimates of monotone parameters. *The Annals of Mathematical Statistics*, 26(4):607–616.

Brunk, H. (1958). On the estimation of parameters restricted by inequalities. *The Annals of Mathematical Statistics*, 29(2):437–454.

Burdakov, O., Sysoev, O., Grimvall, A., and Hussian, M. (2006). An o(n2) algorithm for isotonic regression. *Large-Scale Nonlinear Optimization*, pages 25–33.

Chepoi, V., Cogneau, D., and Fichet, B. (1997). Polynomial algorithms for isotonic regression. *Lecture Notes-Monograph Series*, pages 147–160.

Dykstra, R. and Robertson, T. (1982). An algorithm for isotonic regression for two or more independent variables. *The Annals of Statistics*, 10(3):708–716.

Hanson, D., Pledger, G., and Wright, F. (1973). On consistency in monotonic regression. *The Annals of Statistics*, pages 401–421.

Hardy, G. H. and Ramanujan, S. (1918). Asymptotic formulaæ in combinatory analysis. *Proceedings of the London Mathematical Society*, 2(1):75–115.

Henderson, H. V. and Searle, S. R. (1981). On deriving the inverse of a sum of matrices. *Siam Review*, 23(1):53–60.

Kott, P. S. (2001). The delete-a-group jackknife. *Journal of Official Statistics*, 17:521–526.

Meyer, M. C. (1999). An extension of the mixed primal–dual bases algorithm to the case of more constraints than dimensions. *Journal of Statistical Planning and Inference*, 81(1):13–31.

Meyer, M. C. (2006). Consistency and power in tests with shape-restricted alternatives. *Journal of statistical planning and inference*, 136(11):3931–3947.

Meyer, M. C. (2013). A simple new algorithm for quadratic programming with applications in statistics. *Communications in Statistics: Simulation and Computation*, 42(5):1126–1139.

Meyer, M. C. and Wang, J. C. (2012). Improved power of one-sided tests. *Statistics & Probability Letters*, 82(8):1619–1622.

Meyer, M. C. and Woodroofe, M. (2000). On the degrees of freedom in shape-restricted regression. *Annals of Statistics*, pages 1083–1104.

Pal, J. (2008). Spiking problem in monotone regression: Penalized residual sum of squares. *Statistics & Probability Letters*, 78(12):1548–1556.

Qian, S. and Eddy, W. (1996). An algorithm for isotonic regression on ordered rectangular grids. *Journal of Computational and Graphical Statistics*, 5(3):225–235.

Raubertas, R. F., Charles Lee, C.-I., and Nordheim, E. V. (1986). Hypothesis tests for normal means constrained by linear inequalities. *Communications in Statistics-Theory and Methods*, 15(9):2809–2833.

Resnick, S. (1992). *Adventures in stochastic processes*. Birkhäuser Boston.

Robertson, T., Wright, F., Dykstra, R., and Robertson, T. (1988). *Order restricted statistical inference*, volume 229. Wiley New York.

Sampson, A., Singh, H., and Whitaker, L. (2003). Order restricted estimators: some bias results. *Statistics & probability letters*, 61(3):299–308.

Sampson, A. R. and Whitaker, L. R. (1988). Positive dependence, upper sets, and multidimensional partitions. *Mathematics of operations research*, 13(2):254–264.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling*. Springer-Verlag, New York.

Shao, J. and Wu, C. J. (1989). A general theory for jackknife variance estimation. *The Annals of Statistics*, 17:1176–1197.

Spouge, J., Wan, H., and Wilbur, W. (2003). Least squares isotonic regression in two dimensions. *Journal of optimization theory and applications*, 117(3):585–605.

VanEeden, C. (1956). Maximum likelihood estimation of ordered probabilities. *Indigationes Mathematicae*, 18(3):444–455.

Wu, J., Meyer, M. C., and Opsomer, J. D. (2015). Penalized isotonic regression. *Journal of Statistical Planning and Inference*, 161:12–24.

APPENDIX

In the first lemma, we establish conditions under which the greatest convex minorant (GCM) of two plots will have the same corner points. Given one set of points $\{(x_t, y_t)\}$, $t = 1, \ldots, T$, and its GCM, we determine ϵ_1 and ϵ_2 so that the plot of $\{(w_t, z_t)\}$, $t = 1, \ldots, T$ will have the same corner points if $\max_t |w_t - x_t| < \epsilon_1$ and $\max_t |z_t - y_t| < \epsilon_2$.

Lemma 3. *Suppose* $(0,0), (x_1, y_1), \ldots, (x_t, y_t), \ldots, (x_T, y_T)$ *are points where* $0 < x_1 < \cdots < x_T$. *For* $t = 1, \ldots, T$, *let* (x_t, g_t) *be the points of the corresponding greatest convex minorant and* θ_t *be the slope of the segment extending to the left at* (x_t, g_t). *Let* $\mathcal{T} = \{1, \ldots, t, \ldots, T\}$ *and use* $\mathcal{J} = \{j_1, \ldots, j_S\}$ *to denote the subset of indices of the* S *corner points of* (x_t, g_t), *i.e., the points with different left-derivative and right-derivative slopes, where* $0 < j_1 < j_2 < \cdots < j_S = T$. *Denote* $\delta_1 = \min_{s=1,\ldots,S-1}(\theta_{j_{s+1}} - \theta_{j_s})$, $\delta_2 = \min(x_{i+1} - x_i)$ *for* $i = 1, \ldots, T-1$, $\delta_3 = \min(y_i - g_i)$ *for* $i \in \mathcal{J}^c$, $R_x = x_T$ *and* $R_y = \max_t y_t - \min_t y_t$ *for* $t = 1, \ldots, T$. *Given points* (w_t, z_t), *for* $t = 1, \ldots, T$, *if* $\max_t |w_t - x_t| < \epsilon_1$ *and* $\max_t |z_t - y_t| < \epsilon_2$ *where*

$$\epsilon_1 < \min\left\{\frac{\delta_2}{2}, \frac{\delta_2\delta_3}{4R_y}, \frac{\delta_1\delta_2^2}{4R_y}\right\} \text{ and } \epsilon_2 < \min\left\{\frac{\delta_2\delta_3 - 4R_y\epsilon_1}{4R_x}, \frac{\delta_1\delta_2^2 - 4R_y\epsilon_1}{4R_x}\right\},$$

then the greatest convex minorant of (w_t, z_t) *has the same set of corner points as that of* (x_t, y_t).

Proof. By the properties of the GCM, the quantities δ_1, δ_2, and δ_3 are all strictly positive, and hence ϵ_1 and ϵ_2 are also positive constants. Suppose $\{(w_t, z_t)\}$, $t = 1, \ldots, T$ are another set of points where $\max_t |w_t - x_t| < \epsilon_1$ and $\max_t |z_t - y_t| < \epsilon_2$.

We first show that the slope of the line segment connecting $(w_{j_{s+1}}, z_{j_{s+1}})$ and (w_{j_s}, z_{j_s}) is larger than the slope of line segment connecting (w_{j_s}, z_{j_s}) and $(w_{j_{s-1}}, z_{j_{s-1}})$, namely

$$\frac{z_{j_{s+1}} - z_{j_s}}{w_{j_{s+1}} - w_{j_s}} - \frac{z_{j_s} - z_{j_{s-1}}}{w_{j_s} - w_{j_{s-1}}} > 0, \tag{26}$$

where $s = 1, \ldots, S-1$. Since $\theta_{j_{s+1}} - \theta_{j_s} > \delta_1$, then $(y_{j_{s+1}} - y_{j_s})(x_{j_s} - x_{j_{s-1}}) - (y_{j_s} - y_{j_{s-1}})(x_{j_{s+1}} - x_{j_s}) > \delta_1(x_{j_{s+1}} - x_{j_s})(x_{j_s} - x_{j_{s-1}})$ and

$$\begin{aligned}
&(z_{j_{s+1}} - z_{j_s})(w_{j_s} - w_{j_{s-1}}) - (z_{j_s} - z_{j_{s-1}})(w_{j_{s+1}} - w_{j_s}) \\
&> (y_{j_{s+1}} - y_{j_s} - 2\epsilon_2)(x_{j_s} - x_{j_{s-1}} - 2\epsilon_1) - (y_{j_s} - y_{j_{s-1}} + 2\epsilon_2)(x_{j_{s+1}} - x_{j_s} + 2\epsilon_1) \\
&= (y_{j_{s+1}} - y_{j_s})(x_{j_s} - x_{j_{s-1}}) - (y_{j_s} - y_{j_{s-1}})(x_{j_{s+1}} - x_{j_s}) \\
&- 2\epsilon_1(y_{j_{s+1}} - y_{j_s} + y_{j_s} - y_{j_{s-1}}) - 2\epsilon_2(x_{j_s} - x_{j_{s-1}} + x_{j_{s+1}} - x_{j_s}) \\
&> \delta_1(x_{j_{s+1}} - x_{j_s})(x_{j_s} - x_{j_{s-1}}) - 2\epsilon_1(y_{j_{s+1}} - y_{j_s} + y_{j_s} - y_{j_{s-1}}) - 2\epsilon_2(x_{j_s} - x_{j_{s-1}} + x_{j_{s+1}} - x_{j_s}) \\
&> \delta_1\delta_2^2 - 4\epsilon_1 R_y - 4\epsilon_2 R_x \\
&> 0.
\end{aligned}$$

Next, let $\mathcal{J}_s$ be the set containing indices between j_s and j_{s+1} for $s = 1, \ldots, S-1$, and notice that the corner points partition $\mathcal{T}$ into $S-1$ parts. If j_s and j_{s+1} are consecutive numbers in $\mathcal{T}$, then $\mathcal{J}_s$ is an empty set and (26) shows the relationship of slopes at $(w_{j_{s+1}}, z_{j_{s+1}})$ and (w_{j_s}, z_{j_s}). If $\mathcal{J}_s$ is not empty, then, for any $j \in \mathcal{J}_s$, it will be shown that the point (w_j, z_j) is above the line connecting (w_{j_s}, z_{j_s}) and $(w_{j_{s+1}}, z_{j_{s+1}})$, namely

$$z_j - z_{j_s} - \frac{z_{j_{s+1}} - z_{j_s}}{w_{j_{s+1}} - w_{j_s}}(w_j - w_{j_s}) > 0. \tag{27}$$

According to the definition of δ_3 and greatest convex minorant, $y_j - g_j > \delta_3$, that is

$$y_j - y_{j_s} - \frac{y_{j_{s+1}} - y_{j_s}}{x_{j_{s+1}} - x_{j_s}}(x_j - x_{j_s}) > \delta_3$$

and $(y_j - y_{j_s})(x_{j_{s+1}} - x_{j_s}) - (y_{j_{s+1}} - y_{j_s})(x_j - x_{j_s}) > \delta_3(x_{j_{s+1}} - x_{j_s})$. Then,

$$\begin{aligned}
&(z_j - z_{j_s})(w_{j_{s+1}} - w_{j_s}) - (z_{j_{s+1}} - z_{j_s})(w_j - w_{j_s}) \\
&\quad > (y_j - y_{j_s} - 2\epsilon_2)(x_{j_{s+1}} - x_{j_s} - 2\epsilon_1) - (y_{j_{s+1}} - y_{j_s} + 2\epsilon_2)(x_j - x_{j_s} + 2\epsilon_1) \\
&\quad = (y_j - y_{j_s})(x_{j_{s+1}} - x_{j_s}) - (y_{j_{s+1}} - y_{j_s})(x_j - x_{j_s}) \\
&\quad - 2\epsilon_1(y_j - y_{j_s} + y_{j_{s+1}} - y_{j_s}) - 2\epsilon_2(x_{j_{s+1}} - x_{j_s} + x_j - x_{j_s}) \\
&\quad > \delta_3(x_{j_{s+1}} - x_{j_s}) - 2\epsilon_1(y_j - y_{j_s} + y_{j_{s+1}} - y_{j_s}) - 2\epsilon_2(x_{j_{s+1}} - x_{j_s} + x_j - x_{j_s}) \\
&\quad > \delta_2\delta_3 - 4\epsilon_1 R_y - 4\epsilon_2 R_x \\
&\quad > 0.
\end{aligned}$$

Combining (26) and (27) implies that $\mathcal{J}$ is the set containing all indices of corner points of the greatest convex minorant of points (w_t, z_t). Thus the greatest convex minorant of points (w_t, z_t) has the same corner points as that of points (x_t, y_t).

■

Lemma 4. *For any domain U_t, let $AV(\tilde{y}_{s_t})$ and $\hat{V}(\tilde{y}_{s_t})$ be as defined in* (15) *and* (16)*, respectively, with $i = t = j$. Under Assumptions 1-5, $\tilde{y}_{s_t}$ is asymptotically normal and $n\left(\hat{V}(\tilde{y}_{s_t}) - AV(\tilde{y}_{s_t})\right) = o_p(1)$.*

Proof. We first write

$$\begin{aligned}
\hat{V}(\tilde{y}_{s_t}) &= \frac{1}{\hat{N}_t^2}\sum_{k,l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{(y_k - \bar{y}_{U_t})z_{tk}}{\pi_k}\frac{(y_l - \bar{y}_{U_t})z_{tl}}{\pi_l} + \frac{1}{\hat{N}_t^2}\sum_{k,l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{(\bar{y}_{U_t} - \tilde{y}_{s_t})^2 z_{tk}z_{tl}}{\pi_k\pi_l} \\
&+ \frac{2}{\hat{N}_t^2}\sum_{k,l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{(y_k - \bar{y}_{U_t})z_{tk}}{\pi_k}\frac{(\bar{y}_{U_t} - \tilde{y}_{s_t})z_{tl}}{\pi_l}
\end{aligned}$$

$$= \quad A_1 + B + 2C,$$

then

$$\begin{aligned}
\hat{V}(\tilde{y}_{s_t}) - AV(\tilde{y}_{s_t}) &= \frac{1}{\hat{N}_t^2} \sum_{k,l \in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{(y_k - \bar{y}_{U_t}) z_{tk}}{\pi_k} \frac{(y_l - \bar{y}_{U_t}) z_{tl}}{\pi_l} \\
&- \frac{1}{N_t^2} \sum_{k,l \in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{(y_k - \bar{y}_{U_t}) z_{tk}}{\pi_k} \frac{(y_l - \bar{y}_{U_t}) z_{tl}}{\pi_l} \\
&+ \frac{1}{N_t^2} \sum_{k,l \in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{(y_k - \bar{y}_{U_t}) z_{tk}}{\pi_k} \frac{(y_l - \bar{y}_{U_t}) z_{tl}}{\pi_l} \\
&- \frac{1}{N_t^2} \sum_{k,l \in U} \Delta_{kl} \frac{(y_k - \bar{y}_{U_t}) z_{tk}}{\pi_k} \frac{(y_l - \bar{y}_{U_t}) z_{tl}}{\pi_l} \\
&+ B + 2C \\
&= A_1 - A_2 + A_3 - A_4 + B + 2C.
\end{aligned}$$

By Assumptions 3,

$$\sum_{k,l \in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{(y_k - \bar{y}_{U_t}) z_{tk}}{\pi_k} \frac{(y_l - \bar{y}_{U_t}) z_{tl}}{\pi_l} = O_p\left(\frac{1}{n}\right),$$

and since $(\hat{N}_t - N_t)/N = o_p(1)$, then

$$\frac{N_t^2}{\hat{N}_t^2} - 1 = \frac{N_t^2/N^2}{N_t^2/N^2 + \hat{N}_t^2/N^2 - N_t^2/N^2} - 1 = \frac{1}{1 + o_p(1)2N/N_t + o_p(1)N^2/N_t^2} - 1 = o_p(1)$$

$$A_1 - A_2 = \left(\frac{N_t^2}{\hat{N}_t^2} - 1\right) \sum_{k,l \in s} \frac{\Delta_{kl}}{\pi_{kl}} \frac{(y_k - \bar{y}_{U_t}) z_{tk}}{\pi_k} \frac{(y_l - \bar{y}_{U_t}) z_{tl}}{\pi_l} = o_p(1) O_p\left(\frac{1}{n}\right) = o_p\left(\frac{1}{n}\right).$$

Similarly, with $N_t^2/\hat{N}_t^2 = O_p(1)$ and $\tilde{y}_{s_t} - \bar{y}_{U_t} = O_p\left(1/\sqrt{n}\right)$ by Taylor Linearization

$$B = \frac{N_t^2}{\hat{N}_t^2}\frac{1}{N_t^2}\sum_{k,l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{z_{tk}z_{tl}}{\pi_k\pi_l}(\bar{y}_{U_t} - \tilde{y}_{s_t})^2 = O_p(1)O_p\left(\frac{1}{n}\right)O_p\left(\frac{1}{n}\right) = O_p\left(\frac{1}{n^2}\right),$$

and $C = O_p(1)O_p\left(1/n\right)O_p\left(1/\sqrt{n}\right) = O_p(1/n^{3/2})$. Applying Assumption 5 gives $A_3 - A_4 = o_p(1/n)$. Therefore, $n(\hat{V}(\tilde{y}_{s_t}) - AV(\tilde{y}_{s_t})) = o_p(1) + O_p(1/\sqrt{n}) \to 0$ as $n \to \infty$.

■

Table 2.1: The MSE optimal value of m in $\lambda = m\hat{\sigma}rn^{\alpha}$ are shown for $\mu(x) = x$ with $n = 50$.

m	MSE	MSE_1	$MSE_{n/2}$	MSE_n
$\sigma = .5$				
0	0.03113	0.15664	0.02013	0.16290
0.5	0.02335	0.04768	0.02007	0.05084
0.7	0.02173	0.03491	0.02003	0.03686
1.0	0.02028	0.02863	0.01994	0.02949
1.2	0.01981	0.02900	0.01985	0.02947
1.5	0.01970	0.03326	0.01967	0.03328
1.7	0.01995	0.03754	0.01953	0.03737
2.0	0.02074	0.04526	0.01926	0.04497
2.3	0.02197	0.05393	0.01891	0.05357
2.5	0.02299	0.06005	0.01863	0.05967
2.7	0.02417	0.06640	0.01831	0.06599
3.0	0.02623	0.07626	0.01775	0.07579
$\sigma = 1$				
0	0.10432	0.65402	0.05325	0.64224
0.5	0.06956	0.18593	0.05219	0.18249
0.7	0.06182	0.12640	0.05150	0.12451
1.0	0.05393	0.08508	0.05010	0.08434
1.2	0.05052	0.07440	0.04891	0.07398
1.5	0.04759	0.07153	0.04677	0.07155
1.7	0.04688	0.07510	0.04511	0.07542
2.0	0.04741	0.08557	0.04230	0.08594
2.3	0.04959	0.09982	0.03922	0.10029
2.5	0.05186	0.11081	0.03708	0.11134
2.7	0.05473	0.12273	0.03494	0.12334
3.0	0.06007	0.14186	0.03186	0.14265
$\sigma = 2$				
0	0.37733	2.59272	0.16089	2.54552
0.5	0.21999	0.65921	0.15143	0.64255
0.7	0.18923	0.46959	0.14557	0.46228
1.0	0.15861	0.33299	0.13506	0.33248
1.2	0.14585	0.28965	0.12738	0.29123
1.5	0.13597	0.26265	0.11566	0.26603
1.7	0.13469	0.26154	0.10839	0.26514
2.0	0.13823	0.27279	0.09957	0.27647
2.3	0.14429	0.28798	0.09356	0.29242
2.5	0.14807	0.29692	0.09087	0.30151
2.7	0.15133	0.30438	0.08898	0.30905
3.0	0.15508	0.31228	0.08715	0.31736

Table 2.2: The MSE optimal value of m in $\lambda = m\hat{\sigma}rn^{\alpha}$ are shown for $\mu(x) = x$ with $n = 200$.

m	MSE	MSE_1	$MSE_{n/2}$	MSE_n
$\sigma = .5$				
0.0	0.01163	0.16983	0.00774	0.15457
0.5	0.00899	0.02973	0.00774	0.02744
0.7	0.00851	0.01861	0.00774	0.01753
1.0	0.00809	0.01329	0.00774	0.01278
1.2	0.00794	0.01285	0.00774	0.01243
1.5	0.00788	0.01405	0.00774	0.01372
1.7	0.00791	0.01550	0.00774	0.01524
2.0	0.00807	0.01828	0.00774	0.01807
2.3	0.00833	0.02146	0.00774	0.02135
2.5	0.00856	0.02375	0.00774	0.02370
2.7	0.00882	0.02614	0.00774	0.02614
3.0	0.00928	0.02987	0.00774	0.02992
$\sigma = 1$				
0.0	0.03809	0.65358	0.01914	0.66405
0.5	0.02667	0.11179	0.01911	0.11497
0.7	0.02437	0.06646	0.01909	0.06931
1.0	0.02202	0.03911	0.01904	0.04105
1.2	0.02096	0.03202	0.01900	0.03332
1.5	0.01992	0.02890	0.01892	0.02944
1.7	0.01952	0.02951	0.01885	0.02969
2.0	0.01929	0.03263	0.01874	0.03247
2.3	0.01943	0.03740	0.01858	0.03697
2.5	0.01971	0.04117	0.01846	0.04065
2.7	0.02012	0.04529	0.01832	0.04472
3.0	0.02095	0.05195	0.01807	0.05137
$\sigma = 2$				
0.0	0.13159	2.66643	0.05184	2.55417
0.5	0.08428	0.51013	0.05135	0.49717
0.7	0.07448	0.32833	0.05099	0.32153
1.0	0.06386	0.19537	0.05024	0.19390
1.2	0.05864	0.14992	0.04958	0.14999
1.5	0.05288	0.11338	0.04838	0.11320
1.7	0.05015	0.10075	0.04742	0.10053
2.0	0.04745	0.09211	0.04577	0.09208
2.3	0.04618	0.09149	0.04384	0.09151
2.5	0.04602	0.09402	0.04243	0.09408
2.7	0.04637	0.09831	0.04093	0.09837
3.0	0.04776	0.10726	0.03856	0.10733

Table 2.3: The MSE optimal value of m in $\lambda = m\hat{\sigma}rn^{\alpha}$ are shown for $\mu(x) = x^2$ with $n = 50$.

m	MSE	MSE_1	$MSE_{n/2}$	MSE_n
$\sigma = .5$				
0.0	0.03134	0.15891	0.01923	0.16139
0.3	0.02531	0.06087	0.01912	0.07908
0.5	0.02288	0.03477	0.01900	0.05484
0.7	0.02131	0.02256	0.01884	0.04512
1.0	0.02009	0.01564	0.01854	0.04498
1.2	0.01985	0.01453	0.01828	0.05013
1.5	0.02017	0.01557	0.01782	0.06179
1.7	0.02075	0.01735	0.01747	0.07116
2.0	0.02213	0.02107	0.01690	0.08671
2.3	0.02401	0.02568	0.01630	0.10327
2.5	0.02553	0.02913	0.01590	0.11454
3.0	0.03012	0.03880	0.01494	0.14355
$\sigma = 1$				
0.0	0.10503	0.63872	0.05195	0.63553
0.3	0.07849	0.24062	0.05123	0.29041
0.5	0.06750	0.13944	0.05049	0.18833
0.7	0.05983	0.08950	0.04946	0.13725
1.0	0.05261	0.05649	0.04750	0.10910
1.2	0.04994	0.04779	0.04597	0.10716
1.5	0.04843	0.04483	0.04345	0.11713
1.7	0.04884	0.04714	0.04173	0.12897
2.0	0.05128	0.05467	0.03925	0.15159
2.3	0.05567	0.06551	0.03714	0.17831
2.5	0.05958	0.07416	0.03604	0.19766
3.0	0.07213	0.09891	0.03487	0.24854
$\sigma = 2$				
0.0	0.38112	2.60717	0.15589	2.52745
0.3	0.25634	0.87089	0.15066	1.00636
0.5	0.21022	0.52607	0.14526	0.64953
0.7	0.17949	0.36578	0.13894	0.47892
1.0	0.15176	0.25827	0.12839	0.36648
1.2	0.14240	0.22726	0.12132	0.33986
1.5	0.13924	0.21386	0.11282	0.33828
1.7	0.14255	0.21803	0.10959	0.35096
2.0	0.15089	0.23135	0.10792	0.37484
2.3	0.15894	0.24457	0.10869	0.39346
2.5	0.16331	0.25176	0.10992	0.40135
3.0	0.17150	0.26642	0.11410	0.40946

Table 2.4: The MSE optimal value of m in $\lambda = m\hat{\sigma}rn^{\alpha}$ are shown for $\mu(x) = x^2$ with $n = 200$.

m	MSE	MSE_1	$MSE_{n/2}$	MSE_n
$\sigma = .5$				
0.0	0.01175	0.16099	0.00756	0.16268
0.3	0.00957	0.03646	0.00756	0.05445
0.50	0.00881	0.01729	0.00756	0.03050
0.7	0.00834	0.01016	0.00755	0.02184
1.0	0.00796	0.00635	0.00755	0.01991
1.2	0.00786	0.00554	0.00754	0.02168
1.5	0.00790	0.00546	0.00752	0.02632
1.7	0.00801	0.00585	0.00750	0.03014
2.0	0.00831	0.00684	0.00747	0.03650
2.3	0.00874	0.00817	0.00743	0.04330
2.5	0.00909	0.00919	0.00740	0.04799
3.0	0.01015	0.01209	0.00729	0.06000
$\sigma = 1$				
0.0	0.03824	0.68029	0.01908	0.65004
0.3	0.02895	0.15288	0.01903	0.21695
0.5	0.02556	0.07151	0.01897	0.11732
0.7	0.02326	0.04032	0.01889	0.07356
1.0	0.02106	0.02188	0.01872	0.04960
1.2	0.02018	0.01699	0.01857	0.04594
1.5	0.01949	0.01431	0.01831	0.04831
1.7	0.01939	0.01428	0.01810	0.05291
2.0	0.01966	0.01576	0.01774	0.06220
2.3	0.02039	0.01852	0.01732	0.07339
2.5	0.02110	0.02079	0.01702	0.08151
3.0	0.02356	0.02765	0.01621	0.10320
$\sigma = 2$				
0.0	0.13323	2.63205	0.05262	2.63123
0.3	0.09468	0.66352	0.05224	0.88092
0.5	0.08045	0.35692	0.05178	0.50727
0.7	0.07041	0.22131	0.05115	0.32871
1.0	0.06005	0.12647	0.04988	0.20230
1.2	0.05533	0.09548	0.04884	0.16205
1.5	0.05066	0.07093	0.04702	0.13436
1.7	0.04886	0.06322	0.04567	0.12805
2.0	0.04785	0.05937	0.04357	0.12994
2.3	0.04857	0.06141	0.04146	0.14037
2.5	0.04992	0.06508	0.04015	0.15040
3.0	0.05594	0.07955	0.03740	0.18164

Table 2.5: The MSE optimal value of m in $\lambda = m\hat{\sigma}rn^{\alpha}$ are shown for $\mu(x) = 0$ on $[0, 0.2)$ and $\mu(x) = x^2 - .04$ on $[0.2, 1]$ with $n = 50$.

m	MSE	MSE_1	$MSE_{n/2}$	MSE_n
$\sigma = .5$				
0.0	0.03139	0.16748	0.01869	0.16192
0.3	0.02428	0.04460	0.01847	0.08155
0.5	0.02176	0.02258	0.01821	0.05747
0.7	0.02030	0.01461	0.01787	0.04712
1.0	0.01949	<u>0.01233</u>	0.01723	<u>0.04564</u>
1.2	0.01966	0.01352	0.01673	0.04980
1.5	0.02075	0.01749	0.01592	0.06020
1.7	0.02196	0.02110	0.01540	0.06884
2.0	0.02443	0.02755	0.01470	0.08316
2.3	0.02757	0.03496	0.01420	0.09866
2.5	0.03002	0.04037	0.01402	0.10948
3.0	0.03727	0.05522	0.01435	0.13743
$\sigma = 1$				
0.0	0.10496	0.66860	0.05274	0.63794
0.3	0.07404	0.17500	0.05134	0.30299
0.5	0.06268	0.08959	0.04983	0.20055
0.7	0.05548	0.05610	0.04798	0.14674
1.0	0.04995	<u>0.04162</u>	0.04486	0.11379
1.2	0.04895	0.04243	0.04289	<u>0.10933</u>
1.5	0.05062	0.05155	0.04058	0.11678
1.7	0.05360	0.06106	0.03980	0.12749
2.0	0.06047	0.07880	0.04012	0.14864
2.3	0.06984	0.09979	0.04257	0.17357
2.5	0.07710	0.11482	0.04537	0.19105
3.0	0.09556	0.15101	0.05535	0.22916
$\sigma = 2$				
0.0	0.38265	2.58638	0.16129	2.54066
0.3	0.24099	0.62362	0.15317	1.03543
0.5	0.19460	0.35409	0.14459	0.67389
0.7	0.16733	0.24943	0.13536	0.49394
1.0	0.15009	<u>0.20520</u>	0.12488	0.37297
1.2	0.15038	0.20987	0.12333	0.34180
1.5	0.16245	0.23976	0.13020	0.32993
1.7	0.17345	0.26312	0.13877	0.32923
2.0	0.18992	0.29537	0.15407	0.32717
2.3	0.20565	0.32459	0.17094	0.32051
2.5	0.21604	0.34322	0.18285	0.31458
3.0	0.24304	0.39023	0.21507	<u>0.29732</u>

Table 2.6: The MSE optimal value of m in $\lambda = m\hat{\sigma}rn^{\alpha}$ are shown for $\mu(x) = 0$ on $[0, 0.2)$ and $\mu(x) = x^2 - .04$ on $[0.2, 1]$ with $n = 200$.

m	MSE	MSE_1	$MSE_{n/2}$	MSE_n
$\sigma = .5$				
0.0	0.01163	0.16619	0.00753	0.15565
0.3	0.00901	0.01756	0.00752	0.05327
0.5	0.00825	0.00754	0.00751	0.03004
0.7	0.00784	0.00475	0.00748	0.02132
1.0	0.00762	0.00406	0.00742	0.01909
1.2	0.00768	0.00452	0.00736	0.02056
1.5	0.00799	0.00589	0.00725	0.02476
1.7	0.00833	0.00711	0.00716	0.02834
2.0	0.00902	0.00930	0.00700	0.03435
2.3	0.00990	0.01184	0.00680	0.04086
2.5	0.01058	0.01370	0.00666	0.04536
3.0	0.01258	0.01880	0.00624	0.05700
$\sigma = 1$				
0.0	0.03753	0.66346	0.01900	0.62071
0.3	0.02669	0.07116	0.01886	0.20620
0.5	0.02335	0.02928	0.01867	0.11128
0.7	0.02131	0.01670	0.01839	0.07044
1.0	0.01976	0.01197	0.01783	0.04837
1.2	0.01943	0.01241	0.01738	0.04455
1.5	0.01978	0.01558	0.01662	0.04660
1.7	0.02048	0.01879	0.01608	0.05083
2.0	0.02214	0.02478	0.01531	0.05957
2.3	0.02446	0.03181	0.01465	0.07020
2.5	0.02633	0.03696	0.01433	0.07795
3.0	0.03203	0.05116	0.01411	0.09861
$\sigma = 2$				
0.0	0.13316	2.66617	0.05174	2.62730
0.3	0.08716	0.34337	0.05082	0.90991
0.5	0.07264	0.16097	0.04964	0.52349
0.7	0.06329	0.09303	0.04810	0.33863
1.0	0.05523	0.05750	0.04536	0.21211
1.2	0.05263	0.05149	0.04348	0.17237
1.5	0.05195	0.05483	0.04103	0.14484
1.7	0.05331	0.06182	0.03995	0.13824
2.0	0.05771	0.07688	0.03968	0.13854
2.3	0.06462	0.09593	0.04140	0.14680
2.5	0.07049	0.11035	0.04383	0.15525
3.0	0.08861	0.15021	0.05445	0.18194

Table 2.7: The mean squared error (MSE) optimal value of m in $\lambda = m\hat{\sigma}rn^{\alpha}$ are shown for $y = x_1 + x_2$ on a 10×10 grid with equally spaced covariates on $[0, 1]$.

m	0	.3	.5	1	1.5	2	2.5
$\sigma = .5$							
MSE	0.0421	0.0372	0.0352	0.0328	0.0324	0.0335	0.0357
MSE_1	0.1695	0.0573	<u>0.0420</u>	0.0619	0.1026	0.1471	0.1919
$MSE_{n/2}$	0.0489	0.0477	0.0463	0.0409	0.0340	0.0271	0.0212
MSE_n	0.1623	0.0555	<u>0.0415</u>	0.0635	0.1046	0.1490	0.1937
$\sigma = 1$							
MSE	0.12282	0.10059	0.09064	0.07458	0.06609	0.06252	0.06268
MSE_1	0.68460	0.21196	0.11333	<u>0.07487</u>	0.11552	0.17557	0.24223
$MSE_{n/2}$	0.14377	0.13315	0.12320	0.09447	0.06816	0.04787	0.03472
MSE_n	0.67713	0.20948	0.11222	<u>0.07377</u>	0.11467	0.17497	0.24147
$\sigma = 2$							
MSE	0.42190	0.32300	0.27771	0.20039	0.15399	0.12799	0.11750
MSE_1	2.88918	0.92777	0.49603	0.17091	<u>0.14513</u>	0.20483	0.30052
$MSE_{n/2}$	0.46964	0.41693	0.37026	0.25428	0.16418	0.10771	0.08082
MSE_n	2.95858	0.96689	0.52136	0.17461	<u>0.14395</u>	0.20300	0.29780

Table 2.8: The mean squared error (MSE) optimal value of m in $\lambda = m\hat{\sigma}rn^{\alpha}$ are shown for $y = x_1 + x_2 + x_1x_2$ on a 10×10 grid with equally spaced covariates on $[0, 1]$.

m	0	.3	.5	1	1.5	2	2.5
$\sigma = .5$							
MSE	0.05436	0.04869	0.04752	0.04887	0.05397	0.06163	0.07129
MSE_1	0.16386	<u>0.04366</u>	0.05028	0.11282	0.18608	0.25950	0.33165
$MSE_{n/2}$	0.07416	0.07384	0.07321	0.06893	0.06079	0.05028	0.03969
MSE_n	0.15971	<u>0.08881</u>	0.13932	0.30800	0.47736	0.63886	0.79276
$\sigma = 1$							
MSE	0.14540	0.11751	0.10774	0.09766	0.09990	0.11033	0.12707
MSE_1	0.68437	0.13722	<u>0.08286</u>	0.13069	0.23452	0.35025	0.46883
$MSE_{n/2}$	0.19401	0.18377	0.17199	0.13221	0.09184	0.06135	0.04562
MSE_n	0.66474	<u>0.16838</u>	0.18025	0.39074	0.64362	0.89434	1.13610
$\sigma = 2$							
MSE	0.45653	0.33146	0.28182	0.21054	0.18474	0.18951	0.21794
MSE_1	2.89485	0.59498	0.26586	<u>0.15172</u>	0.27271	0.45488	0.66203
$MSE_{n/2}$	0.57999	0.51017	0.44164	0.27027	0.15435	0.10159	0.10329
MSE_n	2.79342	0.53354	<u>0.29836</u>	0.41878	0.76140	1.14348	1.53057

Table 2.9: The mean squared error (MSE) optimal value of m in $\lambda = m\hat{\sigma}rn^{\alpha}$ are shown for $y = \max(\max(x_1, x_2) - 1/2, 0)$ on a 10×10 grid with equally spaced covariates on $[0, 1]$.

m	0	.3	.5	1	1.5	2	2.5
$\sigma = .5$							
MSE	0.05645	0.04940	0.04686	0.04441	0.04561	0.04941	0.05530
MSE_1	0.19650	0.05242	0.02700	<u>0.01085</u>	0.01317	0.02219	0.03508
$MSE_{n/2}$	0.03397	0.02902	0.02604	0.02474	0.03346	0.05048	0.07391
MSE_n	0.19208	0.05018	0.02598	<u>0.01487</u>	0.02568	0.04568	0.07127
$\sigma = .5$							
MSE	0.16657	0.13751	0.12601	0.11077	0.10762	0.11296	0.12496
MSE_1	0.77456	0.21210	0.11144	0.03793	<u>0.03398</u>	0.05404	0.08710
$MSE_{n/2}$	0.11330	0.09398	0.08120	0.06611	0.07911	0.11431	0.16501
MSE_n	0.76000	0.20566	0.10599	<u>0.04294</u>	0.05670	0.09851	0.15486
$\sigma = .5$							
MSE	0.50639	0.38410	0.33335	0.25715	0.22509	0.22305	0.24427
MSE_1	3.14057	0.87821	0.46271	0.13573	<u>0.08276</u>	0.11772	0.19593
$MSE_{n/2}$	0.40372	0.31771	0.25709	0.16320	0.15751	0.21744	0.32168
MSE_n	3.14928	0.87723	0.45639	0.13855	<u>0.10985</u>	0.17711	0.29200