Topics On Change–Point Estimation Under Adaptive Sampling Procedures

by
Yan Lan

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2007

Doctoral Committee:

Associate Professor Moulinath Banerjee, Co-Chair
Associate Professor George Michailidis, Co-Chair
Associate Professor Bin Nan
Assistant Professor Ji Zhu

To my parents

ACKNOWLEDGEMENTS

My foremost thanks go to my advisors Dr. George Michailidis and Dr. Moulinath Banerjee. I thank them for their advice, patience, support, encouragement, insights and suggestions that helped shape my research skills. This thesis would not be possible without them.

I am grateful to my two other committee members, Dr. Bin Nan and Dr. Ji Zhu, for reviewing my work and providing me with valuable comments.

I would like to extend a word of acknowledgement to the other faculty and the staff in the statistics department for their support and friendship. I also give my gratitude to all my fellow students, especially Xiao Wang, Jing Wang, Lili Yang, Ou Zhao, Anupap Somboonsavatdee, Harsh Singhal, Aijun Zhang, Min Qian, Nengfeng Zhou and Runlong Tang for the various ways in which they contributed to my work and for making the department a pleasant place to be, which has made graduate study all the more enjoyable.

Finally, I would like to thank my parents in China for their long-term support and encouragement. They are the most giving people and have been behind me every step of the way.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1

Motivation and Literature Review

1.1 Change–Point Estimation

The problem of estimating the location of a jump discontinuity (change-point) in an otherwise smooth curve has been extensively studied in the nonparametric regression and survival analysis literature; see for example Dempfle and Stute (2002), Gijbels et al. (1999), Gregoire and Hamrouni (2002), Hall and Molchanov (2003), Kosorok and Song (2007), Loader (1996), Müller (1992), Müller and Song (1997), Pons (2003), Ritov (1990) and references therein. In the classical setting, measurements on all n covariate-response pairs are available *in advance*, and the main issue is to estimate as accurately as possible the location of the change-point. However, there are applications where it is possible to sample the response at any covariate value of the experimenter's choice. The only hard constraint is that the total budget of measurements to be obtained is fixed a priori.

For example, consider the following example from system engineering. There is a stochastic flow of jobs/customers of various types arriving to the system with random service requests. Jobs waiting to be served are placed in queues of infinite capacity. The system's resources are allocated to the various job classes (queues) according to some service policy. This system serves as a canonical queueing model for many applications, including network switches, flexible manufacturing systems, wireless communications, etc. (Hung and Michailidis (2007)). A quantity of great

interest to the system's operator is the average delay of the customers, which is a key performance metric of the quality of service offered by the system.

The average delay of the customers in a two class system as a function of its loading, for a resource allocation policy introduced and discussed in Hung and Michailidis (2007), is shown in Figure 1.1. Specifically, the system was simulated under 134 loading settings and fed by input/service request processes obtained from real network traces and the average delay of 500,000 customers recorded. It can be seen that for loading around 0.8 there is a marked discontinuity in the response, which indicates that under the specified resource allocation policy the service provided to the customers deteriorates. It is of interest to locate the 'threshold' where such a change in the quality of service occurs. It should be pointed out that this threshold would occur at different system loadings for different allocation policies.

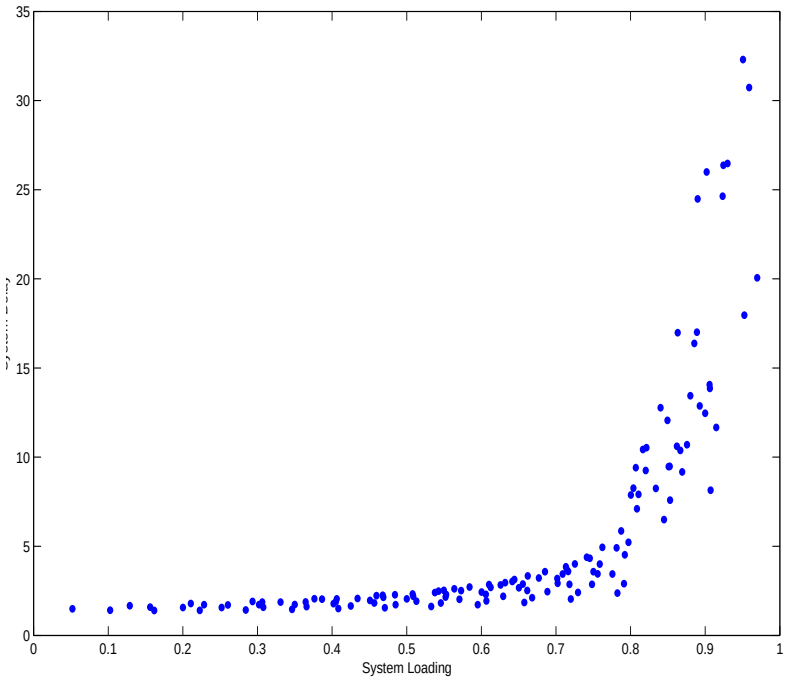


Figure 1.1: Average delay as a function of system loading for a two-class parallel processing system.

A few comments on the setting implied by this example are in order. First, the experimenter can *select* covariate values (in this case the system's loading) and subse-

quently obtain their corresponding sample responses. Second, the sampled responses are *expensive* to obtain; for example, the average delay is obtained by running a fairly large scale discrete event simulation of the system under consideration, involving half a million customers. For systems, comprised of a large number of customer classes, more computationally intensive simulations that can last days must be undertaken. Third, in many situations there is an *a priori fixed budget* of resources; for this example, it may correspond to CPU time, in other engineering applications to emulation time, while in other scientific contexts to real money.

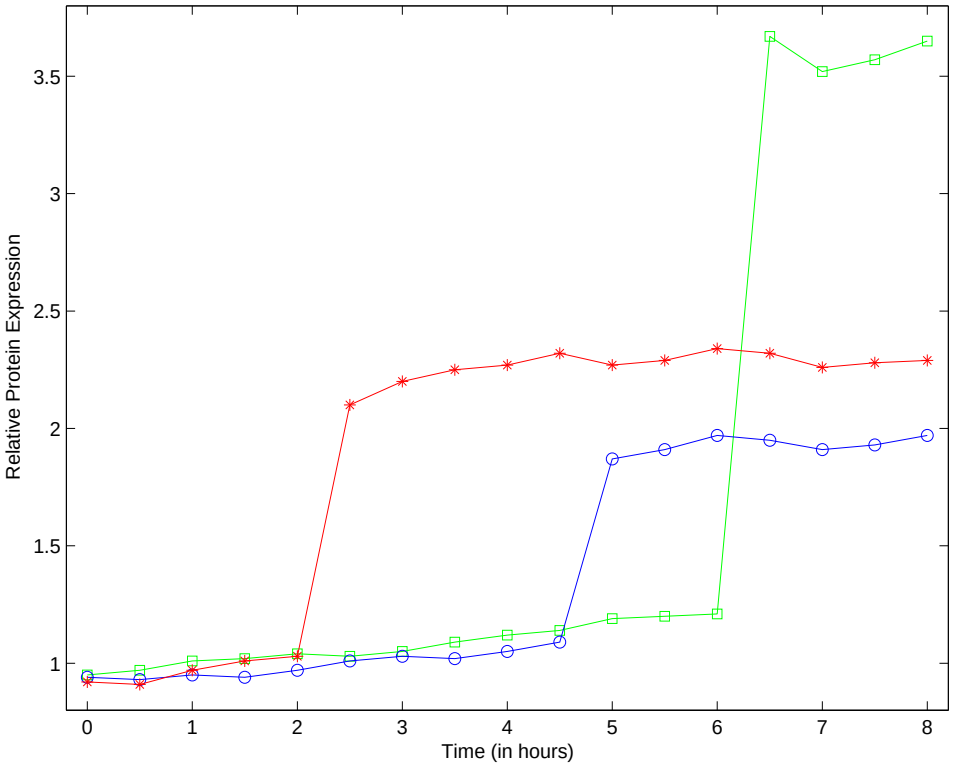


Figure 1.2: Protein expression over time in a bacterium

Another motivating example comes from biology. Three different protein expression levels in a bacterium as shown in Figure 1.2 are examined. Research scientists need to understand more about how protein expression changes over time, which is accomplished by designing appropriate experiments. Initially, the selection of time points is random and the results of the experiment are obtained and recorded. The goal is to identify when significant jump of protein expression level would happen

from the collected experiment data. At each time point, the experiment to observe the protein expression level will take a long time, which means that it is impossible to collect large data sets due to budget constraints. Let us look the protein expression in green (the lightest one) in the picture. Since no prior knowledge about the magnitude of the jump was available, the experiment was conducted every half hour. From the experiment results, it can be seen that the true jump may occur between 6th and 6.5th hours. It informs them that smaller time interval should be to selected. However, they do not have additional budget to do those experiments again. With the same amount of budget, if they selected the time data adaptively, the estimate of change-point over time at which significant jump may occur would be more accurate.

Given the potentially limited budget of points that can be sampled and lack of a priori knowledge about the location of the change-point the following strategy looks promising. A certain portion of the budget is used to obtain an initial estimate of the change-point based on a least squares criterion. Subsequently, a neighborhood around this initial estimate is specified and the remaining portion of the available points are sampled from it, together with their responses, that yield a new estimate of the change-point. Intuition suggests that if the first stage estimate is fairly accurate, the more intensive sampling in its neighborhood ought to produce a more accurate estimate than the one that would have been obtained by laying out the entire budget of points in a uniform fashion. Obviously, the procedure with its 'zoom-in' characteristics can be extended beyond two stages.

The goal of Chapter 3 is to formally introduce such multistage adaptive procedures for change-point estimation and examine their properties. In particular, the following important issues are studied and resolved: (i) the selection of the size of the neighborhoods, (ii) the rate of convergence of the multi-stage least squares estimate, together with its asymptotic distribution and (iii) allocation of the available budget at each stage.

The proposed procedure should be contrasted with the well studied sequential techniques for change-point detection, since the underlying setting exhibits marked differences. In its simplest form, the sequential change-point detection problem can be formulated as follows: there is a process that generates a sequence of independent observations $X_1, X_2, \cdots$ from some distribution F_0. At some unknown point in time τ, the distribution changes and hence observations $X_\tau, X_{\tau+1}, \cdots$ are generated from F_1. The objective is to raise an alarm as soon as the data generating mechanism switches to new distribution. This problem originally arose in statistical quality control and over the years has found important applications in other fields. Being a canonical problem in sequential analysis, many detection procedures have been proposed in the literature over the years in discrete and continuous time, under various assumptions on the distribution of τ and the data generating mechanism. The literature on this subject is truly enormous; a comprehensive treatment of the problem can be found in the book by Basseville and Nikiforov (1993), while some recent developments and new challenges are discussed in the review paper by Lai (2001). An important difference in our setting is the control that the experimenter exercises over the data generation process and also the absence of physical time, a crucial element in the sequential change-point problem.

Since parametric regression models with discontinuities are considered, we are usually interested in improving the performance of the entire regression functions , which leads us in developing adaptive strategies for estimating the regression function in Chapter 4. Two criterions are defined and corresponding optimal problems are studied and solved. The first criterion is related to the optimal allocation of samples, while the second one is used to select the optimal allocation of the available budget at the first stage.

1.2 Jump Boundary Curve Detection

In Section 1.1, we discussed jump discontinuities in an otherwise smooth function. We briefly review the extension to 2-dimensional surfaces and their jump boundary curves (as shown in Figure 1.3)

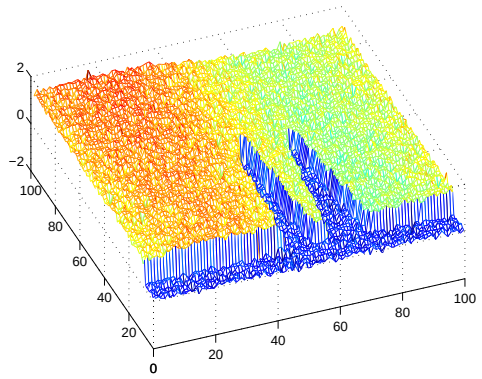


Figure 1.3: Jump Boundary Curve

Surface fitting is a fundamental problem in many applications. For example, meteorologists are interested in fitting the equi-temperature surfaces in high sky or deep ocean. Geologists often need to recover the mine surfaces from mineral samples. In some situations, the related surface is discontinuous and the locations of discontinuity are curves in the design space, called *jump boundary curves*. Of particular interest is the 3-dimensional problem defined as follows due to its applications in geology (Qiu(2002)) :

$$y = f_1(x_1, x_2)I_{\{x_2 \geq g(x_1)\}} + f_2(x_1, x_2)I_{\{x_2 < g(x_1)\}} + \varepsilon$$

where y is the response surface defined by 2-dimensional step functions, $f_1(x_1, x_w)$ and $f_2(x_1, x_2)$, respectively, with $g(\cdot)$ being the jump boundary curve we are interested in identifying and estimating and ε a mean zero, homoskedastic error term.

We are interested in developing computationally efficient methodology that detects and fits the underlying boundary curve, which usually has a complex shape.

The goal is to start with a small initial design, which is subsequently refined, as more information about the boundary curve is obtained. The methodology is illustrated on a number of simulated data sets.

The identification of a jump boundary curve has attracted some attention over the last few years and we provide next a review of some of the approaches proposed in the literature. It should be noted that the edge detection problem in image processing has a similar flavor to the one under study and hence some approaches prove relevant. Chu et al. (1998) looked at edge detection problems and their approach is based on M-estimation techniques. Qiu (2002) looks at identifying the jump boundary curve through local smoothing techniques, while Hall and Molchanov (2003) propose a sequential method for the same problem. Some other works include spline smoothing procedures and wavelet transformation method (see Qiu(2005)). We give a more detailed summary of these approaches in Chapter 5.

CHAPTER 2

Preliminaries and The Classical Procedure

2.1 The Classical Problem

In this study, we focus on parametric models for the regression function of the type:

$$Y_i = \mu(X_i) + \epsilon_i, \quad i = 1, 2, \ldots, n$$

where

$$\mu(x) = \psi_l(\beta_l, x)1(x \leq d^0) + \psi_u(\beta_u, x)1(x > d^0) \tag{2.1}$$

with $\psi_l(\beta_l, x)$ and $\psi_u(\beta_u, x)$ both (at least) twice continuously differentiable in β and infinitely differentiable in x and $\psi_l(\beta_l, d^0) \neq \psi_u(\beta_u, d^0)$, so that d^0 is the unique point of discontinuity – a change point – of the regression function.

The ϵ_i's are assumed to be i.i.d. symmetric mean 0 errors with common (unknown) error variance σ^2 and are independent of the X_i's which are i.i.d. and are distributed on $[0, 1]$ according to some common density $f_X(\cdot)$. The simplest possible parametric candidate for $\mu(x)$, which we will focus on largely to illustrate the key ideas in the thesis, is the simple step function: $\mu(x) = \alpha_0\, 1(x \leq d^0) + \beta_0\, 1(x > d^0)$.

Estimating d^0 based on the above data is coined as the "classical problem". A standard way to estimate the parameters (β_l, β_u, d^0) is to solve a least squares problem. We start by introducing some necessary notation. Let $\mathbb{P}_n$ denote the empirical distribution of the data vector $\{X_i, Y_i\}_{i=1}^n$ and P the true distribution of (X_1, Y_1).

For a function f defined on the space $[0,1]\times\mathbb{R}$ (in which the vector (X_1, Y_1) assumes values) and a measure Q defined on the Borel σ–field on $[0,1]\times\mathbb{R}$, we denote $\int f\, dQ$ as Qf. We now turn our attention to the least squares problem.

The objective is to minimize $\mathbb{P}_n\left[(y-\psi_l(\alpha,x))^2\, 1(x\le d)+(y-\psi_u(\beta_u,x))^2\, 1(x>d)\right]$ over all (α,β,d), with $0\le d\le 1$. Let $(\tilde{\beta}_{l,n}, \tilde{\beta}_{u,n}, \tilde{d}_n)$ denote a vector of minimizers. Note that we refer to "a vector" of minimizers, since there will be, in general, multiple tri–vectors that minimize the criterion function. The asymptotic properties of such a vector can be studied by invoking either the methods of Pons (2003) or those (in Chapter 14) of Kosorok (2006) and Kosorok and Song (2006). We do not provide the details, but state the results that are essential to the *multistage learning procedures* that we formulate in the next chapter. We clarify next the meaning of a minimizer of a right–continuous real–valued function with left limits (say f) defined on an interval I. Specifically, any point $z\in I$ that satisfies $f(z)\wedge f(z-) = \min_{w\in I} f(w)$ is defined to be a minimizer of f. Also, in order to discuss the results for the classical procedure and those for the proposed multistage procedures, we need to define a family of compound Poisson processes that arise in the description of the asymptotic properties of the estimators of the change point.

A family of compound Poisson processes: For a positive constant Λ, let $\nu^+(\cdot)$ be a Poisson process on $[0,\infty)$ with right continuous sample paths, with $\nu^+(s)\sim \text{Poi}(\Lambda s)$ for $s>0$. Let $\tilde{\nu}^+(\cdot)$ be another independent Poisson process on $[0,\infty)$ with left–continuous sample paths, with $\tilde{\nu}^+(s)\sim \text{Poi}(\Lambda s)$ and define a (right–continuous) Poisson process on $(-\infty,0]$ by $\{\nu^-(s) = -\tilde{\nu}^+(-s) : s\in(-\infty,0]\}$. Let $\{\eta_i\}_{i=1}^{\infty}$ and $\{\eta_{-i}\}_{i=1}^{\infty}$ be two independent sequences of i.i.d. random variables where each η_j (j assumes both positive and negative values) is distributed like η, η being a symmetric random variable with finite variance ρ^2. Given a positive constant A, define families of random variables $\{V_i^+\}_{i=1}^{\infty}$ and $\{V_i^-\}_{i=1}^{\infty}$ where, for each $i\ge 1$, $V_i^+ = A/2+\eta_i$ and $V_i^- = -A/2+\eta_{-i}$. Set $V_0^+ = V_0^- \equiv 0$. Next, define compound Poisson processes

$\mathbb{M}_1$ and $\mathbb{M}_2$ on $(-\infty, \infty)$ as follows: $\mathbb{M}_1(s) = (\sum_{0 \le i \le \nu^+(s)} V_i^+)\, 1(s \ge 0)$ and $\mathbb{M}_2(s) = (\sum_{0 \le i \le \nu^-(s)} V_i^-)\, 1(s \le 0)$. Finally, define the two–sided compound Poisson process $\mathbb{M}_{A,\eta,\Lambda}(s) = \mathbb{M}_1(s) - \mathbb{M}_2(s)$. It is not difficult to see that $\mathbb{M}$, almost surely, has a minimizer (in which case it has multiple minimizers, since the sample paths are piecewise constant). Let $d_l(A, \eta, \Lambda)$ denote the smallest minimizer of $\mathbb{M}_{A,\eta,\Lambda}$ and $d_u(A, \eta, \Lambda)$ its largest one, which are, almost surely, well defined. Then, the following relation holds:

$$(d_l(A,\eta,\Lambda), d_u(A,\eta,\Lambda)) \equiv_d \left(d_l\left(\frac{A}{\rho}, \frac{\eta}{\rho}, \Lambda\right), d_u\left(\frac{A}{\rho}, \frac{\eta}{\rho}, \Lambda\right)\right) \tag{2.2}$$

$$\equiv_d \frac{1}{\Lambda}\left(d_l\left(\frac{A}{\rho}, \frac{\eta}{\rho}, 1\right), d_u\left(\frac{A}{\rho}, \frac{\eta}{\rho}, 1\right)\right). \tag{2.3}$$

For the"classical problem", the following proposition holds.

Proposition 2.1. *Consider the model described at the beginning of Chapter 2. Suppose that X has a positive bounded density on $[0,1]$ and that d^0 is known to lie in the interval $[\epsilon_0, 1-\epsilon_0]$ for some small $\epsilon_0 > 0$. Let $(\hat\beta_{l,n}, \hat\beta_{u,n}, \hat d_n)$ denote that minimizing tri–vector $(\tilde\beta_{l,n}, \tilde\beta_{u,n}, \tilde d_n)$, for which the third component is minimal. Then, $(\sqrt{n}(\hat\beta_{l,n} - \beta_l), \sqrt{n}(\hat\beta_{u,n} - \beta_u), n(\hat d_n - d^0))$ is $O_p(1)$. Furthermore, the first two components of this vector are asymptotically independent of the third and*

$$\begin{aligned} n(\hat d_n - d^0) &\to_d d_l(|\ \mu(d^0+) - \mu(d^0)\ |, \epsilon_1, f_X(d^0)) \\ &\equiv_d \frac{1}{f_X(d^0)}\, d_l\left(\frac{|\ \mu(d^0+) - \mu(d^0)\ |}{\sigma}, \frac{\epsilon_1}{\sigma}, 1\right). \end{aligned}$$

Heteroscedastic errors: The proposition can be generalized readily to cover the case of heteroscedastic errors. A generalization of the classical model to the heteroscedastic case is as follows: We observe n i.i.d. observations from the model $Y = \mu(X) + \sigma(X)\,\tilde\epsilon$ where $\mu(x)$ is as defined in (2.1), $\tilde\epsilon$ and X are independent, $\tilde\epsilon$ is symmetric about 0 with unit variance and $\sigma^2(x)$ is a variance function (assumed continuous). As in the homoscedastic case, an unweighted least squares procedure is used to estimate the parameters (β_l, β_u, d^0). As before, letting $(\hat\beta_{l,n}, \hat\beta_{u,n}, \hat d_n)$ denote

that minimizing tri–vector $(\tilde{\beta}_{l,n}, \tilde{\beta}_{u,n}, \tilde{d}_n)$, for which the third component is minimal, we have:

$$\begin{aligned} n(\hat{d}_n - d^0) &\to_d d_l(\mid \mu(d^0+) - \mu(d^0) \mid, \sigma(d^0)\,\tilde{\epsilon}, f_X(d^0)) \\ &\equiv_d \frac{1}{f_X(d^0)}\, d_l\left(\frac{\mid \mu(d^0+) - \mu(d^0) \mid}{\sigma(d^0)}, \tilde{\epsilon}, 1\right). \end{aligned}$$

2.2 Weak Convergence of the Estimates from the Classical Procedure.

As described in Proposition 2.1, the weak convergence of the estimates from the classical procedure can be established. Since this proposition can be extended to the estimate from our proposed two-stage procedure, we leave those (similar) details for the next chapter. In this section, we outline some of the features of weak convergence in the classical model.

We start with the simple step function model:

$$\mu(x) = \alpha_0 I(x \le d^0) + \beta_0 I(x > d^0).$$

Our goal is to minimize

$$\mathbb{P}_n\left[(y-\alpha)^2\, 1(x \le d) + (y-\beta)^2\, 1(x > d)\right]$$

over all (α, β, d), with $0 \le d \le 1$). Let $(\hat{\alpha}_n, \hat{\beta}_n, \hat{d}_n)$ denote a vector of minimizers. We consider $\hat{d}_n$ as the minimal minimizer for all discussions in this section.

2.2.1 α_0 and β_0 are known

WLOG, we suppose $\beta_0 > \alpha_0$ for all discussions in this thesis. If α_0 and β_0 are known, the estimate of d^0 is obtained by,

$$\begin{aligned} \hat{d}_n &= \text{argmin } \mathbb{P}_n[(y-\alpha_0)^2 I(x \le d) + (y-\beta_0)^2 I(x > d)] \\ &= \text{argmin } \mathbb{P}_n[\{(y-\alpha_0)^2 - (y-\beta_0)^2\} I(x \le d)] \\ &= \text{argmin } \mathbb{P}_n\left[2\left(y - \frac{\alpha_0+\beta_0}{2}\right)(\beta_0-\alpha_0) I(x \le d)\right] \end{aligned}$$

Proposition 2.2. *We have:*

$$n(\hat{d}_n - d^0) \rightarrow_d d_l(\mid \mu(d^0+) - \mu(d^0) \mid, \epsilon_1, f_X(d^0)) \equiv d_l(\mid \alpha_0 - \beta_0 \mid, \epsilon_1, f_X(d^0)).$$

Kosorok (2006) verified that

$$n(\hat{d}_n - d^o) = O_p(1)$$

The intuition for the rate is as follows: Suppose there is an interval [0,1] and we distribute n points on an evenly spaced grid. The resolution of the resulting grid is $1/n$ and this determines the rate of convergence of the estimate.

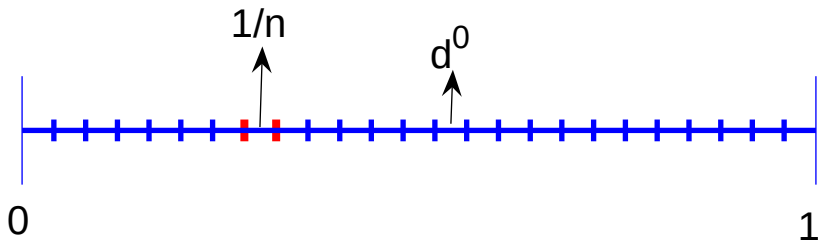


Figure 2.1: Rate of convergence of the estimate from classical procedure

Introduce the local variable $s \equiv n(d - d^0)$ and consider the normalized process:

$$\begin{aligned} \mathbb{M}_n(s) &\triangleq \sum_{i=1}^{n} \left(y_i - \frac{\alpha_0 + \beta_0}{2} \right) \left(I\left(x_i \leq d^o + \frac{s}{n} \right) - I(x_i \leq d^o) \right) \\ &= \mathbb{M}_n^+(s) - \mathbb{M}_n^-(s) \end{aligned}$$

where

$$\mathbb{M}_n^+(s) = \mathbb{M}_n(s)1(s \geq 0) \quad \text{and} \quad \mathbb{M}_n^-(s) = -\mathbb{M}_n(s)1(s \leq 0).$$

For this simple model, we use the same notation as in Section 2.1. Set $A = \mid \beta_0 - \alpha_0 \mid$, $\Lambda = f_X(d^0)$, and let η have the same distribution as ϵ_1, so that $Var(\eta) = \sigma^2$. Consider the compound Poisson process $\mathbb{M}(A, \eta, \Lambda)$. It is easy to see that V_1^+ and V_1^{-1} have characteristic functions $L^+(t) = E[e^{it\left(Y - \frac{\beta_0+\alpha_0}{2}\right)} \mid X = d^{0+}]$ and $L^-(t) = E[e^{it\left(Y - \frac{\beta_0+\alpha_0}{2}\right)} \mid X = d^0]$ respectively. It can be shown that on every compact interval, the process $\mathbb{M}_n$ converges weakly to $\mathbb{M}_{A,\eta,\Lambda}$ in the Skorokhod topology; to this end,

one needs to establish the weak convergence of the finite dimensional distributions of $\mathbb{M}_n$ to those of $(\mathbb{M}_{A,\eta,\Lambda}, \mathbb{J})$ and also the tightness of $\mathbb{M}_n$ (restricted to the compact interval) as a process. Since $n(\hat{d}_n - d^0) = O_p(1)$, one would like to invoke a continuous mapping argument and conclude that $\hat{h}_3 \equiv n(\hat{d}_n - d^0)$, the smallest minimizer of $\mathbb{M}_n$ converges to the smallest minimizer of $(\mathbb{M}_{A,\eta,\Lambda}, \mathbb{J})$. However, the convergence of $\mathbb{M}_n$ in the Skorohod topology is not sufficient to guarantee the convergence of the argmin, since the limit process does not have a unique minimizer. Distributional convergence of $\mathbb{M}_n$ needs to be established in a stronger sense, one that involves convergence of the pure jump process corresponding to $\mathbb{M}_n$ to that corresponding to $\mathbb{M}$. Since the details of this type of an argument are provided in connection with the two–stage procedure in the next section, we not harp on the details here. We only show the proof of one-dimensional weak convergence of $\mathbb{M}_n$ and skip the remainder.

Suppose $s \in [0, K]$, we define

$$\begin{aligned} Q_n(t) &= E[e^{it\mathbb{M}_n^+(s)}] \\ &= E\left[e^{it\left(Y - \frac{\alpha_0+\beta_0}{2}\right)\left(I\left(X \le d^0 + \frac{s}{n}\right) - I(X \le d^o)\right)}\right]^n \end{aligned}$$

Consider:

$$E\left[\exp\left(it\left[\left(Y - \frac{\alpha_0+\beta_0}{2}\right)\left(I\left(X \le d^0 + \frac{s}{n}\right) - I(X \le d^0)\right)\right]\right)\right]$$

$$\begin{aligned} &= \int_{d^0}^{d^0+\frac{s}{n}} E\left[\exp\left(it\left(Y - \frac{\alpha_0+\beta_0}{2}\right)\right)\Big| X = x\right] p_X(x)dx + \int_{(0,d^0]\bigcup(d^0+\frac{s}{n},1)} p_X(x)dx \\ &= 1 - \int_{d^0}^{d^0+\frac{s}{n}} p_X(x)dx + u\left(d^0 + \frac{s}{n}\right) - u(d^0) \\ &= 1 - \left(F_X\left(d^0 + \frac{s}{n}\right) - F_X(d^0)\right) + \left(u\left(d^0 + \frac{s}{n}\right) - u(d^0)\right) \end{aligned}$$

where

$$u(w) = \int_0^w E\left[\exp\left(it\left(Y - \frac{\alpha_0+\beta_0}{2}\right)\right)\Big| X = x\right] p_X(x)dx$$

So

$$Q_n(t) = \left[1 - \left(F_X\left(d^0 + \frac{s}{n}\right) - F_X(d^0)\right) + \left(u\left(d^0 + \frac{s}{n}\right) - u(d^0)\right)\right]^n$$

$$F_X\left(d^0+\frac{s}{n}\right)-F_X(d^0)=f_X(d^0)\frac{s}{n}+O\left(\frac{1}{n^2}\right) \qquad f_X\equiv p_X$$

$$u\left(d^0+\frac{s}{n}\right)-u(d^0)=\frac{s}{n}u'(d^0)+O\left(\frac{1}{n^2}\right)$$

$$u'(d^0)=f_X(d^0)E[e^{it\left(Y-\frac{\alpha_0+\beta_0}{2}\right)}|X=d^0]=f_X(d^0)L^+(t).$$

So:

$$\begin{aligned} Q_n(t) &= \left[1-f_X(d^0)\frac{s}{n}+\frac{s}{n}f_X(d^0)L(t)+O\left(\frac{1}{n^2}\right)\right]^n \\ &= \left[1-\frac{s}{n}f_x(d^0)(1-L^+(t))+O\left(\frac{1}{n^2}\right)\right]^n \\ &\longrightarrow \exp[-sf_X(d^0)(1-L^+(t))]=E[e^{it\mathbb{M}_1(s)}] \end{aligned}$$

since:

$$\begin{aligned} E[e^{it\mathbb{M}_1(s)}] &= E[e^{it\sum_{0\le k\le \nu^+(s)}V_k}] \\ &= \sum_{l=0}^{\infty}E[e^{it(V_1+\ldots+V_l)}]\frac{e^{-sf_X(d^0)}(sf_X(d^0))^l}{l!} \\ &= \sum_{l=0}^{\infty}\frac{(L^+(t))^l(sf_X(d^0))^l}{l!}e^{-sf_X(d^0)} \\ &= e^{-sf_X(d^0)e^{L^+(t)sf_X(d^0)}} \\ &= e^{-sf_X(d^0)(1-L^+(t))} \end{aligned}$$

Similarly, when $s\in[-K,0]$, we can obtain that

$$E\left[e^{it\mathbb{M}_n^-(s)}\right]\to E\left[e^{it\mathbb{M}_2(s)}\right].$$

Thus, one-dimensional weak convergence is established.

2.2.2 α_0 and β_0 are unknown

The estimate of (α_0,β_0,d^0) can be obtained as:

$$\begin{aligned} (\hat{\alpha_n},\hat{\beta}_n,\hat{d}_n) &= \text{argmin } \mathbb{P}_n[(y-\alpha)^2I(x\le d)+(y-\beta)^2I(x>d)] \\ &= \text{argmin } n\mathbb{P}_n[(y-\alpha)^2I(x\le d)+(y-\beta)^2I(x>d)] \\ &\triangleq \text{argmin } \tilde{\mathbb{M}}_n(\alpha,\beta,d) \end{aligned}$$

Check that

$$\begin{aligned}
& \operatorname{argmin} \tilde{\mathbb{M}}_n(\alpha, \beta, d) \\
= \; & \operatorname{argmin} \big\{ n\mathbb{P}_n[((y-\alpha)^2 - (y-\alpha_0)^2) I(x \le d)] \\
& + n\mathbb{P}_n[((y-\beta)^2 - (y-\beta_0)^2) I(x > d)] \\
& + n\mathbb{P}_n[(y-\alpha_0)^2 (I(x \le d) - I(x \le d^0)) + (y-\beta_0)^2 (I(x > d) - I(x > d^0))] \big\} \\
= \; & \operatorname{argmin} \big\{ n\mathbb{P}_n[((y-\alpha)^2 - (y-\alpha_0)^2) I(x \le d)] \\
& + n\mathbb{P}_n[((y-\beta)^2 - (y-\beta_0)^2) I(x > d)] \\
& + n\mathbb{P}_n[((y-\alpha_0)^2 - (y-\beta_0)^2)(I(x \le d) - I(x \le d^0))] \big\} \\
\triangleq \; & \operatorname{argmin} \{ I_{n,1} + I_{n,2} + I_{n,3} \} .
\end{aligned}$$

Also, note that

$$\begin{aligned}
(\alpha_0, \beta_0, d^0) &= \operatorname{argmin} \; P[(y-\alpha)^2 I(x \le d) + (y-\beta)^2 I(x > d)] \\
&= \operatorname{argmin} \big\{ nP[((y-\alpha)^2 - (y-\alpha_0)^2) I(x \le d)] \\
&\quad + nP[((y-\beta)^2 - (y-\beta_0)^2) I(x > d)] \\
&\quad + nP[((y-\alpha_0)^2 - (y-\beta_0)^2)(I(x \le d) - I(x \le d^0))] \big\}
\end{aligned}$$

From Chapter 14 of Kosorok (2006), we know that $n(\hat{d}_n - d^0) = O_p(1), \sqrt{n}(\hat{\alpha_n} - \alpha_0) = O_p(1)$ and $\sqrt{n}(\hat{\beta}_n - \beta_0) = O_p(1)$. As before, introduce local variables (h_1, h_2, h_3) such that $(\alpha, \beta, d) = (\alpha_0 + h_1/\sqrt{n}, \beta_0 + h_2/\sqrt{n}, d^0 + h_3/n)$. Consider $\tilde{\mathbb{M}}_n$ now as a function of (h_1, h_2, h_3), i.e. consider $\tilde{\mathbb{M}}_n(\alpha_0 + h_1/\sqrt{n}, \beta_0 + h_2/\sqrt{n}, d^0 + h_3/n)$. Then

$$(\hat{h}_1, \hat{h}_2, \hat{h}_3) \equiv (\sqrt{n}(\hat{\alpha_n} - \alpha_0), \sqrt{n}(\hat{\beta}_n - \beta_0), n(\hat{d}_n - d^0)),$$

which minimizes $\tilde{\mathbb{M}}_n$ is $O_p(1)$. On every compact rectangle $[-K, K] \times [-K, K] \times [-K, K]$, the process $\tilde{\mathbb{M}}_n$ is asymptotically equivalent to the process $\mathbb{M}_n^\star$ in an ap-

propriate metric, where

$$\begin{aligned}\mathbb{M}_n^\star &= -2\mathbb{P}_n\sqrt{n}h_1(y-\alpha_0)I(x\le d^o)+h_1^2F_X(d^0)\\ &\quad -2\mathbb{P}_n\sqrt{n}h_2(y-\beta_0)I\left(x>d^0+\frac{h_3}{n}\right)+h_2^2(1-F_X(d^0))\\ &\quad +n\mathbb{P}_n\left[2(\beta_0-\alpha_0)\left(y-\frac{\alpha_0+\beta_0}{2}\right)(I(x\le d^0+h_3/n)-I(x\le d^0))\right].\end{aligned}$$

We sketch below a semi–heuristic argument. Consider,

$$\begin{aligned}I_{n,1}(h_1,h_2,h_3) &= n\mathbb{P}_n[((y-\alpha)^2-(y-\alpha_0)^2)I(x\le d)]\\ &= n\mathbb{P}_n[(2y-\alpha-\alpha_0)(\alpha_0-\alpha)I(x\le d)]\\ &= \sum_{i=1}^n\left(2y_i-\alpha_0-\frac{h_1}{\sqrt{n}}-\alpha_0\right)\left(-\frac{h_1}{\sqrt{n}}\right)I\left(x_i\le d^o+\frac{h_3}{n}\right)\\ &= -\frac{1}{\sqrt{n}}h_1\sum_{i=1}^n\left(2y_i-2\alpha_0-\frac{h_1}{\sqrt{n}}\right)I\left(x_i\le d^o+\frac{h_3}{n}\right)\\ &= -2\sqrt{n}h_1\cdot\frac{1}{n}\sum_{i=1}^n(y_i-\alpha_0)I\left(x_i\le d^o+\frac{h_3}{n}\right)\\ &\quad +h_1^2\frac{1}{n}\sum_{i=1}^n I\left(x_i\le d^o+\frac{h_3}{n}\right)\end{aligned}$$

Now,

$$\begin{aligned}&P\sqrt{n}\left[\left(y-\alpha_0)(I\left(x\le d^o+\frac{h_3}{n}\right)-I(x\le d^o)\right)\right]\\ &= \sqrt{n}\int_{d^0}^{d^0+\frac{h_3}{n}}(\beta_0-\alpha_0)\,p_X(x)dx\\ &= \sqrt{n}(\beta_0-\alpha_0)\left[F\left(d^o+\frac{h_3}{n}\right)-F(d^o)\right]\\ &= \sqrt{n}(\beta_0-\alpha_0)\left(\frac{h_3}{n}f_X(d^o)+O\left(\frac{1}{n^2}\right)\right)\longrightarrow 0,\end{aligned}$$

and hence

$$\begin{aligned} I_{n,1}(h_1,h_2,h_3) &= -2\sqrt{n}h_1\mathbb{P}_n(y-\alpha_0)I(x\le d^0)+h_1^2\frac{1}{n}\sum_{i=1}^{n}I\left(x_i\le d^0+\frac{h_3}{n}\right)\\ &\quad +2\sqrt{n}h_1\mathbb{P}_n(y-\alpha_0)\left[I\left(x\le d^0+\frac{h_3}{n}\right)-I(x\le d^0)\right]\\ &= -2\mathbb{P}_n\sqrt{n}h_1(y-\alpha_0)I(x\le d^0)+h_1^2\frac{1}{n}\sum_{i=1}^{n}I\left(x_i\le d^0+\frac{h_3}{n}\right)\\ &\quad +2P\sqrt{n}h_1(y-\alpha_0)\left[I\left(x\le d^0+\frac{h_3}{n}\right)-I(x\le d^0)\right]\\ &\quad +2(\mathbb{P}_n-P)\sqrt{n}h_1(y-\alpha_0)\left[I\left(x\le d^0+\frac{h_3}{n}\right)-I(x\le d^0)\right]\\ &= -2\mathbb{P}_n\sqrt{n}h_1(y-\alpha_0)I(x\le d^0)+h_1^2F_X(d^0)+o_P^K(1). \end{aligned}$$

Similarly,

$$\begin{aligned} I_{n,2}(h_1,h_2,h_3) &= n\mathbb{P}_n[((y-\beta)^2-(y-\beta_0)^2)I(x>d)]\\ &= -2\mathbb{P}_n\sqrt{n}h_2(y-\beta_0)I\left(x>d^0+\frac{h_3}{n}\right)+h_2^2\frac{1}{n}\sum_{i=1}^{n}I\left(x_i>d^0+\frac{h_3}{n}\right)\\ &= -2\mathbb{P}_n\sqrt{n}h_2(y-\beta_0)I\left(x>d^0+\frac{h_3}{n}\right)+h_2^2(1-F_X(d^0))+o_P^K(1) \end{aligned}$$

and

$$\begin{aligned} I_{n,3}(h_1,h_2,h_3) &= n\mathbb{P}_n[((y-\alpha_0)^2-(y-\beta_0)^2)(I(x\le d)-I(x\le d^0))]\\ &= n\mathbb{P}_n\left[2(\beta_0-\alpha_0)\left(y-\frac{\alpha_0+\beta_0}{2}\right)(I(x\le d^0+h_3/n)-I(x\le d^0))\right]. \end{aligned}$$

It can then be deduced that on every $[-K,K]\times[-K,K]\times[-K,K]$ $\mathbb{M}_n^\star \to_d Q \equiv -2h_1Z_1+h_1^2F_X(d^0)-2h_2Z_2+h_2^2(1-F_X(d^0))+M_{A,\eta,\Lambda}(h_3)$ as a process, where Z_1, Z_2 and $M_{A,\eta,\Lambda}$ are all independent, $M_{A,\eta,\Lambda}$ is a compound Poisson process with $A=\mid \beta_0-\alpha_0 \mid$, $\Lambda=f_X(d^0)$, and $Var(\eta)=\sigma^2$. Z_1 and Z_2 are mean zero Gaussian with respective variances $\sigma^2F_X(d^0)$ and $\sigma^2(1-F_X(d^0))$. So the process $\tilde{\mathbb{M}}_n$ restricted to $[-K,K]\times[-K,K]\times[-K,K]$ has the same distributional limit. Continuous mapping arguments then yield that $(\hat{h}_1,\hat{h}_2,\hat{h}_3)\to_d(h_1^{lim},h_2^{lim},h_3^{lim})$ where $(h_1^{lim},h_2^{lim})=\text{argmin}_{h_1,h_2}[-2h_1Z_1+h_1^2F_X(d^0)-2h_2Z_2+h_2^2(1-F_X(d^0))]$ while $h_3^{lim}=d_l(A,\eta,\Lambda)$.

The independence of Z_1, Z_2 and $M_{A,\eta,\Lambda}$ is a consequence of the asymptotic independence of the processes:

$$\begin{aligned}
\mathbb{A}_n(h_1) &= \sum_{i=1}^{n} \frac{h_1}{\sqrt{n}}(y_i - \alpha_0) I(x_i \leq d^0) \\
\mathbb{B}_n(h_2) &= \sum_{i=1}^{n} \frac{h_2}{\sqrt{n}}(y_i - \beta_0) I\left(x_i > d^0 + \frac{h_3}{n}\right) \\
\mathbb{C}_n(h_3) &= \sum_{i=1}^{n} \left(y_i - \frac{\alpha_0 + \beta_0}{2}\right)\left[I\left(x_i \leq d^0 + \frac{h_3}{n}\right) - I(x_i \leq d^0)\right]
\end{aligned}$$

We conclude this section by sketching how asymptotic independence is established. To keep things simple, we just consider one dimensional marginals (a full proof would require consideration of finite dimensional marginals of all possible orders). Consider

$$\begin{aligned}
\hat{F}_n(t) &= E[\exp\{it(a\mathbb{A}_n + b\mathbb{B}_n + c\mathbb{C}_n)\}] \\
&= \left[E\left[\exp\left\{it\left[\frac{ah_1}{\sqrt{n}}(Y - \alpha_0) I(X \leq d^o) + \frac{bh_2}{\sqrt{n}}(Y - \beta_0) I\left(X > d^o + \frac{h_3}{n}\right)\right.\right.\right.\right. \\
&\quad \left.\left.\left.\left. + \; c\left(Y - \frac{\alpha_0 + \beta_0}{2}\right)\left(I\left(X \leq d^o + \frac{h_3}{n}\right) - I(X \leq d^o)\right)\right]\right\}\right]\right]^n \\
&\equiv J^n
\end{aligned}$$

where

$$\begin{aligned}
J &= \int E\left[\exp\left\{it\left[\frac{ah_1}{\sqrt{n}}(Y - \alpha_0) I(X \leq d^0) + \frac{bh_2}{\sqrt{n}}(Y - \beta_0) I\left(X > d^0 + \frac{h_3}{n}\right)\right.\right.\right. \\
&\quad \left.\left.\left. + c\left(Y - \frac{\alpha_0 + \beta_0}{2}\right)\left(I\left(X \leq d^0 + \frac{h_3}{n}\right) - I(X \leq d^0)\right)\right]\right\} \Big| X = x\right] p_X dx \\
&= \int_0^{d^0} E[e^{it\left(\frac{ah_1}{\sqrt{n}}(Y - \alpha_0)\right)} | X = x] p_X dx \\
&\quad + \int_{d^0}^{d^o + \frac{h_3}{n}} E[e^{it\left(c\left(Y - \frac{\alpha_0 + \beta_0}{2}\right)\right)} | X = x] p_X dx \\
&\quad + \int_{d^0 + \frac{h_3}{n}}^{1} E[e^{it\left(\frac{bh_2}{\sqrt{n}}(Y - \beta_0)\right)} | X = x] p_X dx \\
&\triangleq J_1 + J_2 + J_3
\end{aligned}$$

where

$$\begin{aligned} J_1 &= \int_0^{d^0} \left[1 - \frac{1}{2}\left(\frac{ah_1 t}{\sqrt{n}}\right)^2 \sigma^2 + O\left(\frac{t^2}{n}\right)\right] p_X dx \\ &= \left[1 - \frac{a^2 {h_1}^2 \sigma^2 t^2}{2n} + O\left(\frac{t^2}{n}\right)\right] F_X(d^0) \\ J_3 &= \int_{d^0+\frac{h_3}{n}}^{1} \left[1 - \frac{1}{2}\left(\frac{bh_2 t}{\sqrt{n}}\right)^2 \sigma^2 + O\left(\frac{t^2}{n}\right)\right] p_X dx \\ &= \left[1 - \frac{b^2 {h_2}^2 \sigma^2 t^2}{2n} + O\left(\frac{t^2}{n}\right)\right] \left[1 - F_X\left(d^0 + \frac{h_3}{n}\right)\right] \end{aligned}$$

For J_2,

$$\begin{aligned} J_2 &= \int_0^{d^0+\frac{h^3}{n}} E[e^{itc\left(Y - \frac{\alpha_0+\beta_0}{2}\right)} | X = x] p_X dx - \int_0^{d^0} E[e^{itc\left(Y - \frac{\alpha_0+\beta_0}{2}\right)} | X = x] p_X dx \\ &= u\left(d^0 + \frac{h_3}{n}\right) - u(d^0) \qquad \text{where } \; u(w) = \int_0^w E(e^{itc\left(Y - \frac{\alpha_0+\beta_0}{2}\right)} | X = x) p_X dx \\ &= \frac{h_3}{n} u'(d^0) + O\left(\frac{1}{n^2}\right) \\ &= \frac{h_3}{n} f_X(d^0) E(e^{itc\left(Y - \frac{\alpha_0-\beta_0}{2}\right)} | X = d^0) + O\left(\frac{1}{n^2}\right) \\ &= \frac{h_3}{n} f_X(d^0) L^+(ct) + O\left(\frac{1}{n^2}\right) \end{aligned}$$

Then

$$\begin{aligned} \hat{F}_n(t) &= \left[\left(1 - \frac{a^2 {h_1}^2 \sigma^2 t^2}{2n}\right) F_X(d^0) + \left(1 - \frac{b^2 {h_2}^2 \sigma^2 t^2}{2n}\right)\left(1 - F_X\left(d^0 + \frac{h_3}{n}\right)\right)\right. \\ &\quad \left. + \frac{h_3}{n} f_X(d^0) L(t) + O\left(\frac{1}{n}\right)\right]^n \\ &= \left[1 - \left[\frac{a^2 {h_1}^2 \sigma^2 t^2}{2n} F_X(d^0) + \frac{b^2 {h_2}^2 \sigma^2 t^2}{2n}\left(1 - F_X\left(d^0 + \frac{h_3}{n}\right)\right)\right.\right. \\ &\quad \left.\left. + \left(F_X\left(d^0 + \frac{h_3}{n}\right) - F_X(d^0)\right) - \frac{h_3}{n} f_X(d^0) L(t)\right] + O\left(\frac{1}{n}\right)\right]^n \\ &= \left[1 - \left[\frac{a^2 {h_1}^2 \sigma^2 t^2}{2n} F_X(d^0) + \frac{b^2 {h_2}^2 \sigma^2 t^2}{2n}\left(1 - F_X\left(d^0 + \frac{h_3}{n}\right)\right)\right.\right. \\ &\quad \left.\left. + \frac{h_3}{n} f_X(d^0) - \frac{h_3}{n} f_X(d^0) L^+(t)\right] + O\left(\frac{1}{n}\right)\right]^n \\ &= \left[1 - \left[\frac{a^2 {h_1}^2 \sigma^2 t^2}{2n} F_X(d^0) + \frac{b^2 {h_2}^2 \sigma^2 t^2}{2n}\left(1 - F_X\left(d^0 + \frac{h_3}{n}\right)\right)\right.\right. \\ &\quad \left.\left. + \frac{h_3}{n} f_X(d^0)(1 - L^+(t))\right] + O\left(\frac{1}{n}\right)\right]^n \\ &\longrightarrow e^{-\frac{a^2 {h_1}^2 \sigma^2 t^2}{2} F_X(d^0)} \cdot e^{-\frac{b^2 {h_2}^2 \sigma^2 t^2}{2}(1 - F_X(d^0))} \cdot e^{-h_3 f_X(d^0)(1 - L^+(ct))} \end{aligned}$$

Note that

$$\begin{aligned}
E[e^{it(a\mathbb{A}_n)}] &= \left[E[e^{it\left(\frac{ah_1}{\sqrt{n}}(y-\alpha_0)I(x\leq d^0)\right)}]\right]^n \\
&= \left[\int_0^{d^0} E[e^{it\left(\frac{ah_1}{\sqrt{n}}(y-\alpha_0)\right)}|X=x]p_X dx + \int_{d^0}^1 p_X dx\right]^n \\
&= \left[\left[1-\frac{1}{2}\left(\frac{ah_1 t}{\sqrt{n}}\right)^2\sigma^2 + O\left(\frac{t^2}{n}\right)\right]F_X(d^0) + 1 - F_X(d^0)\right]^n \\
&= \left[1-\frac{a^2{h_1}^2\sigma^2t^2}{2n}F_X(d^0) + O\left(\frac{1}{n}\right)\right]^n \\
&\longrightarrow \exp\{-\frac{a^2{h_1}^2\sigma^2t^2}{2}F_X(d^0)\} \\
E[e^{it(b\mathbb{B}_n)}] &= \left[E[e^{it\left(\frac{bh_2}{\sqrt{n}}(y-\beta_0)I\left(x>d^0+\frac{h_3}{n}\right)\right)}]\right]^n \\
&= \left[\int_{d^0+\frac{h_3}{n}}^1 E[e^{it\left(\frac{bh_2}{\sqrt{n}}(y-\beta_0)\right)}|X=x]p_X dx + \int_0^{d^0+\frac{h_3}{n}} p_X dx\right]^n \\
&= \left[\left(1-F_X\left(d^0+\frac{h_3}{n}\right)\right)\left(1-\frac{b^2{h_2}^2\sigma^2t^2}{2n}+O\left(\frac{t^2}{n}\right)\right)\right. \\
&\qquad \left. +F_X\left(d^0+\frac{h_3}{n}\right)\right]^n
\end{aligned}$$

$$\begin{aligned}
&= \left[1-\frac{b^2{h_2}^2\sigma^2t^2}{2n}\left(1-F_X\left(d^0+\frac{h_3}{n}\right)\right)+O\left(\frac{1}{n}\right)\right]^n \\
&\longrightarrow \exp\{-\frac{b^2{h_2}^2\sigma^2t^2}{2}(1-F_X(d^0))\}
\end{aligned}$$

and

$$E[e^{it(c\mathbb{C}_n)}] \longrightarrow e^{-h_3 f_X(d^0)(1-L^+(ct))}$$

as shown in the case that α_0 and β_0 are known. Now, the asymptotic independence follows from the fact that the joint characteristic function, in the limit, splits as the product of the limiting marginal characteristic functions.

Remark: These ideas can be readily extended to general (smooth) parametric models on either side of the change point and, in particular, higher order polynomials. We omit a discussion.

2.3 Confidence Intervals

We compare the performance of exact confidence intervals based on the asymptotic properties of the estimate from the classical procedure to those proposed in Ferger(2004).

For these comparisons, simulations were run for a stump model with $\alpha_0 = 0.5$, $\beta_0 = 1.5$, $d^0 = 0.5$ and sample sizes $n = 50, 100, 500, 1000$ with 2000 replicates for each n. Confidence intervals for d^0 based on the minimal minimizer $\hat{d}_{n,l}$, the maximal minimizer $\hat{d}_{n,u}$, and the average minimizer $\hat{d}_{n,av} = (\hat{d}_{n,l} + \hat{d}_{n,u})/2$ were constructed. The confidence level was set at $1-q = .95$ and the percentage of replicates for which the true change-point was included in the corresponding intervals, as well as the average length of each interval, were recorded.

In what follows, the symbols $d_l(|\beta_0 - \alpha_0|, \epsilon_1, f_X(d^0))$ and $d_u(|\beta_0 - \alpha_0|, \epsilon_1, f_X(d^0))$ have the same connotations as in Proposition 1.

2.3.1 Conservative Confidence Intervals.

Ferger (2004) proposed an asymptotic confidence interval for d at level $1-q$:

$$\hat{I}_n(q) := (\hat{d}_n - b/\theta_n, \hat{d}_n - a/\theta_n)$$

where $a < b$ are any solution of the inequality

$$\text{Prob}\,(d_u(|\beta_0 - \alpha_0|, \epsilon_1, f_X(d^0)) < b) - \text{Prob}\,(d_l(|\beta_0 - \alpha_0|, \epsilon_1, f_X(d^0)) \leq a) \geq 1 - q$$

Using the results of Ferger (2004), we construct asymptotically conservative confidence interval for d^0 at level $1-q$:

$$\hat{I}_{n,l}(\alpha) = (\hat{d}_{n,l} - b/n, \hat{d}_{n,l} - a/n),$$

$$\hat{I}_{n,u}(\alpha) = (\hat{d}_{n,u} - b/n, \hat{d}_{n,u} - a/n),$$

and

$$\hat{I}_{n,av}(\alpha) = (\hat{d}_{n,av} - b/n, \hat{d}_{n,av} - a/n)$$

	n=50	n=100	n=500	n=1000
$\hat{I}_{n,s}$	97.70%	97.70%	97.70%	97.10%
	(.1211)	(.0606)	(.0121)	(.0061)
$\hat{I}_{n,l}$	97.60%	97.70%	97.30%	97.10%
	(.1211)	(.0606)	(.0121)	(.0061)
$\hat{I}_{n,a}$	99.75%	99.50%	99.70%	99.60%
	(.1211)	(.0606)	(.0121)	(.0061)

Table 2.1: Conservative Confidence Intervals for different sample sizes, classical procedure

where a is the $q/2^{th}$ quantile of $d_l(|\beta_0 - \alpha_0|, \epsilon_1, f_X(d^0))$ and b is the $(1 - q/2)^{th}$ quantile of $d_u(|\beta_0 - \alpha_0|, \epsilon_1, f_X(d^0))$. These quantiles do not seem to be analytically determinable but can certainly be simulated to a reasonable degree of approximation.

In Table 2.1, the coverage probabilities together with the length of the confidence intervals are shown for a number of sample sizes. It can be seen that the recorded coverage exceeds the nominal level of 95% and almost approaching perfect (100%) coverage for the average minimizer.

2.3.2 Exact Confidence Intervals.

On the other hand, since Proposition 1 provides us with the asymptotic distributions of the sample minimizers, we can construct asymptotically exact (level $1 - q$ confidence intervals) as follows:

$$\tilde{I}_{n,l} = (\hat{d}_{n,l} - b_l/n, \hat{d}_{n,l} - a_l/n),$$

$$\tilde{I}_{n,u} = (\hat{d}_{n,u} - b_u/n, \hat{d}_{n,u} - a_u/n),$$

$$\tilde{I}_{n,av} = (\hat{d}_{n,av} - b_{av}/n, \hat{d}_{n,av} - a_{av}/n)$$

where $a_l, b_l, a_u, b_u, a_{av}$ and b_{av} are the exact quantiles (a_l, a_u and a_{av} correspond to $q/2^{th}$ quantiles and b_l, b_u and b_{av} correspond to $(1 - q/2)^{th}$ quantiles) of $d_l(|\beta_0 - \alpha_0|, \epsilon_1, f_X(d^0))$, $d_u(|\beta_0 - \alpha_0|, \epsilon_1, f_X(d^0))$ and $(d_l(|\beta_0 - \alpha_0|, \epsilon_1, f_X(d^0)) + d_u(|\beta_0 - \alpha_0|, \epsilon_1, f_X(d^0)))/2$, respectively.

In Table 2.2, the coverage probabilities together with the length of the confidence intervals are shown for a number of sample sizes. It can be seen that the coverage

	n=50	n=100	n=500	n=1000
$\tilde{I}_{n,s}$	94.85%	95.20%	94.95%	94.75%
	(.0607)	(.0303)	(.0061)	(.0030)
$\tilde{I}_{n,l}$	94.80%	95.20%	94.35%	94.10%
	(.0604)	(.0302)	(.0060)	(.0030)
$\tilde{I}_{n,a}$	95.55%	94.85%	95.20%	94.40%
	(.0616)	(.0308)	(.0062)	(.0031)

Table 2.2: Exact Confidence Intervals for different sample sizes, classical procedure

probabilities are fairly close to their nominal values. Further, their length is almost half of those obtained by Ferger's (2004) method.

CHAPTER 3

Multi-Stage Procedure for Change–Point Estimation

3.1 The Two-stage Procedure and the Asymptotic Properties of the Estimate.

We first describe a two–stage procedure for estimating the (unique) change–point. In what follows, we consider a regression scenario where the response Y_x generated at covariate level x can be written as $Y_x = \mu(x)+\epsilon$, where ϵ is a symmetric error variable with finite variance and μ is the regression function. The errors corresponding to different covariate levels are i.i.d. We first focus on the simple regression function $\mu(x) = \alpha_0\, 1(x \leq d^0) + \beta_0\, 1(x > d^0)$ and discuss generalizations to more complex parametric models later. We are allowed to sample n covariate–response pairs at most and are free to sample a response from any covariate level that we like.

- Step 1: At stage one, $\lambda\, n$ covariate values are sampled uniformly from $[0, 1]$ and responses are obtained. Denote the observed data by $\{X_i, Y_i\}_{i=1}^{n_1}$, $n_1 = \lambda n$ and the corresponding estimated location of the change point by $\hat{d}_{n_1}$.

- Step 2: Sample the remaining $n_2 = (1-\lambda)n$ covariate–response pairs $\{U_i, W_i\}_{i=1}^{n_2}$, where:

$$W_i = \mu(U_i) + \varepsilon_i, \quad U_i \sim \text{Unif } [\hat{a}_{n_1}, \hat{b}_{n_1}]$$

  and $[\hat{a}_{n_1}, \hat{b}_{n_1}] = [\hat{d}_{n_1} - Kn_1^{-\gamma}, \hat{d}_{n_1} + Kn_1^{-\gamma}]$, $0 < \gamma < 1$ and K is some constant. Obtain an updated estimate of the change point based on the n_2 covariate–response pairs from stage 2, which is denoted by $\hat{d}_{n_2}$.

We discuss the basic procedure in some more detail. Let $(\hat{\alpha}_{n_1}, \hat{\beta}_{n_1}, \hat{d}_{n_1})$ denote the parameter estimates obtained from stage one. Let $\mathbb{P}_{n_2}$ denote the empirical measure of the data points $\{U_i, W_i\}_{i=1}^{n_2}$. The updated estimates are computed by minimizing

$$\mathbb{P}_{n_2}[\{(w - \hat{\alpha}_{n_1})^2 \, I(u \leq d) + (w - \hat{\beta}_{n_1})^2 \, I(u > d)]$$

which, as is readily seen, is equivalent to minimizing the process

$$\tilde{\mathbb{M}}_{n_2}(d) \equiv \mathbb{P}_{n_2}[\{(w - \hat{\alpha}_{n_1})^2 - (w - \hat{\beta}_{n_1})^2\}(I(u \leq d) - I(u \leq d^0))] \, .$$

The process $\tilde{\mathbb{M}}_{n_2}$ is a piecewise constant right continuous function with left limits. We let $\hat{d}_{n_2,l}$ and $\hat{d}_{n_2,u}$ denote its minimal and maximal minimizers, respectively. Our goal is to determine the joint limit distribution of normalized versions of $(\hat{d}_{n_2,l}, \hat{d}_{n_2,u})$. This is described in the theorems that follow.

Theorem 3.1. *Assume that the error variable ϵ in the regression model has a finite moment generating function in a neighborhood of 0. Then, the random vector $n^{1+\gamma}(\hat{d}_{n_2,l} - d^0, \hat{d}_{n_2,u} - d^0)$ is $O_p(1)$.*

Remark: The proof of this theorem is fairly technical and particularly long and thus deferred to Section 3.1.3. However, a few words regarding the intuition behind the *accelerated rate* of convergence are in order. For simplicity, consider sampling procedures where instead of sampling from a uniform distribution on the interval of interest, sampling takes place on a uniform grid on the interval. The interval from which sampling takes place at the second stage has length $2K \, n_1^{-\gamma}$. Since the n_2 covariate values are equispaced over this interval, the resolution of the resulting grid at which responses are measured is $O(n_1^{-\gamma}/n_2) = O(n^{-(1+\gamma)})$ and this determines the rate of convergence of the two stage estimator (just as the rate of convergence in the classical procedure where n covariates are equispaced over $[0, 1]$ is given by the resolution of the resulting grid in that situation, which is simply (n^{-1})).

We next describe the limit distributions of the normalized estimates considered in Theorem 3.1.

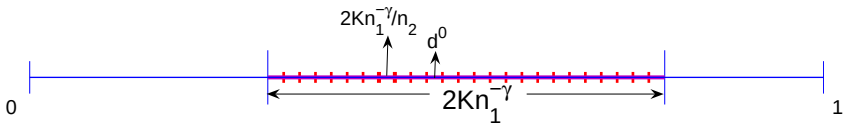


Figure 3.1: Rate of the convergence of the estimate from two-stage procedure

Theorem 3.2. *Set* $C(K,\lambda,\gamma) = (2K)^{-1}\,(\lambda/(1-\lambda))^{\gamma}$. *The random vector* $n_2^{1+\gamma}(\hat{d}_{n_2,l} - d^0, \hat{d}_{n_2,u} - d^0)$ *converges in distribution to*

$$(d_l(|\ \alpha_0 - \beta_0\ |, \epsilon, C(K,\lambda,\gamma)), d_u(|\ \alpha_0 - \beta_0\ |, \epsilon, C(K,\lambda,\gamma)))\,.$$

Remark: The asymptotic distributions of the 'zoom-in' estimators are given by the minimizers of a compound Poisson process. The underlying Poisson process is basically the limiting version of the count process $\{\mathcal{P}_n(s) :\ s \in \mathbb{R}\}$, where $\mathcal{P}_n(s)$ counts the number of U_i's in the interval $(d^0, d^0 + s/n_2^{1+\gamma}] \cup (d^0 + s/n_2^{1+\gamma}, d^0]$. It can be readily checked that marginally, $\mathcal{P}_n(s)$, converges in distribution to a Poisson random variable with mean $C(K,\lambda,\gamma)s$, using the Poisson approximation to the Binomial distribution. On the other hand, the size of the jumps of the compound Poisson process is basically determined by $|\alpha_0 - \beta_0|/\sigma$, the signal-to-noise ratio in the model.

We draw a QQ plot (as shown in Figure 3.2) to illustrate the quality of the approximation for the estimate. We use the average minimizer of the minimal minimizer and the maximal minimizer here. The horizontal axis corresponds to the quantiles of the average minimizer of the limit process and the vertical axis corresponds to the quantiles of the estimates from the simulated two-stage procedure with 1000 replicates. The estimates match the straight line perfectly.

From the simulation results, we compare the histograms (Figure 3.3) of the estimates from classical procedure and the estimates from a two-stage procedure. The

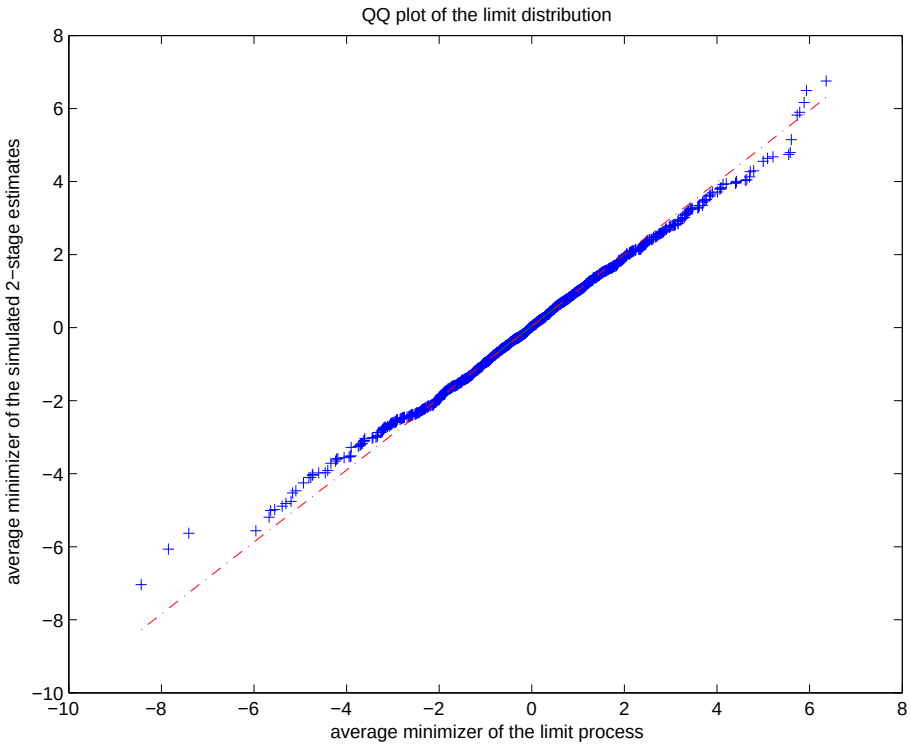
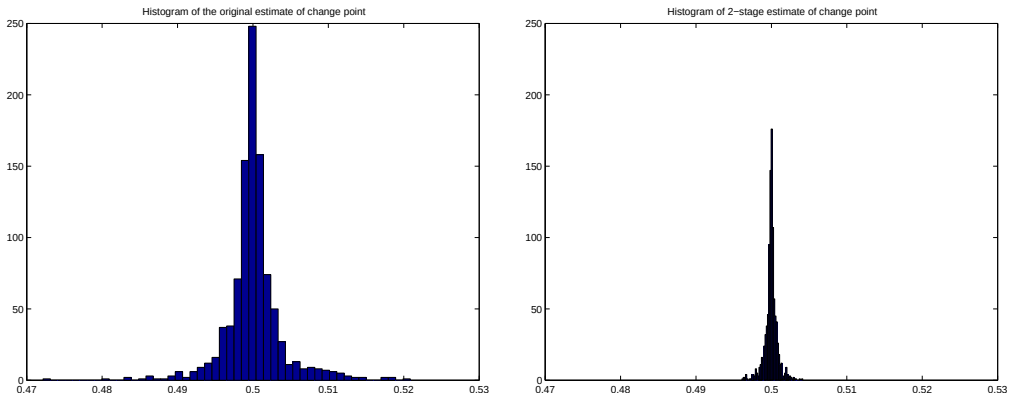


Figure 3.2: Quality of the Approximation, $\alpha_0 = .5$, $\beta_0 = 1.5$, $d^0 = .5$, $\sigma = .2$, n=1000, $K = 1$, $\lambda = .5$



Classical One-stage Procedure Estimates | Two-stage Procedure Estimates

Figure 3.3: Histograms of estimate of change-point

variation of our two-stage estimates is much smaller.

General parametric models: These results admit ready extensions to the case where the function $\mu(x)$ is as defined in (2.1). As in the case of a piecewise constant μ, $n_1 \equiv \lambda\, n$ points are initially used to obtain least squares estimates of (β_l, β_u, d^0), which we denote by $(\hat{\beta}_{l,n_1}, \hat{\beta}_{u,n_1}, \hat{d}_{n_1})$. Step 2 of the two–stage procedure is identical and the updated estimate $\hat{d}_{n_2}$ is computed by minimizing the criterion function

$$\mathbb{P}_{n_2}[\{(w - \psi_l(\hat{\beta}_{l,n_1}, u)^2\, I(u \leq d) + (w - \psi_u(\hat{\beta}_{u,n_1}, u)^2\, I(u > d)]$$

which is equivalent to minimizing

$$\tilde{\mathbb{M}}_{n_2}(d) = \mathbb{P}_{n_2}[\{(w - \psi_l(\hat{\beta}_{l,n_1}, u))^2 - (w - \psi_u(\hat{\beta}_{u,n_1}, u))^2\}(I(u \leq d) - I(u \leq d^0))]\,.$$

Letting $\hat{d}_{n_2,l}$ and $\hat{d}_{n_2,u}$ denote the smallest and largest argmins of $\tilde{\mathbb{M}}_{n_2}$ respectively (as in the piecewise constant function case), we have the following Proposition.

Proposition 3.1. *The random vector* $n_2^{1+\gamma}(\hat{d}_{n_2,l} - d^0, \hat{d}_{n_2,u} - d^0)$ *converges in distribution to*

$$(d_l(|\ \psi_l(\beta_l, d^0) - \psi_u(\beta_u, d^0)\ |, \epsilon, C(K, \lambda, \gamma)), d_u(|\ \psi_l(\beta_l, d^0) - \psi_u(\beta_u, d^0)\ |, \epsilon, C(K, \lambda, \gamma)))\,.$$

The heteroscedastic case: Similar results continue to hold for a heteroscedastic regression model. We formulate the heteroscedastic setting as follows. At any given covariate level x, the observed response $Y_x = \mu(x) + \sigma(x)\,\tilde{\epsilon}$ with $\mu(x)$ as defined in (2.1), $\sigma^2(x)$ is a (continuous) variance function and $\tilde{\epsilon}$ is a symmetric error variable with unit variance. The errors corresponding to different covariate values are independent. Using the same two stage procedure as described above, the following proposition obtains.

Proposition 3.2. *We have*

$$n_2^{1+\gamma}(\hat{d}_{n_2,l} - d^0, \hat{d}_{n_2,u} - d^0) \to_d \big(d_l(|\ \psi_l(\beta_l, d^0) - \psi_u(\beta_u, d^0)\ |, \sigma(d^0)\,\tilde{\epsilon}, C(K, \lambda, \gamma)),$$
$$d_u(|\ \psi_l(\beta_l, d^0) - \psi_u(\beta_u, d^0)\ |, \sigma(d^0)\,\tilde{\epsilon}, C(K, \lambda, \gamma))\big)\,.$$

Remark: With choice of a constant variance function, $\sigma^2(x) \equiv \sigma^2$, the heteroscedastic model reduces to the homoscedastic one. We, nevertheless, present results for these two situations separately. We also subsequently derive our results for the homoscedastic case, the derivations extending almost trivially to the heteroscedastic case.

3.1.1 Some Generalizations

We briefly discuss some generalizations of the two–stage procedure. The first of these considers more general neighborhoods of the initial estimate of the change–point and the second generalizes the two–stage procedure to multiple stages.

More general neighborhoods: Instead of considering a polynomially decaying neighborhood of the initial estimate in Step 2, one could consider a more general neighborhood of the form $[\hat{d}_{n_1} - k_n\, n^{-1}, \hat{d}_{n_1} + k_n\, n^{-1}]$ where $k_n = o(n^{-1})$ and $k_n \to \infty$. In this case $(\hat{d}_{n_2,1} - d^0, \hat{d}_{n_2,u} - d^0)$ is $O_p(k_n/n^2)$, so that, in theory, rates logarithmically close to n^2 can be achieved (set $k_n = \log n$). The procedure discussed at the beginning of Section 3 is a special case of this scenario with $k_n = K\,\lambda^{-\gamma}\, n^{1-\gamma}$. For the heteroscedastic model described above,

$$\frac{n^2}{k_n}\left(\hat{d}_{n_2,1} - d^0, \hat{d}_{n_2,u} - d^0\right) \to_d \left(d_l(|\ \psi_l(\beta_l, d^0) - \psi_u(\beta_u, d^0)\ |, \sigma(d^0)\,\tilde{\epsilon}, (1-\lambda)/2),\right.$$
$$\left. d_u(|\ \psi_l(\beta_l, d^0) - \psi_u(\beta_u, d^0)\ |, \sigma(d^0)\,\tilde{\epsilon}, (1-\lambda)/2)\right)\ .$$

Multi–stage procedures: Consider a generalization of the two stage procedure to k stages in the setting of the heteroscedastic model with a general parametric regression function μ. Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be the proportions of points used at each stage (where $\lambda_1 + \lambda_2 + \ldots + \lambda_k = 1$) and let $n_i = \lambda_i\, n$. Also, fix sequences of numbers $0 < \gamma_{(k-1)} < \ldots < \gamma_{(1)} < 1$ and $K_1, K_2, \ldots, K_{k-1}$ (with $K_i > 0$). Having used n_1 points to construct the initial estimate $\hat{d}_{n_1}$, in the qth $(2 \le q \le k)$ stage, define the sampling neighborhood as $[\hat{d}_{n_{q-1}} - K_{q-1} n_{q-1}^{-((q-2)+\gamma_{q-1})}, \hat{d}_{n_{q-1}} + K_{q-1} n_{q-1}^{-((q-2)+\gamma_{q-1})}]$, sample n_q covariate–response pairs $\{w_i, u_i\}_{i=1}^{n_q}$ from this neighborhood: $W_i = \mu(U_i) + \epsilon_i$ and

update the estimate of the change–point to $\hat{d}_{n_q}$. Let $(\hat{d}_{n_k,l}, \hat{d}_{n_k,u})$ denote the smallest and largest estimates at stage k. It can be shown that $n_k^{(k-1)+\gamma_{(k-1)}}((\hat{d}_{n_k,l}-d^0), (\hat{d}_{n_k,u}-d^0))$ is $O_p(1)$ and converges in distribution to (d_l, d_u), where (d_l, d_u) is the vector of the smallest and the largest argmins of the process $\mathbb{M}(|\ \psi_l(\beta_l, d^0) - \psi_u(\beta_u, d^0)\ |\ , \sigma(d^0)\,\tilde{\epsilon}, C_k)$, with $C_k = (1/2K_{k-1})(\lambda_{k-1}/\lambda_k)^{((k-2)+\gamma_{(k-1)})}$.

3.1.2 Proof of Theorem 3.2

For the proof of this theorem (and the proof of Lemma 3.2 in Section 3.1.3) we denote the process $\mathbb{M}_{|\alpha_0-\beta_0|,\epsilon,C(K,\lambda,\gamma)}$ simply by $\mathbb{M}$ and its smallest and largest minimizers simply by (d_l, d_u). Our proof of this theorem will rely on continuous mapping for the argmin functional. For the sake of concreteness, in what follows, we assume that $\alpha_0 < \beta_0$. Under this assumption, with probability increasing to 1 as n (and consequently n_1) goes to infinity, $\hat{\alpha}_{n_1} < \hat{\beta}_{n_1}$ and d^0 belongs to the set $[\hat{d}_{n_1} - K\, n_1^{-\gamma}, \hat{d}_{n_1} + K\, n_1^{-\gamma}]$. On this set $(\hat{d}_{n_2,l}, \hat{d}_{n_2,u})$ can be obtained by minimizing (the equivalent) criterion function:

$$\mathbb{P}_{n_2}\left[\left(w - \frac{\hat{\alpha}_{n_1} + \hat{\beta}_{n_1}}{2}\right)(I(u \le d) - I(u \le d^0))\right]$$

and d^0 is characterized as:

$$d^0 \;=\; \text{argmin}\ \mathbf{P}\left[\left(w - \frac{\hat{\alpha}_{n_1} + \hat{\beta}_{n_1}}{2}\right)(I(u \le d) - I(u \le d^0))\right]$$

where $\mathbf{P}$ is the distribution of (W, U). Therefore, in what follows, we take:

$$\tilde{\mathbb{M}}_{n_2}(d) = \mathbb{P}_{n_2}\left[\left(w - \frac{\hat{\alpha}_{n_1} + \hat{\beta}_{n_1}}{2}\right)(I(u \le d) - I(u \le d^0))\right],$$

and $\hat{d}_{n_2,l}$ and $\hat{d}_{n_2,u}$ to be the smallest and largest argmins of this stochastic process. Set $(\xi_{n,l}, \xi_{n,u}) = n_2^{1+\gamma}(\hat{d}_{n_2,l} - d^0, \hat{d}_{n_2,u} - d^0)$. Then $(\xi_{n,l}, \xi_{n,u})$ is the vector of smallest and largest argmins of the stochastic process:

$$\begin{aligned}\mathbb{M}_{n_2}(s) &= \sum_{i=1}^{n_2}\left[\left(w_i - \frac{\hat{\alpha}_{n_1} + \hat{\beta}_{n_1}}{2}\right)\left(I\left(u_i \le d^0 + \frac{s}{n_2^{1+\gamma}}\right) - I\left(u_i \le d^0\right)\right)\right] \\ &= \mathbb{M}_{n_2}^{+}(s) - \mathbb{M}_{n_2}^{-}(s)\end{aligned}$$

where

$$\mathbb{M}_{n_2}^{+}(s) = \mathbb{M}_{n_2}(s)\,1(s \geq 0) \;\text{ and }\; \mathbb{M}_{n_2}^{-}(s) = -\mathbb{M}_{n_2}(s)\,1(s \leq 0)\,.$$

We now introduce some notation that is crucial to the subsequent development. Let $\mathcal{S}$ denote the class of piecewise constant right continuous functions with left limits (from $\mathbb{R}$ to $\mathbb{R}$) that are continuous at every integer point, assume the value 0 at 0 and possess finitely many jumps in every compact interval $[-C, C]$ where $C > 0$ is an integer. Let $\tilde{f}$ denote the pure jump process (of jump size 1) corresponding to the function f; i.e. $\tilde{f}$ is the piecewise constant right continuous function with left limits, such that for any $s > 0$, $\tilde{f}(s)$ counts the number of jumps of the function f in the interval $[0, s]$, while for $s < 0$, $\tilde{f}(s)$ counts the number of jumps in the set $(s, 0)$.

For any positive integer $C > 0$, let $\mathcal{D}[-C, C]$ denote the class of all right continuous functions with left limits with domain $[-C, C]$ equipped with the Skorokhod topology and let $\mathcal{D}[-C, C] \times \mathcal{D}[-C, C]$ denote the corresponding product space. Finally, let $\mathcal{D}_C^0$ denote the (metric) subspace of $\mathcal{D}([-C, C]) \times \mathcal{D}([-C, C])$ that comprises all function pairs of the form $(f\,|_{[-C,C]}, \tilde{f}\,|_{[-C,C]})$ for $f \in \mathcal{S}$. We have the following Lemma which is proved in Section 3.1.3.

Lemma 3.1. *Let $\{f_n\}$ and f_0 be functions in $\mathcal{S}$, such that for every positive integer C, $(f_n\,|_{[-C,C]}, \tilde{f}_n\,|_{[-C,C]})$ converges to $(f_0\,|_{[-C,C]}, \tilde{f}_0\,|_{[-C,C]})$ in $\mathcal{D}_0^C$ where f_0 satisfies the property that no two flat stretches of f_0 have the same height. Let $l_{n,C}$ and $u_{n,C}$ denote the smallest and the largest minimizers of f_n on $[-C, C]$, and $l_{0,C}$ and $u_{0,C}$ denote the corresponding functionals for f_0. Then $(l_{n,C}, u_{n,C}) \to (l_{0,C}, u_{0,C})$.*

Consider the sequence of stochastic processes $\mathbb{M}_{n_2}(s)$ and let $\mathbb{J}_{n_2}(s)$ denote the corresponding jump processes. We have:

$$\begin{aligned}
\mathbb{J}_{n_2}(s) &= \text{sign}(s) \sum_{i=1}^{n_2} \left[\left(I\left(U_i \leq d^0 + \frac{s}{n_2^{1+\gamma}} \right) - I\left(U_i \leq d^0 \right) \right) \right] \\
&= \mathbb{J}_{n_2}^{+}(s) + \mathbb{J}_{n_2}^{-}(s)
\end{aligned}$$

where

$$\mathbb{J}_{n_2}^+(s) = \mathbb{J}_{n_2}(s)\,1(s \geq 0) \;\text{ and }\; \mathbb{J}_{n_2}^-(s) = \mathbb{J}_{n_2}(s)\,1(s \leq 0)\,.$$

The jump process corresponding to $\mathbb{M}(s)$ is denoted by $\mathbb{J}(s)$ and is given by $\nu^+(s)1(h \geq 0)+\nu^-(s)1(h \leq 0)$. For each n,$\{\mathbb{M}_{n_2}(s) : s \in \mathbb{R}\}$ lives in $\mathcal{S}$ with probability one. Also, with probability 1, $\{\mathbb{M}(s) : s \in \mathbb{R}\}$ lives in $\mathcal{S}$. Also, on a set of probability one (which does not depend on C), for every positive integer C, $((\mathbb{M}_{n_2}(s), \mathbb{J}_{n_2}(s)) : s \in [-C, C])$ belongs to $\mathcal{D}_0^C$ and so does $((\mathbb{M}(s), \mathbb{J}(s)) : s \in [-C, C])$. Let $(\xi_{n,C,l}, \xi_{n,C,u})$ denote the smallest and largest argmin of $\mathbb{M}_{n_2}$ restricted to $[-C, C]$ and let $(d_{C,l}, d_{C,u})$ denote the corresponding functionals for $\mathbb{M}$ restricted to $[-C, C]$. We prove in the Appendix:

Lemma 3.2. *For every $C > 0$, $((\mathbb{M}_{n_2}(s), \mathbb{J}_{n_2}(s)) : s \in [-C, C])$ converges in distribution to $((\mathbb{M}(s), \mathbb{J}(s)) : s \in [-C, C])$ in the space $\mathcal{D}_C^0$.*

Consider the function h that maps an element (a pair of functions) of $\mathcal{D}_0^C$ to the two dimensional vector given by the smallest argmin and the largest argmin of the first component of the element. Using the fact that almost surely no two flat stretches of $\mathbb{M}$ have the same height, it follows by Lemma 3.1 that the process $((\mathbb{M}(s), \mathbb{J}(s)) : s \in [-C, C])$ belongs, almost surely, to the continuity set of the function h. This, coupled with the distributional convergence established in Lemma 3.2 leads to the conclusion that

$$(\hat{\xi}_{n,C,l}, \hat{\xi}_{n,C,u}) \to_d (d_{C,l}, d_{C,u})\,. \tag{3.1}$$

We will show that $(\xi_{n,l}, \xi_{n,u}) \to (d_l, d_u)$. To this end, we use the following lemma from Prakasa Rao (1969).

Lemma 3.3. *Suppose that $\{W_{n\epsilon}\}, \{W_n\}$ and $\{W_\epsilon\}$ are three sets of random vectors such that*

(i) $\lim_{\epsilon \to 0} \limsup_{n \to \infty} P[W_{n\epsilon} \neq W_n] = 0\,,$ *(ii)* $\lim_{\epsilon \to 0} P[W_\epsilon \neq W] = 0$ *and (iii) For every $\epsilon > 0$, $W_{n\epsilon} \to_d W_\epsilon$ as $n \to \infty$.*

Then $W_n \to_d W$, as $n \to \infty$.

Before applying the lemma, we first note the following facts: (a) The sequence of (smallest and largest minimizers) $(\xi_{n,l}, \xi_{n,u})$ is $O_p(1)$, and (b) The minimizers (d_l, d_u) are $O_p(1)$. Now, in the above lemma, set $\epsilon = 1/C$, $W_{n\epsilon} = (\xi_{n,C,l}, \xi_{n,C,u})$, $W_\epsilon = (d_{C,l}, d_{C,u})$, $W_n = (\xi_{n,l}, \xi_{n,u})$ and $W = (d_l, d_u)$. Condition (iii) is established in (3.1). From (a) and (b) it follows that Conditions (i) and (ii) of the lemma are satisfied. We conclude that $(\xi_{n,l}, \xi_{n,u}) \rightarrow_d (d_l, d_u)$. □

Remark: It is instructive to compare the obtained result on the convergence of the (non–unique) argmin functional to that considered in Ferger (2004). Ferger deals with the convergence of the argmax functional under the Skorokhod topology in Theorems 2 and 3 of his paper. Since the argmax functional is *not continuous* under the Skorokhod topology, an exact result on distributional convergence cannot be achieved. Instead, asymptotic upper and lower bounds are obtained on the distribution function of the argmax in terms of the smallest maximizer and the largest maximizer of the limit process (page 88 of Ferger (2004)). The result we obtain here is, admittedly, in a more specialized set–up than the one considered in his paper, but it is stronger since we are able to show *exact* distributional convergence of argmins. This is achieved at the cost of some extra effort: establishing the joint convergence of the original processes, whose argmins are of interest, and their jump processes, and subsequently invoking continuous mapping. Under this stronger mode of convergence, the argmin functional indeed turns out to be continuous, as Lemma 3.1 shows (the arguments employed are similar in spirit to those in section 14.5.1 of Kosorok (2006)). This result allows us to construct asymptotic confidence intervals that have *exact* coverage at any given level, as opposed to the conservative intervals proposed in Ferger (2004). That the exact confidence intervals buy us significant precision over the conservative ones is evident from the reported simulation results discussed in Section 3.3.

3.1.3 Proof of Theorem 3.1

Proof of Lemma 3.1: Let $f_n(t) = \sum_{i=1}^{\infty} v_{n,i}\, 1(t \in [a_{n,i}, a_{n,i+1})) + \sum_{i=1}^{\infty} v_{n,-i}\, 1(t \in [a_{n,-(i+1)}, a_{n,-i}))$ where $0 < a_{n,1} < a_{n,2} < \ldots$ and $0 > a_{n,-1} > a_{n,-2} > \ldots$, and let $f_0(t) = \sum_{i=1}^{\infty} v_i\, 1(t \in [a_i, a_{i+1})) + \sum_{i=1}^{\infty} v_{-i}\, 1(t \in [a_{-(i+1)}, a_{-i}))$. On $[-C, C]$, the limit function f_0 has finitely many jump points, say, m_r to the right of 0 and m_l to the left of 0. The function $\tilde{f}_n$ assumes the value 0 on $[a_{n,-1}, a_{n,1})$ and the value i on $[a_{n,i}, a_{n,i+1})$ and $[a_{n,-(i+1)}, a_{n,-i})$, for all $i \geq 1$. The function $\tilde{f}_0$ assumes the value 0 on $[a_{-1}, a_1)$ and the value i on $[a_i, a_{i+1})$ and $[a_{-(i+1)}, a_{-i})$, for all $i \geq 1$.

For any $\epsilon > 0$, consider, for $1 \leq i \leq m_r$, the points $a_i - \epsilon, a_i + \epsilon$. Since these are continuity points of f_0, $\tilde{f}_n$ must converge to $\tilde{f}_0$ at these finitely many points. Since $\tilde{f}_n$ and $\tilde{f}_0$ only assume integer values, for all sufficiently large n, $\tilde{f}_n(a_i - \epsilon) = \tilde{f}_0(a_i - \epsilon) = i - 1$ and $\tilde{f}_n(a_i + \epsilon) = \tilde{f}_0(a_i + \epsilon) = i$ for all $1 \leq i \leq m_r$ and $\tilde{f}_0(C) = \tilde{f}_n(C) = m_r$. It follows that the function $\tilde{f}_n$ has exactly m_r jump discontinuities on $[0, C]$ for all sufficiently large n; furthermore, since $\tilde{f}_n$ jumps between $a_i - \epsilon$ and $a_i + \epsilon$ for all i, $a_{n,i}$, the i'th largest jump location to the right of 0 satisfies $\mid a_{n,i} - a_i \mid \leq \epsilon$ for $1 \leq i \leq m_r$ for n large enough. A similar phenomenon happens to the left of 0, with the number of jumps of $\tilde{f}_n$ in $[-C, 0]$ being exactly m_l for all sufficiently large n, and the jump locations $\{a_{n,-i}\}_{i=1}^{m_l}$ converging to the jump locations $\{a_{-i}\}_{i=1}^{m_l}$ of $\tilde{f}_0$ on $[-C, 0]$.

Let $[a_j, a_{(j+1)})$ (with $j > 0$) denote the unique stretch on which the restriction of $\tilde{f}_0$ to $[-C, C]$ is minimized (that the minimizing stretch is unique is guaranteed by our assumptions). The value on this stretch is v_j. Now, consider the points $\{(a_m + a_{m+1})/2 : -m_l \leq m \leq m_r\} \cup \{(-C + a_{-m_l})/2, (a_{m_r} + C)/2\}$. These are continuity points of f_0 (and $\tilde{f}_0$) and by what has been shown in the previous paragraph, for all sufficiently large n, these are continuity points of $\tilde{f}_n$ with $(a_m + a_{m+1})/2$ lying in the stretch with $a_{n,m}$ and $a_{n,m+1}$ as extremities. Since f_n converges to f_0 in the Skorokhod metric on $[-C, C]$ it follows that $f_n((a_m + a_{m+1})/2) \to f_0((a_m + a_{m+1})/2)$,

for all m. Also $f_n((-C+a_{-m_l})/2$ and $f_n((a_{m_r}+C)/2)$ converge to $f_0((-C+a_{-m_l})/2)$ and $f_0((a_{m_r}+C)/2)$ respectively. Since v_j is the smallest value of f on $[-C,C]$ and is separated from the remaining possible levels of f_0 on $[-C,C]$ and $v_{n,j}$, the constant value of f_n on the stretch $[a_{n,j}, a_{n,(j+1)})$, converges to v_j, while the constant value of f_n on any other stretch converges to the constant value of f_0 on the corresponding stretch for the limit function, it follows that for all sufficiently large n, v_j^n is separated from the other possible values of f_n and is the smallest value of f_n on $[-C,C]$. It follows that $l_{n,C} = a_{n,j}$ and $u_{n,C} = a_{n,j+1}$ and these converge to a_j and a_{j+1} which are simply $l_{0,C}$ and $l_{u,C}$. A similar argument works if the stretch on which the minimum of f_0 on $[-C,C]$ is attained lies to the left of 0. □

Proof of Lemma 3.2: We first note that $\mathcal{D}_C^0$ (which we view as a metric subspace of $D[-C,C]\times D[-C,C]$) is a measurable subset of $D[-C,C]\times D[-C,C]$. To establish convergence in distribution in the space $\mathcal{D}_C^0$, it therefore suffices to establish convergence in distribution in the larger space $D[-C,C]\times D[-C,C]$ (see the discussion in Example 3.1 of Billingsley (1999)). This can be achieved by (a) Establishing finite dimensional convergence: showing that $\{\mathbb{M}_{n_2}(h_i), \mathbb{J}_{n_2}(h_i)\}_{i=1}^l \to \{\mathbb{M}(h_i), \mathbb{J}(h_i)\}_{i=1}^l$ for all $h_1, h_2, \ldots, h_l$ in $[-C,C]$. (b) Verifying tightness of the processes $(\mathbb{M}_{n_2}(h), \mathbb{J}_{n_2}(h))$ under the product topology. But this boils down to verifying marginal tightness.

Let

$$L^+(t) = E[e^{it\left(w-\frac{\alpha_0+\beta_0}{2}\right)} \mid U = d^{0+}] \equiv \lim_{d\to d^0+} E[e^{it\left(w-\frac{\alpha_0+\beta_0}{2}\right)} \mid U = d]$$

and $L^-(t) = E[e^{it\left(w-\frac{\alpha_0+\beta_0}{2}\right)} \mid U = d^0]$. It is not difficult to see that L^+ is the characteristic function of the V_i^+'s while L^- is the characteristic function of the V_i^-'s. In order to establish finite–dimensional convergence, we first show that for a fixed s, $\mathbb{M}_{n_2}(s)$ converges in distribution to $\mathbb{M}(s)$. We do this via characteristic

functions. Consider $\phi_s(t)$, the characteristic function of $\mathbb{M}(s)$ (with $s > 0$). We have:

$$\begin{aligned} E[e^{it\mathbb{M}_1(s)}] &= E[e^{it\sum_{0\le k\le \nu^+(s)} V_k^+}] \\ &= \sum_{l=0}^{\infty} E[e^{it(V_1^+ \ldots + V_l^+)}] \frac{e^{-\frac{s}{2K}(\frac{\lambda}{1-\lambda})^\gamma} \left(\frac{s}{2K}(\frac{\lambda}{1-\lambda})^\gamma\right)^l}{l!} \\ &= \sum_{l=0}^{\infty} \frac{(L^+(t))^l \left(\frac{s}{2K}(\frac{\lambda}{1-\lambda})^\gamma\right)^l}{l!} e^{-\frac{s}{2K}(\frac{\lambda}{1-\lambda})^\gamma} \\ &= e^{-\frac{s}{2K}(\frac{\lambda}{1-\lambda})^\gamma} e^{L(t)\frac{s}{2K}(\frac{\lambda}{1-\lambda})^\gamma} \\ &= e^{-\frac{s}{2K}(\frac{\lambda}{1-\lambda})^\gamma (1-L(t))} \end{aligned}$$

We show that $Q_{n_2,s}(t) \equiv E[e^{it\,\mathbb{M}_{n_2}(s)}]$ converges to $\phi_s(t)$. Let $\xi_{n_1} = n_1(\hat{d}_{n_1} - d^0), \eta_{n_1,1} = \sqrt{n_1}(\hat{\alpha}_{n_1} - \alpha_0), \eta_{n_1,2} = \sqrt{n_1}(\hat{\beta}_{n_1} - \beta_0)$. We have:

$$Q_{n_2,s}(t) = \int Q^\star_{n_2,s}(t, \eta_1, \eta_2, \xi)\, dZ_{n_1}(\eta_1, \eta_2, \xi)\,,$$

where Z_{n_1} is the joint distribution of $(\eta_{n_1,1}, \eta_{n_1,2}, \xi_{n_1})$ and

$$\begin{aligned} & Q^\star_{n_2,s}(t, \eta_1, \eta_2, \xi) = E[e^{it\mathbb{M}^+_{n_2}(s)} | \eta_{n_1,1} = \eta_1, \eta_{n_1,2} = \eta_2, \xi_{n_1} = \xi] \\ = \; & E\left[e^{it\left(W_1 - \frac{\hat{\alpha}_{n_1} + \hat{\beta}_{n_1}}{2}\right)\left(I\left(U_1 \le d^0 + \frac{s}{n_2^{1+\gamma}}\right) - I(U_1 \le d^0)\right)} \Big| \eta_{n_1,1} = \eta_1, \eta_{n_1,2} = \eta_2, \xi_{n_1} = \xi \right]^{n_2}. \end{aligned}$$

Let $\epsilon > 0$ be pre–assigned. By Proposition 1, we can find $L > 0$ such that for all sufficiently large n, $Z_{n_1}([-L, L]^3) \ge 1 - \epsilon/3$. Using the fact that characteristic functions are bounded by 1, it follows immediately that for all $n \ge N_0$ (for some N_0),

$$\begin{aligned} \mid Q_{n_2,s}(t) - \phi_s(t) \mid &\le \int_{[-L,L]^3} \mid Q^\star_{n_2,s}(t, \eta_1, \eta_2, \xi) - \phi_s(t) \mid d\, Z_{n_1}(\eta_1, \eta_2, \xi) + 2\,\epsilon/3 \\ &\le \sup_{(\eta_1,\eta_2,\xi)\in[-L,L]^3} \mid Q^\star_{n_2,s}(t, \eta_1, \eta_2, \xi) - \phi_s(t) \mid + 2\,\epsilon/3\,. \end{aligned}$$

For this fixed L, we now show that for all sufficiently large n

$$D_n \equiv \sup_{(\eta_1,\eta_2,\xi)\in[-L,L]^3} \mid Q^\star_{n_2,s}(t, \eta_1, \eta_2, \xi) - \phi_s(t) \mid \le \epsilon/3\,,$$

whence it follows that eventually $\mid Q_{n_2,s}(t) - \phi_s(t) \mid \le \epsilon$. To show the uniform convergence of $Q^\star_{n_2,s}(t, \eta_1, \eta_2, \xi)$ to $\phi_s(t)$ over the compact rectangle $[-L, L]^3$ we proceed as follows.

For given L and C, it is the case that for all sufficiently large n, for any $\xi \in [-L, L]$ and any $0 < s < C$,

$$d^0 + \xi/n_1 - K\, n_1^{-\gamma} < d^0 < d^0 + s/n_2^{1+\gamma} < d^0 + \xi/n_1 + K\, n_1^{-\gamma}\,.$$

Let $P_{n_2,d}(s) \equiv \Pr(d^0 \le U_1 \le d^0 + s/n_2^{1+\gamma} \mid \hat{d}_{n_1} = d)$. Consider the conditional characteristic function $Q^{\star}_{n_2,s}(t, \eta_1, \eta_2, \xi)$, for $(\eta_1, \eta_2, \xi) \in [-L, L]^3$. It follows from the above display that for all sufficiently large n (depending only on L and C),

$$\begin{aligned}
& Q^{\star}_{n_2}(t, \eta_1, \eta_2, \xi) \\
= \;& \left[[1 - P_{n_2, d^0 + \frac{\xi}{n_1}}(s)] + \int_{d^0}^{d^0 + \frac{s}{n_2^{1+\gamma}}} E\left[e^{it\left(W_1 - \frac{\hat{\alpha}_{n_1} + \hat{\beta}_{n_1}}{2}\right)} \Big| U_1 = u \right] p_{U_1}(u) du \right]^{n_2} \\
& \left(\text{where} \quad \hat{\alpha}_{n_1} = \alpha_0 + \frac{\eta_1}{\sqrt{n_1}}, \quad \hat{\beta}_{n_1} = \beta_0 + \frac{\eta_2}{\sqrt{n_1}} \right) \\
= \;& \left[[1 - P_{n_2, d^0 + \frac{\xi}{n_1}}(s)] + \frac{n_1^{\gamma}}{2K} \int_{d^0}^{d^0 + \frac{s}{n_2^{1+\gamma}}} E\left[e^{it\left(W_1 - \frac{\hat{\alpha}_{n_1} + \hat{\beta}_{n_1}}{2}\right)} \Big| U_1 = u \right] du \right]^{n_2} \\
= \;& \left[1 - \frac{1}{n_2} \frac{s}{2K} \left(\frac{\lambda}{1-\lambda} \right)^{\gamma} + \frac{n_1^{\gamma}}{2K} \int_{d^0}^{d^0 + \frac{s}{n_2^{1+\gamma}}} E\left[e^{it\left(W_1 - \frac{\hat{\alpha}_{n_1} + \hat{\beta}_{n_1}}{2}\right)} \Big| U_1 = u \right] du \right]^{n_2} \\
= \;& \left[1 - \frac{1}{n_2} \frac{s}{2K} \left(\frac{\lambda}{1-\lambda} \right)^{\gamma} + \frac{n_1^{\gamma}}{2K} \right. \\
& \left. \times \int_0^s E\left[\exp\left\{ it \left(W_1 - \frac{\alpha_0 + \frac{\eta_1}{\sqrt{n_1}} + \beta_0 + \frac{\eta_2}{\sqrt{n_1}}}{2} \right) \right\} \Big| U_1 = d^0 + \frac{v}{n_2^{1+\gamma}} \right] \frac{1}{n_2^{1+\gamma}} dv \right]^{n_2} \\
= \;& \left[1 - \frac{1}{n_2} \frac{s}{2K} (\frac{\lambda}{1-\lambda})^{\gamma} \right. \\
& \left. + \frac{1}{2Kn_2} \left(\frac{\lambda}{1-\lambda} \right)^{\gamma} e^{-\frac{it(\eta_1+\eta_2)}{2\sqrt{n_1}}} \int_0^s E\left[e^{it\left(W_1 - \frac{\alpha_0+\beta_0}{2}\right)} \Big| U = d^0 + \frac{v}{n_2^{1+\gamma}} \right] dv \right]^{n_2} \\
= \;& \left[1 - \frac{1}{n_2} \frac{s}{2K} \left(\frac{\lambda}{1-\lambda} \right)^{\gamma} (1 - B_{n_1,n_2,\eta_1,\eta_2}(s)) \right]^{n_2}
\end{aligned}$$

where

$$B_{n_1,n_2,\eta_1,\eta_2}(s) = \frac{1}{s} e^{-\frac{it(\eta_1+\eta_2)}{2\sqrt{n_1}}} \int_0^s E\left[e^{it\left(W_1 - \frac{\alpha_0+\beta_0}{2}\right)} \Big| U_1 = d^0 + \frac{v}{n_2^{1+\gamma}} \right] dv$$

and

$$D_n = \sup_{(\eta_1,\eta_2) \in [-L,L]^2} \left| \left[1 - \frac{1}{n_2} \frac{s}{2K} \left(\frac{\lambda}{1-\lambda} \right)^{\gamma} (1 - B_{n_1,n_2,\eta_1,\eta_2}(s)) \right]^{n_2} - \phi_s(t) \right| .$$

Let:

$$z_n(\eta_1,\eta_2) = -\frac{s}{2K}(\frac{\lambda}{1-\lambda})^\gamma\,(1 - B_{n_1,n_2,\eta_1,\eta_2}(s))\,.$$

It is easy to see that:

$$\tilde{D}_n \equiv \sup_{(\eta_1,\eta_2)\in[-L,L]^2}\ \mid z_n(\eta_1,\eta_2) - z_0\ \mid \to 0\,,$$

where $z_0 = -(s/2K)(\lambda/(1-\lambda))^\gamma(1-L^+(t))$. Consider now:

$$D_n = \sup_{(\eta_1,\eta_2)\in[-L,L]^2}\left|\left(1+\frac{1}{n_2}\,z_n(\eta_1,\eta_2)\right)^{n_2} - e^{z_0}\right|\,.$$

This is dominated by $I_n + II_n$ where:

$$I_n = \left|\left(1+\frac{1}{n_2}\,z_0\right)^{n_2} - e^{z_0}\right| \to 0\,,$$

and

$$II_n = \sup_{(\eta_1,\eta_2)\in[-L,L]^2}\ \left|\left(1+\frac{1}{n_2}\,z_n(\eta_1,\eta_2)\right)^{n_2} - \left(1+\frac{1}{n_2}\,z_0\right)^{n_2}\right|\,.$$

Since $\tilde{D}_n$ goes to 0, for all sufficiently large n, $\mid z_0 \mid \vee(\sup_{\eta_1,\eta_2)\in[-L,L]^2}\mid z_n(\eta_1,\eta_2)\mid)$ is bounded by a constant, say M. Straightforward algebra shows that for all sufficiently large n,

$$\begin{aligned} II_n &\leq (\sup_{(\eta_1,\eta_2)\in[-L,L]^2}\ \mid z_n(\eta_1,\eta_2) - z_0\ \mid)\left(\sum_{j=1}^{n_2}\binom{n_2}{j}\frac{j\,M^{j-1}}{n_2^j}\right) \\ &= (\sup_{(\eta_1,\eta_2)\in[-L,L]^2}\ \mid z_n(\eta_1,\eta_2) - z_0\ \mid)\left(1+\frac{M}{n_2}\right)^{n_2-1} \to 0\,. \end{aligned}$$

Thus $D_n \to 0$ and the uniform convergence of $Q^*_{n_2}(t,\eta_1,\eta_2,\xi)$ to $\phi_s(t) = e^{z_0}$ on $[-L,L]^3$ is established.

We now establish the weak convergence of the finite dimensional distributions of $(\mathbb{M}_{n_2},\mathbb{J}_{n_2})$ to those of $(\mathbb{M},\mathbb{J})$. For convenience, we restrict ourselves only to the set $[0,C]$. Let J be a positive integer and consider $0 = s_0 < s_1 < s_2 < \ldots < s_J \leq C$. Let $c_1, c_2, \ldots, c_J$ and $d_1, d_2, \ldots, d_J$ be constants. We show that:

$$\begin{aligned} A_n &\equiv E\left(e^{it\sum_{j\leq J}(c_j(\mathbb{M}^+_{n_2}(s_j)-\mathbb{M}^+_{n_2}(s_{j-1}))+d_j(\mathbb{J}^+_{n_2}(s_j)-\mathbb{J}^+_{n_2}(s_{j-1})))}\right) \\ &\to A \equiv E\left(e^{it\sum_{j\leq J}(c_j(\mathbb{M}(s_j)-\mathbb{M}(s_{j-1}))+d_j(\mathbb{J}(s_j)-\mathbb{J}(s_{j-1})))}\right) \end{aligned}$$

for any vector of constants $(c_1, c_2, \ldots, c_j, d_1, d_2, \ldots, d_j)$. By the Cramer–Wold device it follows that:

$$\left(\{\mathbb{M}_{n_2}(s_i) - \mathbb{M}_{n_2}(s_{i-1})\}_{i=1}^{J}, \{\mathbb{J}_{n_2}(s_i) - \mathbb{J}_{n_2}(s_{i-1})\}_{i=1}^{J}\right)$$

$$\rightarrow_d \left(\{\mathbb{M}(s_i) - \mathbb{M}(s_{i-1})\}_{i=1}^{J}, \{\mathbb{J}(s_i) - \mathbb{J}(s_{i-1})\}_{i=1}^{J}\right) ,$$

establishing the claim. As before:

$$A_n = \int K^{\star}_{n_2}(t, \eta_1, \eta_2, \xi)\ dZ_{n_1}(\eta_1, \eta_2, \xi)$$

where

$$\begin{aligned} & K^{\star}_{n_2}(t, \eta_1, \eta_2, \xi) \\ = \ & E[e^{it\sum_{j\leq J}(c_j(\mathbb{M}^{+}_{n_2}(s_j) - \mathbb{M}^{+}_{n_2}(s_{j-1})) + d_j(\mathbb{J}^{+}_{n_2}(s_j) - \mathbb{J}^{+}_{n_2}(s_{j-1})))} \big| \eta_{n_1,1} = \eta_1, \eta_{n_1,2} = \eta_2, \xi_{n_1} = \xi] \, . \end{aligned}$$

Proceeding as before, the convergence of A_n to A follows if we establish the uniform convergence of $K^{\star}_{n_2}(t, \eta_1, \eta_2, \xi)$ to A on a compact rectangle of the form $[-L, L]^3$. We have:

$$\begin{aligned} & K^{\star}_{n_2}(t, \eta_1, \eta_2, \xi) \\ = \ & E(e^{it\sum_{j\leq J}(c_j(\mathbb{M}^{+}_{n_2}(s_j) - \mathbb{M}^{+}_{n_2}(s_{j-1})) + d_j(\mathbb{J}^{+}_{n_2}(s_j) - \mathbb{J}^{+}_{n_2}(s_{j-1})))} \Big| \eta_{n_1,1} = \eta_1, \eta_{n_1,2} = \eta_2, \xi_{n_1} = \xi) \\ = \ & E\left[\exp\left\{it\sum_{j\leq J}\left(\sum_{i=1}^{n_2}\left(c_j\left(W_i - \frac{\hat{\alpha}_{n_1} + \hat{\beta}_{n_1}}{2}\right) + d_j\right)\left(I\left(U_i \leq d^0 + \frac{s_j}{n_2^{1+\gamma}}\right)\right.\right.\right.\right. \\ & \left.\left.\left.\left. - I\left(U_i \leq d^0\right)\right) - \sum_{i=1}^{n_2}\left(c_j\left(W_i - \frac{\hat{\alpha}_{n_1} + \hat{\beta}_{n_1}}{2}\right) + d_j\right)\left(I\left(U_i \leq d^0 + \frac{s_{j-1}}{n_2^{1+\gamma}}\right)\right.\right.\right. \\ & \left.\left.\left. - I(U_i \leq d^0)\right)\right)\right\} \Big| \eta_{n_1,1} = \eta_1, \eta_{n_1,2} = \eta_2, \xi_{n_1} = \xi\Bigg] \\ = \ & \int E\left[\exp\left\{it\sum_{j\leq J}\left(\sum_{i=1}^{n_2}\left(c_j\left(W_i - \frac{\hat{\alpha}_{n_1} + \hat{\beta}_{n_1}}{2}\right) + d_j\right)\left(I\left(U_i \leq d^0 + \frac{s_j}{n_2^{1+\gamma}}\right)\right.\right.\right.\right. \\ & \left.\left.\left.\left. - I\left(U_i \leq d^0 + \frac{s_{j-1}}{n_2^{1+\gamma}}\right)\right)\right)\right\} \Big| \eta_{n_1,1} = \eta_1, \eta_{n_1,2} = \eta_2, \xi_{n_1} = \xi\Bigg] \end{aligned}$$

$$
\begin{aligned}
&= \int E\left[\exp\left\{it\sum_{i=1}^{n_2}\left(\sum_{j\le J}\left(c_j\left(W_i-\frac{\hat\alpha_{n_1}+\hat\beta_{n_1}}{2}\right)+d_j\right)\left(I\left(U_i\le d^0+\frac{s_j}{n_2^{1+\gamma}}\right)\right.\right.\right.\right.\\
&\quad \left.\left.\left.\left.-I\left(U_i\le d^0+\frac{s_{j-1}}{n_2^{1+\gamma}}\right)\right)\right)\right\}\Big|\eta_{n_1,1}=\eta_1,\eta_{n_1,2}=\eta_2,\xi_{n_1}=\xi\Big]\\
&= \int E\left[\exp\left\{it\left(\sum_{j\le J}\left(c_j\left(W_1-\frac{\hat\alpha_{n_1}+\hat\beta_{n_1}}{2}\right)+d_j\right)\left(I\left(U_1\le d^0+\frac{s_j}{n_2^{1+\gamma}}\right)\right.\right.\right.\right.\\
&\quad \left.\left.\left.\left.-I\left(U_1\le d^0+\frac{s_{j-1}}{n_2^{1+\gamma}}\right)\right)\right)\right\}\Big|\eta_{n_1,1}=\eta_1,\eta_{n_1,2}=\eta_2,\xi_{n_1}=\xi\Big]^{n_2}.
\end{aligned}
$$

As previously, for all sufficiently large n (depending possibly only on C and L), for all $\xi\in[-L,L]$,

$$
\begin{aligned}
&K^\star_{n_2}(t,\eta_1,\eta_2,\xi)\\
&= \left[\int_{[d^0,d^0+s_J/n_2^{1+\gamma}]^C} p_{U_1}(u)\,du\right.\\
&\quad \left.+\sum_{j\le J}\int_{d^0+\frac{s_{j-1}}{n_2^{1+\gamma}}}^{d^0+\frac{s_j}{n_2^{1+\gamma}}} E\left[e^{it\left(c_j\left(W_1-\frac{\alpha_0+\eta_1/\sqrt{n_1}+\beta_0+\eta_2/\sqrt{n_2}}{2}\right)+d_j\right)}\Big|U_1=u\right]p_{U_1}(u)du\right]^{n_2}\\
&= \left[1-\frac{s_J-s_0}{n_2^{1+\gamma}}\,\frac{n_1^\gamma}{2K}\right.\\
&\quad \left.+\sum_{j\le J}\int_{d^0+\frac{s_{j-1}}{n_2^{1+\gamma}}}^{d^0+\frac{s_j}{n_2^{1+\gamma}}} E\left[e^{it\left(c_j\left(W_1-\frac{\alpha_0+\eta_1/\sqrt{n_1}+\beta_0+\eta_2/\sqrt{n_2}}{2}\right)+d_j\right)}\Big|U_1=u\right]p_{U_1}(u)du\right]^{n_2}\\
&= \left[1+\sum_{j\le J}\left[-\frac{s_j-s_{j-1}}{2K\,n_2}\left(\frac{\lambda}{1-\lambda}\right)^\gamma+\frac{n_1^\gamma}{2K}\cdot\frac{1}{n_2^{1+\gamma}}\right.\right.\\
&\quad \times\int_{s_{j-1}}^{s_j} E\left[\exp\left\{it\left(c_j\left(W_1-\frac{\alpha_0+\frac{\eta_1}{\sqrt{n_1}}+\beta_0+\frac{\eta_2}{\sqrt{n_1}}}{2}\right)+d_j\right)\right\}\right.\\
&\quad \left.\left.\left.\Big|U_1=d^0+\frac{v}{1+n_2^{1+\gamma}}\right]dv\right]\right]^{n_2}\\
&= \left[1+\sum_{j\le J}\left[-\frac{(s_j-s_{j-1})}{2Kn_2}\left(\frac{\lambda}{1-\lambda}\right)^\gamma\right.\right.\\
&\qquad\qquad \left.\left.+\frac{1}{2Kn_2}\left(\frac{\lambda}{1-\lambda}\right)^\gamma(s_j-s_{j-1})B^*_{n_1,n_2,\eta_1,\eta_2}(s_j,s_{j-1})\right]\right]^{n_2}\\
&= \left[1+\sum_{j\le J}\left[-\frac{(s_j-s_{j-1})}{2Kn_2}\left(\frac{\lambda}{1-\lambda}\right)^\gamma(1-B^*_{n_1,n_2,\eta_1,\eta_2}(s_j,s_{j-1}))\right]\right]^{n_2}
\end{aligned}
$$

where

$$\begin{aligned} & B^*_{n_1,n_2,\eta_1,\eta_2}(s_j, s_{j-1}) \\ = \; & e^{-it\frac{\eta_1+\eta_2}{2\sqrt{n_1}}} \int_{s_{j-1}}^{s_j} \frac{1}{s_j - s_{j-1}} E\left[e^{it\left(c_j\left(W_1 - \frac{\alpha_0+\beta_0}{2}\right)+d_j\right)} \middle| U_1 = d^0 + \frac{v}{n_2^{1+\gamma}} \right] dv \end{aligned}$$

and

$$\sup_{(\eta_1,\eta_2,\xi)\in[-L,L]^3} \mid K^\star_{n_2}(t,\eta_1,\eta_2,\xi) - A \mid = \sup_{(\eta_1,\eta_2)\in[-L,L]^2} \mid K^\star_{n_2}(t,\eta_1,\eta_2,\xi) - A \mid \, .$$

Setting

$$z_n(\eta_1,\eta_2) = \left[-\sum_{j\le J} \frac{(s_j - s_{j-1})}{2K} \left(\frac{\lambda}{1-\lambda}\right)^\gamma (1 - B^*_{n_1,n_2,\eta_1,\eta_2}(s_j, s_{j-1})) \right]$$

we see that:

$$\sup_{(\eta_1,\eta_2)\in[-L,L]^2} \mid z_n(\eta_1,\eta_2) - z_0 \mid \to 0\,,$$

where

$$z_0 = \exp\left[-\frac{1}{2K} \left(\frac{\lambda}{1-\lambda}\right)^\gamma \sum_{j\le J} (s_j - s_{j-1})(1 - e^{itd_j} L^+(c_j t)) \right] ,$$

since for each $1 \le j \le J$,

$$B^*_{n_1,n_2,\eta_1,\eta_2}(s_j, s_{j-1}) \longrightarrow e^{itd_j} E\left[e^{itc_j\left(w - \frac{\alpha_0+\beta_0}{2}\right)} \middle| U = d^{0+} \right] = e^{itd_j} L^+(c_j t)$$

uniformly over $(\eta_1,\eta_2) \in [-L,L]^2$. It follows that

$$\begin{aligned} & \sup_{(\eta_1,\eta_2)\in[-L,L]^2} \left| K^\star_{n_2}(t,\eta_1,\eta_2,\xi) - e^{z_0} \right| \\ = \; & \sup_{(\eta_1,\eta_2)\in[-L,L]^2} \left| \left(1 + \frac{1}{n_2} z_n(\eta_1,\eta_2)\right)^{n_2} - e^{z_0} \right| \to 0\,. \end{aligned}$$

But $e^{z_0} = A$, as can be verified by direct computation. By the property of indepen-

dent increments of compound Poisson processes, we have:

$$
\begin{aligned}
& E\left[e^{it\sum_{j\le J}(c_j(\mathbb{M}(s_j)-\mathbb{M}(s_{j-1}))+d_j(\mathbb{J}(s_j)-\mathbb{J}(s_{j-1})))}\right] \\
= & \prod_{j\le J} E\left[e^{it(c_j(\mathbb{M}(s_j)-\mathbb{M}(s_{j-1}))+d_j(\mathbb{J}(s_j)-\mathbb{J}(s_{j-1})))}\right] \\
= & \prod_{j\le J}\sum_{l=0}^{\infty}\sum_{m=0}^{l} E\left[e^{it[\sum_{0\le k\le l}(c_j V_k^+ + d_j)-\sum_{0\le k\le m}(c_j V_k^+ + d_j)]}\right]\cdot P(\nu^+(s_j)=l,\nu^+(s_{j-1})=m) \\
= & \prod_{j\le J}\sum_{l=0}^{\infty}\sum_{m=0}^{l} E\left[e^{it(c_j V_{m+1}^+ + \ldots + c_j V_l^+ + (l-m)d_j)}\right] \\
& \times P(\nu^+(s_j)-\nu^+(s_{j-1})=l-m)\cdot P(\nu^+(s_{j-1})=m) \\
= & \prod_{j\le J}\sum_{l=0}^{\infty}\sum_{m=0}^{l}[e^{itd_j}L^+(c_jt)]^{l-m}\cdot\frac{e^{-\left(\frac{s_j-s_{j-1}}{2K}\right)\left(\frac{\lambda}{1-\lambda}\right)^\gamma}}{(l-m)!}\left[\left(\frac{s_j-s_{j-1}}{2K}\right)\left(\frac{\lambda}{1-\lambda}\right)^\gamma\right]^{l-m} \\
& \times\frac{e^{-\left(\frac{s_{j-1}}{2K}\right)\left(\frac{\lambda}{1-\lambda}\right)^\gamma}}{m!}\left[\left(\frac{s_{j-1}}{2K}\right)\left(\frac{\lambda}{1-\lambda}\right)^\gamma\right]^{m} \quad (\text{set} \quad l=m+l') \\
= & \prod_{j\le J}\sum_{m=0}^{\infty}\frac{e^{-\left(\frac{s_{j-1}}{2K}\right)\left(\frac{\lambda}{1-\lambda}\right)^\gamma}}{m!}\left[\left(\frac{s_{j-1}}{2K}\right)\left(\frac{\lambda}{1-\lambda}\right)^\gamma\right]^{m}\sum_{l'=0}^{\infty}\Bigg([e^{itd_j}L(c_jt)]^{l'} \\
& \times\frac{e^{-\left(\frac{s_j-s_{j-1}}{2K}\right)\left(\frac{\lambda}{1-\lambda}\right)^\gamma}}{l'!}\left[\left(\frac{s_j-s_{j-1}}{2K}\right)\left(\frac{\lambda}{1-\lambda}\right)^\gamma\right]^{l'}\Bigg) \\
= & \prod_{j\le J} e^{-\left(\frac{s_j-s_{j-1}}{2K}\right)\left(\frac{\lambda}{1-\lambda}\right)^\gamma(1-e^{itd_j}L^+(c_jt))} \\
= & e^{-\sum_{j\le J}\frac{s_j-s_{j-1}}{2K}\left(\frac{\lambda}{1-\lambda}\right)^\gamma(1-e^{itd_j}L^+(c_jt))} \\
\equiv & A
\end{aligned}
$$

This finishes the proof of finite dimensional convergence. This derivation can be extended readily to allow for s_j's that can also assume negative values; we avoid this here since the derivation involves no new ideas but becomes somewhat more cumbersome.

We finally show that the process $\mathbb{M}_{n_2}$ restricted to $[-C,C]$ is tight. We know that $(\hat{\alpha}_{n_1},\hat{\beta}_{n_1},\hat{d}_{n_1}) \longrightarrow_p (\alpha_0,\beta_0,d^0)$ and $n_1(\hat{d}_{n_1}-d^0)=O_p(1)$. Let

$$\Omega_n = \left\{|\hat{\alpha}_{n_1}-\alpha_0|\le\Delta, |\hat{\beta}_{n_1}-\beta_0|\le\Delta, \hat{d}_{n_1}-\frac{K}{n_1^\gamma} < d^0 - \frac{C}{n_2^{1+\gamma}} < d^0 + \frac{C}{n_2^{1+\gamma}} < d_{n_1}+\frac{K}{n_1^\gamma}\right\}$$

Clearly, $P(\Omega_n) \longrightarrow 1$. The event Ω_n can be written as $(\hat{\alpha}_{n_1}, \hat{\beta}_{n_1}, \hat{d}_{n_1}) \in R_n$, where $H_n(R_n) \longrightarrow 1$, H_n being the joint distribution of $(\hat{\alpha}_{n_1}, \hat{\beta}_{n_1}, \hat{d}_{n_1})$. Note that $\mathbb{M}_{n_2}\mathbf{1}(\Omega_n)$ is also a process in $D(\mathbb{R})$. We verify tightness of $\mathbb{M}_{n_2}\mathbf{1}(\Omega_n)$ restricted to $[-C, C]$. To this end, we verify (the analogue of) Condition (13.14) on Page 143 of Billingsley (1999), with $\beta = 1/2$ and $\alpha = 1$. Once again, let $0 \leq s_1 \leq s \leq s_2 \leq C$,

$$
\begin{aligned}
& E[|\mathbb{M}^+_{n_2}(s) - \mathbb{M}^+_{n_2}(s_1)| \cdot |\mathbb{M}^+_{n_2}(s_2) - \mathbb{M}^+_{n_2}(s)|\mathbf{1}(\Omega_n)] \\
= & \int_{R_n} E[|\mathbb{M}^+_{n_2}(s) - \mathbb{M}^+_{n_2}(s_1)| \cdot |\mathbb{M}^+_{n_2}(s_2) - \mathbb{M}^+_{n_2}(s)| \big| (\hat{\alpha}_{n_1}, \hat{\beta}_{n_1}, \hat{d}_{n_1}) = (\alpha, \beta, d)] dH_n(\alpha, \beta, d)
\end{aligned}
$$

where

$$
\begin{aligned}
& E[|\mathbb{M}^+_{n_2}(s) - \mathbb{M}^+_{n_2}(s_1)| \cdot |\mathbb{M}^+_{n_2}(s_2) - \mathbb{M}^+_{n_2}(s)| \Big| (\hat{\alpha}_{n_1}, \hat{\beta}_{n_1}, \hat{d}_{n_1}) = (\alpha, \beta, d)] \\
= & E_{\alpha,\beta,d}\left[\left| \sum_{i=1}^{n_2} \left(W_i - \frac{\alpha+\beta}{2} \right) \left(I\left(U_i \leq d^0 + \frac{s}{n_2^{1+\gamma}} \right) - I\left(U_i \leq d^0 + \frac{s_1}{n_2^{1+\gamma}} \right) \right) \right| \right. \\
& \left. \times \left| \sum_{i=1}^{n_2} \left(W_i - \frac{\alpha+\beta}{2} \right) \left(I\left(U_i \leq d^0 + \frac{s_2}{n_2^{1+\gamma}} \right) - I\left(U_i \leq d^0 + \frac{s}{n_2^{1+\gamma}} \right) \right) \right| \right] \\
= & E_{\alpha,\beta,d}\left[\left| \sum_{i \neq j} I\left(d^0 + \frac{s_1}{n_2^{1+\gamma}} \leq U_i < d^0 + \frac{s}{n_2^{1+\gamma}} \right) I\left(d^0 + \frac{s}{n_2^{1+\gamma}} \leq U_j < d^0 + \frac{s_2}{n_2^{1+\gamma}} \right) \right. \right. \\
& \left. \left. \left(W_i - \frac{\alpha+\beta}{2} \right) \left(W_j - \frac{\alpha+\beta}{2} \right) \right| \right] \\
\leq & \sum_{i \neq j} E_{\alpha,\beta,d}\left[\left| I\left(d^0 + \frac{s_1}{n_2^{1+\gamma}} \leq U_i < d^0 + \frac{s}{n_2^{1+\gamma}} \right) I\left(d^0 + \frac{s}{n_2^{1+\gamma}} \leq U_j < d^0 + \frac{s_2}{n_2^{1+\gamma}} \right) \right. \right. \\
& \left. \left. \left(W_i - \frac{\alpha+\beta}{2} \right) \left(W_j - \frac{\alpha+\beta}{2} \right) \right| \right] \\
= & \sum_{i \neq j} \sup_{t \in [d^0 + s_1 n_2^{-(1+\gamma)}, d^0 + s n_2^{-(1+\gamma)}]} E_{\alpha,\beta,d}\left[\left| W_i - \frac{\alpha+\beta}{2} \right| \mid U_i = t \right] \frac{n_1^\gamma}{2K} \frac{s - s_1}{n_2^{1+\gamma}} \\
& \times \sup_{t \in [d^0 + s n_2^{-(1+\gamma)}, d^0 + s_2 n_2^{-(1+\gamma)}]} E_{\alpha,\beta,d}\left[\left| W_j - \frac{\alpha+\beta}{2} \right| \mid U_j = t \right] \frac{n_1^\gamma}{2K} \frac{s_2 - s}{n_2^{1+\gamma}} \\
\leq & \sum_{i \neq j} K_1\left(\frac{s - s_1}{2K n_2} \left(\frac{\lambda}{1-\lambda} \right)^\gamma \right) K_2\left(\frac{s_2 - s}{2K n_2} \left(\frac{\lambda}{1-\lambda} \right)^\gamma \right) \quad \text{for any} (\alpha, \beta, d) \in R_n \\
= & K_1 K_2 \sum_{i \neq j} \frac{(s - s_1)(s_2 - s)}{n_2^2 4k^2} \left(\frac{\lambda}{1-\lambda} \right)^{2\gamma} \\
\leq & c^*(s_2 - s_1)^2 .
\end{aligned}
$$

It follows that

$$
E[|\mathbb{M}^+_{n_2}(s) - \mathbb{M}^+_{n_2}(s_1)| \cdot |\mathbb{M}^+_{n_2}(s_2) - \mathbb{M}^+_{n_2}(s)|\mathbf{1}(\Omega_n)] \leq H_n(R_n) c^*(s_2 - s_1)^2 \leq c^*(s_2 - s_1)^2
$$

which establishes tightness of $\mathbb{M}_{n_2}\mathbf{1}(\Omega_n)$.

Then, given any $\epsilon > 0$, $\forall n > N_1$,

$$\text{Prob}[\omega : \mathbb{M}_{n_2}\mathbf{1}(\Omega_n)(\omega) \in \mathcal{K}] \geq 1 - \varepsilon$$

where $\mathcal{K}$ is a compact set. But $\text{Prob}[\omega : \omega \in \Omega_n] \geq 1 - \varepsilon$ eventually. Therefore, eventually

$$\text{Prob}[\omega \in \Omega_n \text{ and } \mathbb{M}_{n_2}\mathbf{1}(\Omega_n) \in \mathcal{K}] \geq 1 - 2\varepsilon$$

and consequently $\text{Prob}[\mathbb{M}_{n_2} \in \mathcal{K}] \geq 1 - 2\varepsilon$. This establishes the tightness of $\mathbb{M}_{n_2}$ in the space of right continuous left limits endowed functions on $[-C, C]$. Similarly, the tightness of $\mathbb{J}_{n_2}$ can be established. This completes the verification of marginal tightness and therefore joint tightness. □

Before embarking on the proof of Theorem 3.1, we need some auxiliary lemmas. We first state these below.

Lemma 3.4. *Let $\tilde{U}_1, \tilde{U}_2, \ldots, \tilde{U}_n$ be i.i.d. Uniform (0,1) random variables. Then, for all $\lambda > 0$ and for all $0 < \alpha < \beta \leq 1$, we have:*

$$Pr\left(\sup_{\alpha \leq s \leq \beta} \frac{\mid \sqrt{n}(\mathbb{P}_n - P)(1(\tilde{u} \leq s)) \mid}{s} \geq \lambda\right) \leq (\alpha^{-1} - \beta^{-1})\lambda^{-2}\,,$$

where $\mathbb{P}_n$ denotes the empirical measure of the data and P the distribution of $\tilde{U}_1$.

This lemma is due to Ferger (2005).

Lemma 3.5. *Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. random elements assuming values in a space $\mathcal{X}$. Let $\mathcal{F}$ be a class of functions with domain $\mathcal{X}$ and range in $[0, 1]$ with finite VC dimension $V(\mathcal{F})$ and set $V = 2(V(\mathcal{F}) - 1)$. Denoting by $\mathbb{P}_n$ the empirical measure corresponding to the sample and by P the distribution of X_1, we have:*

$$Pr^{\star}\left(\|\sqrt{n}(\mathbb{P}_n - P)\|_{\mathcal{F}} \geq \lambda\right) \leq \left(\frac{D\lambda}{\sqrt{V}}\right)^{V} \exp(-2\,\lambda^2)\,.$$

This lemma is adapted from Talagrand (1994).

Lemma 3.6. *Let $\tilde{U}_1, \tilde{U}_2, \ldots, \tilde{U}_n$ be i.i.d. random variables following the uniform distribution on $(0, 1)$. Let $\tilde{\epsilon}_1, \tilde{\epsilon}_2, \ldots, \tilde{\epsilon}_n$ be i.i.d. mean 0 random variables with finite*

variance σ^2 *that are independent of the* $\tilde{U}_i$*'s. Let* $\beta_n(s) = \sum_{i=1}^n \tilde{\epsilon}_i \, 1(\tilde{U}_i \leq s)$*. Then for any* $0 < \alpha < \beta < 1$*, we have:*

$$Pr\left(\sup_{\alpha \leq s \leq \beta} \frac{\mid \beta_n(s) \mid}{s} \geq \lambda\right) \leq (\alpha^{-1} - \beta^{-1})\lambda^{-2}\sigma^2 \, .$$

The proof of this lemma follows the proof of Theorem 3.1.

Lemma 3.7. (Hajek–Renyi inequality) *Consider independent random variables* $X_1, X_2, \ldots,$ *and define* $S_n = \sum_{i=1}^n X_i$*. Further assume that* $E(X_k) = 0$ *and* $E(X_k^2) < \infty$ *for each* k*. Let* $\{c_k\}$ *be a decreasing sequence of positive numbers. Then, for any* $\epsilon > 0$ *and* $n, m > 0$ *with* $n \leq m$*, we have:*

$$P\left(max_{n \leq k \leq m} \mid S_k \mid c_k \geq \epsilon\right) \leq \frac{1}{\epsilon^2}\left[c_n^2 \sum_{i=1}^n E(X_k^2) + \sum_{k=n+1}^m c_k^2 \, E(X_k^2)\right] .$$

Proof of Theorem 3.1: For simplicity and ease of exposition, in what follows, we assume that n points are used at the first stage to compute estimates $\hat{\alpha}_n, \hat{\beta}_n, \hat{d}_{n_1}$ of the three parameters of interest. At the second stage n i.i.d. U_i's are sampled from the uniform distribution on $\tilde{D}_n \equiv [\hat{d}_{n_1} - Kn^{-\gamma}, \hat{d}_{n_1} + Kn^{-\gamma}]$ and the updated estimate of d^0 is computed as

$$\hat{d}_{n_2} = \text{argmin}_{u \in \tilde{D}_n} \frac{1}{n} \sum_{i=1}^n [W_i - \hat{\alpha}_n 1(U_i \leq u) - \hat{\beta}_n 1(U_i > u)]^2 \equiv \text{argmin}_{u \in \tilde{D}_n} \, SS(u) \, .$$

In the above display $W_i = f(U_i) + \epsilon_i$ where ϵ_i's are i.i.d. error variables. Working under this more restrictive setting (of equal allocation of points at each stage) *does not compromise* the complexity of the arguments involved. Finally, recall that by our assumption, $E[e^{C \mid \epsilon_1 \mid}]$ is finite, for some $C > 0$.

Before proceeding further, a word about the definition of argmin in the above display. The function SS is a right–continuous function endowed with left limits. For this derivation, we take the argmin to be the smallest u in $\tilde{D}_n$ for which $\min(SS(u), SS(u-) = \inf_{x \in \tilde{D}_n} SS(x)$.

Denote by G_n the distribution of $(\hat{\alpha}_n, \hat{\beta}_n, \hat{d}_{n_1})$. Now, given $\epsilon > 0$, find L so large that for all sufficiently large n, say $n \geq N_0$,

$$(\hat{\alpha}_n, \hat{\beta}_n, \hat{d}_{n,1}) \in [\alpha_0 - L/\sqrt{n}, \alpha_0 + L/\sqrt{n}] \times [\beta_0 - L/\sqrt{n}, \beta_0 + L/\sqrt{n}] \times [d^0 - L/n, d^0 + L/n]$$

with probability greater than $1 - \epsilon$. Denote the region on the right side of the above display by R_n. Then, for all $n \geq N_0$,

$$\Pr(n^{1+\gamma} \mid \hat{d}_{n_2} - d^0 \mid > a)$$

$$\leq \int_{R_n} \Pr(n^{1+\gamma} \mid \hat{d}_{n_2} - d^0 \mid > a \mid \hat{\alpha}_n = \alpha, \hat{\beta}_n = \beta, \hat{d}_{n_1} = t)\, dG_n(\alpha, \beta, t) + \epsilon$$

which is dominated by

$$\sup_{(\alpha,\beta,t) \in R_n} \Pr_{t,\alpha,\beta}(n^{1+\gamma} \mid \hat{d}_{n_2} - d^0 \mid > a) + \epsilon\,.$$

By making a large, we will show that for all sufficiently large n (say $n > N_1 > N_0$), the supremum is bounded by ϵ. This will complete the proof.

First note that since N_0 is chosen to be sufficiently large, whenever $n \geq N_0$ and $t \in [d^0 - L/n, d^0 + L/n]$, it is the case that $t - Kn^{-\gamma} < d^0 - Kn^{-\gamma}/2 < d^0 + Kn^{-\gamma}/2 < t + Kn^{-\gamma}]$. It is not difficult to see that

$$\begin{aligned} \hat{d}_{n_2} &= \operatorname{argmin}_{d \in [t - Kn^{-\gamma}, t + Kn^{-\gamma}]} \tilde{\mathbb{P}}_n \left[(w - (\alpha + \beta)/2)(1(u \leq d) - 1(u \leq d^0)) \right] \\ &\equiv \operatorname{argmin}_{d \in [t - Kn^{-\gamma}, t + Kn^{-\gamma}]} \tilde{\mathbb{M}}_n(d) \end{aligned}$$

and

$$\begin{aligned} d^0 &= \operatorname{argmin}_{d \in [t - Kn^{-\gamma}, t + Kn^{-\gamma}]} \tilde{P}_n \left[(w - (\alpha + \beta)/2)(1(u \leq d) - 1(u \leq d^0)) \right] \\ &\equiv \operatorname{argmin}_{d \in [t - Kn^{-\gamma}, t + Kn^{-\gamma}]} \tilde{M}_n(d)\,, \end{aligned}$$

where $\tilde{P}_n$ is the distribution of the pair (W_1, U_1) generated at stage two under first stage parameters (α, β, t) and $\tilde{\mathbb{P}}_n$ is the empirical measure corresponding to n i.i.d. observations from $\tilde{P}_n$. Note that

$$\tilde{M}_n(d) = \{\mid \beta_0 - (\alpha+\beta)/2 \mid\mid d - d^0 \mid 1(d \geq d^0) + \mid \alpha_0 - (\alpha+\beta)/2 \mid\mid d - d^0 \mid 1(d < d^0)\}\,(n^\gamma/2K)\,.$$

Now, for $0 < r \leq K/2$, set $a(r) = \min\{\tilde{M}_n(d) :\mid d - d^0 \mid \geq r\, n^{-\gamma}\}$. Then $a(r) = \min(\mid \beta_0 - (\alpha+\beta)/2) \mid, \mid \alpha_0 - (\alpha+\beta)/2) \mid) r/2K$ and let $b(r) = (a(r) - \tilde{M}_n(d^0))/3 = a(r)/3$. Now, for all $n \geq N_0$, for α, β in the region under consideration, $b(r)$ is readily seen to be uniformly bounded below by $\kappa\, r$ for some constant κ depending only on $\alpha_0, \beta_0, K, N_0$. We then have:

$$\sup_{d \in [t - Kn^{-\gamma}, t + Kn^{-\gamma}]} \mid \tilde{\mathbb{M}}_n(d) - \tilde{M}_n(d) \mid \leq b(r) \Rightarrow \mid \hat{d}_{n_2} - d^0 \mid \leq r\, n^{-\gamma}\,. \tag{3.2}$$

To prove this, assume that the inequality on the left side of the above display holds and consider $d \in [t - Kn^{-\gamma}, t + Kn^{-\gamma}]$ with $|d - d^o| > rn^{-\gamma}$. Then,

$$\tilde{\mathbb{M}}_n(d) \geq \tilde{M}_n(d) - b(r) \geq a(r) - b(r)$$

and

$$\tilde{\mathbb{M}}_n(d^o) \leq \tilde{M}_n(d^0) + b(r)$$

$$\Rightarrow \tilde{\mathbb{M}}_n(d) - \tilde{\mathbb{M}}_n(d^o) \geq a(r) - b(r) - \tilde{M}_n(d^0) - b(r) = b(r) > 0$$

Hence

$$\tilde{\mathbb{M}}_n(d) > \tilde{\mathbb{M}}_n(d^0).$$

Now, since $\hat{d}_{n_2}$ is the smallest $d \in \tilde{D}_n$ for which $\tilde{\mathbb{M}}_n(d) \wedge \tilde{\mathbb{M}}_n(d-) = \inf_{x \in \tilde{D}_n} \tilde{\mathbb{M}}_n(x)$ and $\tilde{\mathbb{M}}_n$ is a (right continuous left limits endowed) piecewise constant function with finitely many flat stretches, it is easy to see that $\tilde{\mathbb{M}}_n(\hat{d}_{n_2}) = \inf_{x \in \tilde{D}_n} \tilde{\mathbb{M}}_n(x)$. Therefore, $\tilde{\mathbb{M}}_n(\hat{d}_{n_2}) \leq \tilde{\mathbb{M}}_n(d^0)$, showing that $\mid \hat{d}_{n_2} - d^0 \mid \leq r\, n^{-\gamma}$ in view of the last display above.

Now, consider

$$\Pr_{\alpha,\beta,t}(\mid \hat{d}_{n_2} - d^0 \mid > r\, n^{-\gamma})$$

$$\leq \Pr_{\alpha,\beta,t}(rn^{-\gamma} < \mid \hat{d}_{n_2} - d^0 \mid \leq \delta\, n^{-\gamma}) + \Pr_{\alpha,\beta,t}(\mid \hat{d}_{n_2} - d^0 \mid > \delta\, n^{-\gamma}) \tag{3.3}$$

$$\equiv P_n(\alpha, \beta, t) + Q_n(\alpha, \beta, t)\,, \tag{3.4}$$

where δ (is sufficiently small, say less than $K/3$) does not depend on t, α, β. We deal with $Q_n(\alpha, \beta, t)$ later. We first consider $P_n(\alpha, \beta, t) \equiv \Pr_{\alpha,\beta,t}(rn^{-\gamma} <\mid \hat{d}_{n_2} - d^0 \mid \leq \delta\, n^{-\gamma})$. Since,

$$\{rn^{-\gamma} < |\hat{d}_{n_2} - d^o| \leq \delta n^{-\gamma}\} \subseteq \left[\cup_{d^0+rn^{-\gamma}<d\leq d^0+\delta n^{-\gamma}}\{\tilde{\mathbb{M}}_n(d) \leq \tilde{\mathbb{M}}_n(d^0)\}\right]$$

$$\cup \left[\cup_{d^0-\delta n^{-\gamma}\leq d<d^0-rn^{-\gamma}}\{\tilde{\mathbb{M}}_n(d) \leq \tilde{\mathbb{M}}_n(d^0)\}\right],$$

we conclude that

$$\begin{aligned} P_n(\alpha,\beta,t) &\leq P_{n,1}(\alpha,\beta,t) + P_{n,2}(\alpha,\beta,t) \\ &\equiv \Pr_{\alpha,\beta,t}(\cup_{d^0+rn^{-\gamma}<d\leq d^0+\delta n^{-\gamma}}\{\tilde{\mathbb{M}}_n(d^0) - \tilde{\mathbb{M}}_n(d) \geq 0\}) \\ &\quad +\Pr_{\alpha,\beta,t}(\cup_{d^0-\delta n^{-\gamma}\leq d<d^0-rn^{-\gamma}}\{\tilde{\mathbb{M}}_n(d^0) - \tilde{\mathbb{M}}_n(d) \geq 0\}). \end{aligned}$$

We first construct an upper bound on $\sup_{(\alpha,\beta,t)\in R_n} P_{n,1}(\alpha,\beta,t)$. For any $d \in (d^0 + rn^{-\gamma}, d^0 + \delta n^{-\gamma}]$ we have:

$$\begin{aligned} \tilde{\mathbb{M}}_n(d^0) - \tilde{\mathbb{M}}_n(d) &= (\tilde{\mathbb{M}}_n(d^0) - \tilde{M}_n(d^0)) - (\tilde{\mathbb{M}}_n(d) - \tilde{M}_n(d)) - (\tilde{M}_n(d) - \tilde{M}_n(d^0)) \\ &= -(\tilde{\mathbb{M}}_n(d) - \tilde{M}_n(d)) - \left|\beta_0 - \frac{\alpha+\beta}{2}\right| \frac{n^\gamma}{2K} \mid d - d^0 \mid . \end{aligned}$$

Hence:

$$0 \leq \tilde{\mathbb{M}}_n(d^0) - \tilde{\mathbb{M}}_n(d) \Rightarrow (2K)^{-1} \left|\beta_0 - \frac{\alpha+\beta}{2}\right| \leq \frac{-(\tilde{\mathbb{M}}_n(d) - \tilde{M}_n(d))}{n^\gamma \mid d - d^0 \mid}.$$

Now, for all $(\alpha, \beta, t) \in R_n$ (with $n \geq N_0$), $\left|\beta_0 - \frac{\alpha+\beta}{2}\right|$ is bounded below by some constant B, whence it follows that:

$$0 \leq \tilde{\mathbb{M}}_n(d^0) - \tilde{\mathbb{M}}_n(d) \Rightarrow \frac{\mid \tilde{\mathbb{M}}_n(d) - \tilde{M}_n(d) \mid}{n^\gamma \mid d - d^0 \mid} \geq \frac{B}{2K}.$$

Thus,

$$\begin{aligned} &\cup_{d^0-\delta n^{-\gamma}<d\leq d^0-rn^{-\gamma}}\{\tilde{\mathbb{M}}_n(d^0) - \tilde{\mathbb{M}}_n(d) \geq 0\} \\ \subset\ &\{\sup_{d^0+rn^{-\gamma}<d\leq d^0+\delta n^{-\gamma}} \frac{\mid \tilde{\mathbb{M}}_n(d) - \tilde{M}_n(d) \mid}{n^\gamma \mid d - d^0 \mid} \geq \tilde{B}\}, \end{aligned}$$

where $\tilde{B} = B/2K$. We thus have:

$$\begin{aligned} P_{n,1}(\alpha,\beta,t) &\leq \Pr_{\alpha,\beta,t}\left[\sup_{d^0+rn^{-\gamma}<d\leq d^0+\delta n^{-\gamma}} \frac{\mid \tilde{\mathbb{M}}_n(d) - \tilde{M}_n(d) \mid}{n^{\gamma} \mid d-d^0 \mid} \geq \tilde{B}\right] \\ &= \Pr_{\alpha,\beta,t}\left(\sup_{d^0+rn^{-\gamma}<d\leq d^0+\delta n^{-\gamma}} \frac{\mid \sqrt{n}(\tilde{\mathbb{P}}_n - \tilde{P}_n) f_{d,\alpha,\beta}(u,w) \mid}{\mid d-d^0 \mid n^{\gamma}} \geq \sqrt{n}\,\tilde{B}\right), \end{aligned}$$

where

$$f_{d,\alpha,\beta}(u,w) = (w - (\alpha+\beta)/2)(1(u\leq d) - 1(u\leq d^0)).$$

Using the fact that for $d > d^0$, $(W_j - (\alpha+\beta)/2)(1(U_j \leq d) - 1(U_j \leq d^0)) = (\beta_0 - (\alpha+\beta)/2)(1(U_j \leq d) - 1(U_j \leq d^0)) + \epsilon_j (1(U_j \leq d) - 1(U_j \leq d^0))$, this upper bound on $P_{n,1}(\alpha,\beta,t)$ is easily seen to be dominated by $I_n + II_n$ where

$$\begin{aligned} & I_n \\ = \; & \Pr_{\alpha,\beta,t}\left(\mid \beta_0 - (\alpha+\beta)/2 \mid \sup_{r<s\leq\delta} \tfrac{|\sqrt{n}(\tilde{\mathbb{P}}_n - \tilde{P}_n)(1(u\leq d^0+sn^{-\gamma}) - 1(u\leq d^0))|}{s} \geq \sqrt{n}\,\tilde{B}/2\right) \end{aligned}$$

which in turn is dominated by

$$I'_n = \Pr_{\alpha,\beta,t}\left(\sup_{r<s\leq\delta} \frac{\mid \sqrt{n}(\tilde{\mathbb{P}}_n - \tilde{P}_n)(1(u\leq d^0+sn^{-\gamma}) - 1(u\leq d^0)) \mid}{s} \geq \frac{\sqrt{n}\,\tilde{B}}{2B'}\right)$$

where, for $n \geq N_0$ and $(\alpha,\beta,t) \in R_n$, B' is an upper bound on $\mid \beta_0 - (\alpha+\beta)/2 \mid$, while

$$II_n = \Pr_t\left(\sup_{r<s\leq\delta} \frac{|(1/\sqrt{n})\sum_{i=1}^n \epsilon_i(1(U_i \leq d^0 + s\,n^{-\gamma}) - 1(U_i \leq d^0))|}{s} \geq \sqrt{n}\tilde{B}/2\right).$$

Since the U_i's are i.i.d. Uniform on $[t - Kn^{-\gamma}, t + Kn^{-\gamma}]$, it is easy to see that I'_n is simply

$$\Pr\left(\sup_{r<s\leq\delta} \frac{\mid \sqrt{n}(\mathbb{Q}_n - Q)(1(\tilde{w}\leq s)) \mid}{s} \geq \frac{\sqrt{n}\,\tilde{B}}{2B'}\right)$$

where $\tilde{W}_1, \tilde{W}_2, \ldots, \tilde{W}_n$ are i.i.d. Unif$[0, 2K]$, $\mathbb{Q}_n$ is the empirical measure of the $\tilde{W}_i$s and Q is the distribution of $\tilde{W}_1$. In terms of $\tilde{U}_1, \tilde{U}_2, \ldots, \tilde{U}_n$, which are i.i.d. Unif$[0,1]$, this expression is simply:

$$\Pr\left(\sup_{r/2K<s\leq\delta/2K} \frac{\mid \sqrt{n}(\mathbb{P}_n - P)(1(\tilde{u}\leq s)) \mid}{s} \geq \frac{\sqrt{n}\,2K\tilde{B}}{2B'}\right).$$

By Lemma 3.4, this is bounded above by a constant (that depends only on $\alpha_0, \beta_0, K, N_0$) times $1/rn$. Now, in terms of $\tilde{U}_1, \cdots, \tilde{U}_n$ and $\tilde{\epsilon}_1, \cdots, \tilde{\epsilon}_n$ (where the $\tilde{\epsilon}_i$'s are defined on the same probability space as the $\tilde{U}_i$'s, but independently of them, and are distributed like the ϵ_i's) II_n is simply:

$$\Pr\left(\sup_{r/2K<s\leq\delta/2K} \frac{\mid (1/\sqrt{n})\ \sum_{i=1}^{n} \tilde{\epsilon}_i(1(\tilde{U}_i \leq s)) \mid}{s} \geq \frac{\sqrt{n}\, 2K\tilde{B}}{2}\right)$$

and this, by Lemma 3.6 , is dominated up to a constant (that only depends on $\alpha_0, \beta_0, \sigma, K, N_0$) by $(1/rn)$. It follows that for some constant C_0, for all $n \geq N_0$,

$$\sup_{(\alpha,\beta,t)\in R_n} P_{n1}(\alpha, \beta, t) \leq \frac{C_0}{rn}\,.$$

A similar (uniform) bound works $P_{n2}(\alpha, \beta, t)$. It follows that

$$\sup_{(\alpha,\beta,t)\in R_n} P_n(\alpha, \beta, t) \leq \frac{C_0}{rn}$$

at the expense of a larger constant C_0. Thus, from (3.4), we have:

$$\sup_{(\alpha,\beta,t)\in R_n} \Pr_{\alpha,\beta,t}(\mid \hat{d}_{n_2} - d^0 \mid > r\, n^{-\gamma}) \leq C_0\,(rn)^{-1} + \sup_{(\alpha,\beta,t)\in R_n} Q_n(\alpha, \beta, t)\,.$$

To find a uniform upper bound on $Q_n(\alpha, \beta, t)$ note that, from (3.2), we have, for all $n > N_0$

$$\begin{aligned}
\Pr_{\alpha,\beta,t}(\mid \hat{d}_{n_2} - d^0 \mid > \delta n^{-\gamma}) &\leq \Pr_{\alpha,\beta,t}(\sup_{d\in[t-Kn^{-\gamma}, t+Kn^{-\gamma}]} \mid \tilde{\mathbb{M}}_n(d) - \tilde{M}_n(d) \mid > b(\delta)) \\
&\leq \Pr_{\alpha,\beta,t}(\sup_{d\in[t-Kn^{-\gamma}, t+Kn^{-\gamma}]} \mid \tilde{\mathbb{M}}_n(d) - \tilde{M}_n(d) \mid > \kappa\,\delta)
\end{aligned}$$

and it suffices to find a uniform upper bound for this last expression. But this is bounded by

$$\begin{aligned}
&\Pr_{\alpha,\beta,t}[\sup_{d\in[t-Kn^{-\gamma}, t+Kn^{-\gamma}]} \mid \sqrt{n}(\tilde{\mathbb{P}}_n - \tilde{P}_n)(\mu(u) - (\alpha+\beta)/2)(1(u \leq d) - 1(u \leq d^0)) \mid \\
&\qquad\qquad > \sqrt{n}\kappa\delta/2] \\
&+ \Pr_{\alpha,\beta,t}[\sup_{d\in[t-Kn^{-\gamma}, t+Kn^{-\gamma}]} \mid n^{-1} \sum_{i=1}^{n} \epsilon_i(1(U_i \leq d) - 1(U_i \leq d^0)) \mid > \kappa\,\delta/2]\,.
\end{aligned}$$

To tackle the first term, we invoke Lemma 3.5. For $(\alpha, \beta, t) \in R_n$, the class $[(\mu(u) - (\alpha+\beta)/2)(1(u \le d) - 1(u \le d^0)) : d \in [t - Kn^{-\gamma}, t + Kn^{-\gamma}]]$ is a bounded VC class of functions (with the bound not depending on α, β, t) and with finite VC dimension, say $\mathcal{V}$ (which does not depend on α, β, t). Hence, we can apply Lemma 3.5 to conclude that:

$$\begin{aligned}\mathrm{Pr}_{\alpha,\beta,t}[\sup_{d\in[t-Kn^{-\gamma},t+Kn^{-\gamma}]} \mid \sqrt{n}(\tilde{\mathbb{P}}_n - \tilde{P}_n)(\mu(u) - (\alpha+\beta)/2)(1(u \le d) - 1(u \le d^0)) \mid \\ > \sqrt{n}\kappa\delta/2] \\ \le \tilde{C}_1 \times (\sqrt{n}\kappa\,\delta)^{2(\mathcal{V}-1)} \exp(-\tilde{C}_2 n\kappa^2\delta^2)\,,\end{aligned}$$

where the constants $\tilde{C}_1$ and $\tilde{C}_2$ depend solely on the VC dimension and the upper bound on the functions. For all sufficiently large n, the right side of the above display is less than $\epsilon/3$. To deal with the second term, we use the results on Pages 132–133 of Van de Geer (2000). We write the second term as:

$$\int \mathrm{Pr}_{\alpha,\beta,t}(\sup_{d\in[t-Kn^{-\gamma},t+Kn^{-\gamma}]} \mid n^{-1}\sum_{i=1}^{n} \epsilon_i(1(u_i \le d) - 1(u_i \le d^0)) \mid > \kappa\,\delta/2) \\ dH_n(u_1, u_2, \ldots, u_n)$$

where H_n is the joint distribution of $(U_1, U_2, \ldots, U_n)$. For each fixed $(u_1, u_2, \ldots, u_n)$, Corollary 8.8 of Van de Geer (2000) can be used to show that for δ sufficiently small and n sufficiently large (where the thresholds do not depend on the u_i's or α, β, t),

$$\begin{aligned}\mathrm{Pr}_{\alpha,\beta,t}(\sup_{d\in[t-Kn^{-\gamma},t+Kn^{-\gamma}]} \mid n^{-1}\sum_{i=1}^{n} \epsilon_i(1(u_i \le d) - 1(u_i \le d^0)) \mid > \kappa\,\delta/2) \\ \le \tilde{C}\exp(-\tilde{C}'\,n\,\delta^2)\,,\end{aligned}$$

for some constants $\tilde{C}$ and $\tilde{C}'$ that do not depend on α, β, t or the points $(u_1, u_2, \ldots, u_n)$. This implies that the second term can be made less than $\epsilon/3$ by choosing n sufficiently large. It follows, that for all sufficiently large n (say $n > N_1 > N_0$) and an appropriate choice of δ, we have:

$$\sup_{(\alpha,\beta,t)\in R_n} \mathrm{Pr}_{\alpha,\beta,t}(\mid \hat{d}_{n_2} - d^0 \mid > r\,n^{-\gamma}) \le C_0\,(rn)^{-1} + 2\epsilon/3\;;$$

the first term on the right side can be made less than $\epsilon/3$ by choosing $r = A/n$ where A is large enough, showing that for all sufficiently large n, we can find A large enough so that:

$$\sup_{(\alpha,\beta,t)\in R_n} \Pr_{\alpha,\beta,t}(n^{1+\gamma} \mid \hat{d}_{n_2} - d^0 \mid > A) < \epsilon\,.$$

It remains to say a few words about the application of Corollary 8.8 of Van de Geer (2000). The W_is in that lemma are our ϵ_is and we can, without loss, take $\mathcal{G} = \{g_d(u) = 1(u \le d) - 1(u \le d^0) : d \in [0,1]\}$. The z_i's are our u_i's. The moment generating function condition on our errors implies that Condition (8.23) is satisfied for some σ_0^2; with $Q_n \equiv \sum_{i=1}^n \delta_{u_i}/n$, it is easy to see that the first condition of the corollary: $\sup_{d\in[0,1]} \|g_d\|_{Q_n} \le R$ is satisfied with $R = 1$. Condition (8.24) is satisfied with $K_2 = 1$. If δ is chosen to satisfy conditions (8.26) and (8.27), it is easy to see that Condition (8.28) will be automatically satisfied for sufficiently large n. This is based on the fact that $H_B(v, \mathcal{G}, Q_n) \le \log(n+1)$ for $v > 1/\sqrt{n}$. To see this, construct brackets $\{[1(u \le u_i) - 1(u \le d^0), 1(u \le u_{i+1}) - 1(u \le d^0)]\}_{i=0}^n$, where $u_0 \equiv 0$ and $u_{n+1} \equiv 1$; any g_d lies in some bracket and the size of each of these brackets with respect to the Q_n norm is no larger than $1/\sqrt{n}$. Given a choice of δ, consider the right side of display (8.28) in Van de Geer (2000). This is given by

$$C_0 \int_{\delta/2^6}^{\sqrt{2}R\sigma_0} H_B^{1/2}(u/\sqrt{2}\sigma_0, \mathcal{G}, Q_n)\, du$$

and for all sufficiently large n, is readily seen to be dominated by

$$\sqrt{2}\,\sigma_0\, C_0 \int_{1/\sqrt{n}}^{R} H_B^{1/2}(v, \mathcal{G}, Q_n)\, dv$$

which in turn is no larger than

$$\sqrt{2}\,\sigma_0\, C_0\, (R - 1/\sqrt{n})\, \sqrt{\log(n+1)}\,.$$

But this quantity is certainly dominated by $\sqrt{n}\,\delta$ for large enough n. □

Proof of Lemma 3.6: Let $\beta_n(s) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\epsilon}_i 1(x_i \leq s)$. Let $\{s_k = \alpha + (\beta - \alpha)\, 2^{-m} : 0 \leq k \leq 2^m\}$, $m \in \mathbb{N}$ be a dyadic partition of $[\alpha, \beta]$. Consider:

$$\begin{aligned} P(\alpha, \beta, \lambda) &= P\left(\sup_{\alpha \leq s \leq \beta} \frac{|\beta_n(s)|}{s} \geq \lambda \right) \\ &= \lim_{m \to \infty} P\left(\max_{1 \leq k \leq 2^m} \frac{|\beta_n(s_k)|}{s_k} \geq \lambda \right) \\ &= \lim_{m \to \infty} \int_{(x_1, \ldots, x_n) \in (0,1)^n} P\left(\max_{1 \leq k < 2^m} \frac{\left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\epsilon}_i 1(x_i \leq s_k) \right|}{s_k} \geq \lambda \right) \\ & \qquad\qquad dx_1 dx_2 \cdots dx_n \end{aligned}$$

For fixed $(x_1, x_2, \ldots, x_n) \in (0,1)^n$, set

$$M_k = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\epsilon}_i 1(x_i \leq s_k)\ ,\ 0 \leq k \leq 2^m\,.$$

Define $X_k = M_k - M_{k-1}$ for $k \geq 1$. Then the X_k's are independent random variables, each with mean 0 and finite variance and $M_k = X_1 + X_2 + \ldots + X_k$. Since $1/s_k$ is a decreasing sequence of constants, we can apply the Hajek-Renyi inequality (Lemma 3.7) to conclude that:

$$\begin{aligned} P\left(\max_{1 \leq k \leq 2^m} \frac{|M_k|}{s_k} \geq \lambda \right) &\leq \frac{1}{\lambda^2} \sum_{1 \leq k \leq 2^m} s_k^{-2} E(M_k - M_{k-1})^2 \\ &= \frac{1}{\lambda^2} \sum_{1 \leq k \leq 2^m} s_k^{-2} (EM_k^2 - EM_{k-1}^2) \end{aligned}$$

Now,

$$\begin{aligned} EM_k^2 &= E[(\frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\epsilon}_i 1(x_i \leq s_k))^2] \\ &= \mathrm{Var}[\frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\epsilon}_i 1(x_i \leq s_k)]^2 \\ &= \frac{1}{n} \sigma^2 \sum_{i=1}^n 1(x_i \leq s_k)\,. \end{aligned}$$

It follows that

$$EM_k^2 - EM_{k-1}^2 = \frac{\sigma^2}{n} \sum_{i=1}^n 1(s_{k-1} < x_i \leq s_k)\,.$$

Therefore

$$\begin{aligned} P(\alpha,\beta,\lambda) &\leq \lim_{m\to\infty}\int_{(x_1,\ldots,x_n)\in(0,1)^n}\left\{\frac{1}{\lambda^2}\sum_{1\leq k<2^m} s_k^{-2}\frac{\sigma^2}{n}\sum_{i=1}^{n}1(s_{k-1}<x_i\leq s_k)\right\} \\ &\qquad\qquad dx_1dx_2\cdots dx_n \\ &= \lim_{m\to\infty}\frac{1}{\lambda^2}\sum_{1\leq k<2^m} s_k^{-2}\frac{\sigma^2}{n}\sum_{i=1}^{n}\int 1(s_{k-1}<x_i\leq s_k)dx_i \\ &= \lim_{m\to\infty}\frac{1}{\lambda^2}\sigma^2\sum_{1\leq k<2^m} s_k^{-2}(s_k-s_{k-1}) \\ &= \frac{\sigma^2}{\lambda^2}\int_\alpha^\beta\frac{1}{s^2}ds \\ &= (\alpha^{-1}-\beta^{-1})\frac{\sigma^2}{\lambda^2} \end{aligned}$$

□

3.2 Strategies for Parameter Allocation in Finite Samples

In this section, we describe strategies for selecting the tuning parameters K, γ and λ used in the procedure. We do this in the setting of the simple regression model $\mu(x) = \alpha_0\, 1(x \leq d^0) + \beta_0\, 1(x > d^0)$ and homoscedastic normal errors, obvious analogues holding in more general settings.

Recall that $(\hat{d}_{n_2,l}, \hat{d}_{n_2,u})$ are the minimal and maximal minimizers at Step 2. Set: $\hat{d}_{2,av} = (\hat{d}_{n_2,l} + \hat{d}_{n_2,u})/2$. In what follows we use this as our second stage estimate of the change–point. Using notation from Theorem 3.2, we have:

$$n_2^{1+\gamma}(\hat{d}_{2,av} - d^0) \to_d \frac{d_l + d_u}{2}\,.$$

It is also not difficult to see that this limit distribution is symmetric about 0.

Henceforth, the notation Argmin will denote the simple average of the minimal and maximal minimizers of a compound Poisson process. The quantity $\mid \alpha_0 - \beta_0 \mid /\sigma$ will be denoted as SNR (signal-to-noise ratio). The higher the SNR, the more advantageous the estimation of the change point at any given sample size would be. By (2.2) and (2.3), we have:

$$\mathbb{M} \equiv \mathbb{M}_{|\alpha_0-\beta_0|,\epsilon,C(K,\lambda,\gamma)} \equiv_d \frac{1}{C(K,\lambda,\gamma)}\,\mathbb{M}_{\mathrm{SNR},Z,1}\,,$$

where Z is standard normal random variable. This is a consequence of the fact that $\epsilon/\sigma \sim N(0,1)$. From Theorem 3.2 and the above display, we have:

$$n^{1+\gamma}(\hat{d}_{2,av} - d^0) \rightarrow_d \frac{1}{(1-\lambda)^{1+\gamma}} 2K \left(\frac{1-\lambda}{\lambda}\right)^{\gamma} \operatorname{Argmin} \mathbb{M}_{\mathrm{SNR},Z,1}$$

$$\equiv_d \frac{2K}{(1-\lambda)\,\lambda^{\gamma}} \operatorname{Argmin} \mathbb{M}_{\mathrm{SNR},Z,1}\,.$$

For a fixed K and γ this immediately provides an optimal allocation for λ. We should choose λ so as to maximize $\gamma \log \lambda + \log(1-\lambda)$ which occurs at $\lambda_{opt} = \gamma/(1+\gamma)$. It can be seen that the approximate standard deviation of $\hat{d}_{2,av}$ is then given by $n^{-(1+\gamma)}(2\,K\,\tau/\Psi(\gamma))$ where $\Psi(\gamma) = \gamma^{\gamma}/(1+\gamma)^{1+\gamma}$. This is actually decreasing in γ.

Consider now a one–stage procedure with the covariates sampled from a density f_X, with the estimate of the change–point once again chosen to be the simple average of the minimal and maximal minimizers; call this $\hat{d}_{av}$. In this case, the standard change–point asymptotics in conjunction with (2.2) and (2.3) give:

$$n\,(\hat{d}_{av} - d^0) \rightarrow_d \frac{1}{f_X(d^0)} \operatorname{Argmin} \mathbb{M}_{\mathrm{SNR},Z,1}\,.$$

This immediately provides an expression of the asymptotic efficiency of the two–stage procedure with respect to the one–stage (in terms of ratios of approximate standard deviations) given by:

$$\mathrm{ARE}_{2,1}(n) = \frac{n^{\gamma}\,\Psi(\gamma)}{2\,K\,f_X(d^0)}\,.$$

In finite samples, how do we choose our K and γ? For any γ, note that the interval from which sampling at Stage 2 takes place is of the form $[\hat{d}_{n_1} - K\,n_1^{-\gamma}, \hat{d}_{n_1} + K\,n_1^{-\gamma}]$. The requirement that this interval contains d^0 with probability increasing to 1 (with increasing n) translates to choosing $K \equiv K_{(n)}$ in such a way that $K_{(n)}\,n_1^{-\gamma} \approx C_\zeta/n_1$ where C_ζ is the upper ζ'th quantile of the distribution of $\operatorname{Argmin} \mathbb{M}_{\mathrm{SNR},Z,1}$ (which can be shown to be symmetric about 0). Here ζ is a very small fraction, say, .0005. In other words, we want to "zoom in" but not "zoom in" so much that we systematically start missing d^0 in our sampling interval.

Now, setting $K \equiv K_{(n)} = C_\zeta / n_1^{1-\gamma}$, writing $n_1 = n\,\gamma/(1+\gamma)$ and using the form of the function Ψ , we get an approximate formula for the ARE:

$$\mathrm{ARE}_{2,1}(n) \approx \frac{n}{2\,C_\zeta\,f_X(d^0)}\;\frac{\gamma}{(1+\gamma)^2}\,.$$

The latter is maximized at $\gamma = 1$, which corresponds to an allocation of 50% points at Stage 1 (and the remainder at Stage 2) and gives the approximate ARE as:

$$\mathrm{ARE}_{2,1}(n) \approx \frac{n}{8\,C_\zeta\,f_X(d^0)}\,. \tag{3.5}$$

It is not difficult to see that the same approximate formula for the ARE holds for some other measures of dispersion, besides the standard deviation. Let

$$\mathrm{ARE}_{2,1,\mathrm{MAD}}(n) \equiv \frac{E \mid \hat{d}_{av} - d^0 \mid}{E \mid \hat{d}_{2,av} - d^0 \mid}$$

and

$$\mathrm{ARE}_{2,1,\mathrm{IQR}}(n) \equiv \frac{\mathrm{IQR}(\hat{d}_{av})}{\mathrm{IQR}(\hat{d}_{2,av})}\,,$$

where both first and second stage estimates are based on samples of size n, and $\mathrm{IQR}(X)$ denotes the inter–quartile range of the distribution of a random variable X. Then, following similar steps to those involved in calculating the ARE based on standard deviations, we conclude that:

$$\mathrm{ARE}_{2,1,\mathrm{MAD}}(n) \approx \mathrm{ARE}_{2,1,\mathrm{IQR}}(n) \approx \frac{n}{8\,C_\zeta\,f_X(d^0)}\,.$$

The accuracy of the above approximation is confirmed empirically through a simulation study. The setting involves a change-point model given by

$$y_i = 0.5I(x \le 0.5) + 1.5I(x > d^0) + \epsilon_i,\;\; x_i \in (0,1). \tag{3.6}$$

The variance σ^2 was chosen so that the SNR defined as $(\beta_0 - \alpha_0)/\sigma = 1, 2, 5$ and 8 and the sample size varies in increments of 50 from 50 to 1500. The results based on an interval corresponding to $\alpha = .0025$ are shown in Figure 3.4. It can be seen that there is great agreement between the theoretical formula for the ARE and the

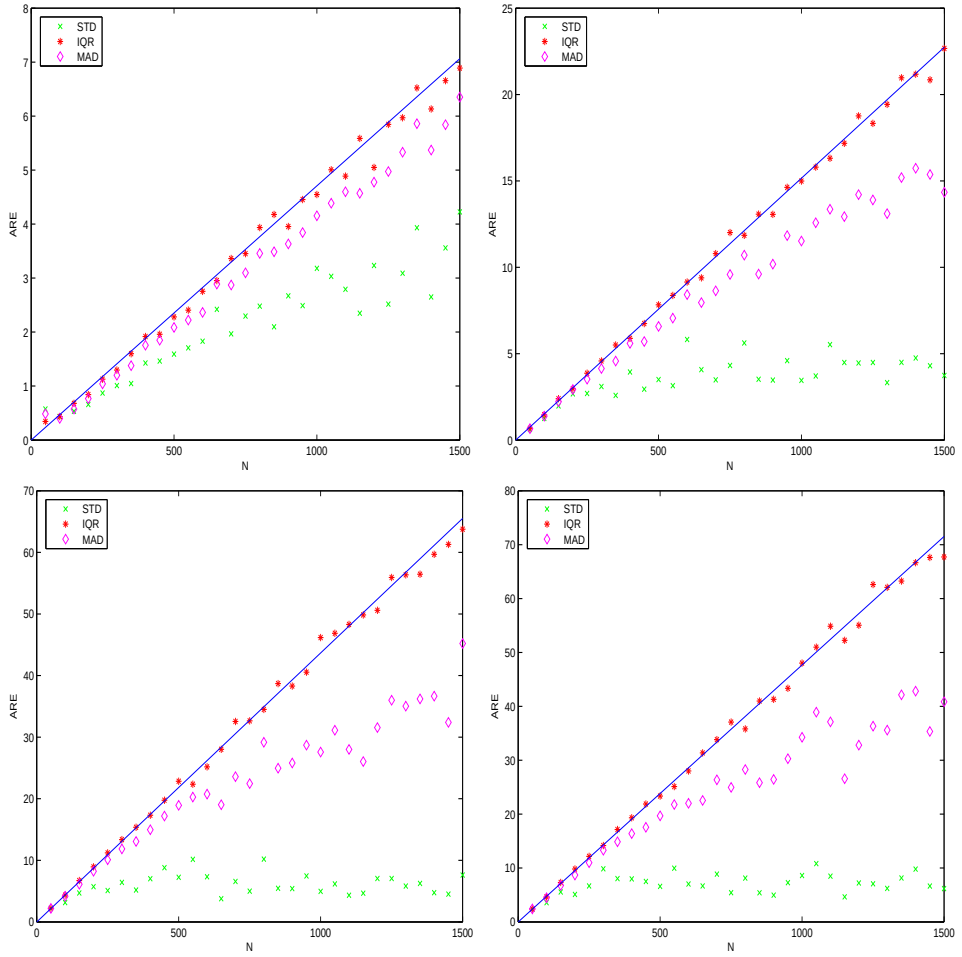


Figure 3.4: Top panels: ARE for standard deviation, IQR and MAD measures for SNR=1 (left) and 2 (right). Bottom panels: Corresponding ARE for SNR=5 (left) and 8 (right).

empirical ARE, especially for the IQR at all SNR levels and to a large extent for the MAD measure. On the other hand, the presence of 'outliers' amongst the first stage estimates introduces too much variability, which in turn leads to inaccuracies for the proposed formula with the standard deviation as a measure of efficiency.

Remark: The formula for the ARE in (3.5) says that the "agnostic" two stage procedure ("agnostic" since the covariates are sampled uniformly at each stage) will eventually, i.e. with increasing n, surpass any one stage procedure, no matter the amount of background information incorporated about the location of the change–point in the one stage process, so long as there is uncertainty about the exact location. One can think of an "oracle–type" one stage procedure where the experimenter samples the covariates from a density that peaks in a neighborhood of d^0 relative to the uni-

form density (corresponding to high values of $f_X(d^0)$). The faster convergence rate of the two stage procedure relative to this one stage procedure guarantees that with increasing n, the ARE will always go to infinity. Further, expression (3.5) provides an approximation to the minimal sample size required for the two-stage procedure to outperform the "classical" one, a result verified through simulations (not shown here).

Remark: A uniform density has been considered up to this point for sampling second stage covariate-response pairs. We examine next the case of using an arbitrary sampling density $g_U(\cdot)$ supported on the interval $[\hat{d}_{n_1} - Kn_1^{-\gamma}, \hat{d}_{n_1} + Kn_1^{-\gamma}]$ and symmetric around $\hat{d}_{n_1}$. A natural choice for such a density is

$$g_U(d^0) = h\left(\frac{d^0 - \hat{d}_{n_1}}{n_1^{-\gamma}K}\right)\frac{n_1^{\gamma}}{K},$$

for a density $h(\cdot)$ supported on $[-1, 1]$ and symmetric about 0. Analogous arguments to those used in the proof of Theorem 3.2 establish that the random vector $n_2^{1+\gamma}(\hat{d}_{n_2,l} - d^0, \hat{d}_{n_2,u} - d^0)$ converges in distribution to

$$(d_l(\mid \alpha_0 - \beta_0 \mid, \epsilon, C(K, \lambda, \gamma, h)), d_u(\mid \alpha_0 - \beta_0 \mid, \epsilon, C(K, \lambda, \gamma, h)))\,,$$

where

$$C(K, \lambda, \gamma, h) = \left(\frac{\lambda}{1-\lambda}^{\gamma}\right)\frac{h(0)}{K}.$$

With the error term normally distributed, as assumed in this Section, we get

$$n^{1+\gamma}(\hat{d}_{2,av} - d^0) \rightarrow_d \frac{K}{h(0)(1-\lambda)\,\lambda^{\gamma}}\ \mathrm{Argmin}\,\mathbb{M}_{\mathrm{SNR},Z,1}\,,$$

and it can be readily checked that the approximate ARE formula reduces to

$$ARE_{2,1}(n) \approx \frac{nh(0)}{4C_{\zeta}f_X(d^0)}. \tag{3.7}$$

It can be seen that the more "peaked" the sampling density g_U (equivalently h) is the greater the efficiency gains. However, one needs to be careful, since the above formula

is obtained through asymptotic considerations. In finite samples, a very peaked density around $\hat{d}_{n_1}$ may not perform well, since bias issues (involving $(d^0 - \hat{d}_{n_1})$) must also be taken into account.

3.2.1 Uniform Sampling Designs

Simulation results indicate that in the presence of a small budget of available points ($n = 20$ or 50), the efficiency of the two-stage estimator can be improved by employing a uniform (equispaced) design in the first stage, as the results in Table 3.1 attest. The results show that with an appropriate choice of the tuning parameters the gains in efficiency become very large, since the ratio of the MSE of the two-stage estimators to the one-stage estimators can take values up to 30. A uniform design renders the two-stage estimator competitive even for SNR=2 as the results in Table 3.2 show. The reason is that such a design reduces the sampling variability of the covariate x, which leads to improved localization of the change point. However, the approximate formula for the ARE discussed above is no longer valid when we use a uniform design at the first stage, since the asymptotics of the one stage estimator are then no longer described by the minimizer of a compound Poisson process. We elaborate on this point below.

Suppose that a uniform design on $[0, 1]$ is used in the first stage. The covariate x is then sampled at the points $x_i = i/(n_1 + 1),\ i = 1, \cdots, n_1$. Consider the case where $d^0 = 1/2$, since this is taken up later on in the simulation study. From the results established thus far and the remark following Theorem 3.2, it is clear that the asymptotic distribution of the 1st stage estimator will be determined by the limiting behavior of the count process $\mathcal{P}_n(s)$, where $\mathcal{P}_n(s)$ is once again the number of covariate values in the interval $(d^0, d^0 + s/n_1] \cup (d^0 + s/n_1, d^0]$. However, in the presence of a uniform sampling design, this becomes a *deterministic* process and is

given by

$$\lfloor \frac{1}{2}(n_1+1) + (1+\frac{1}{n})s \rfloor - \lfloor \frac{1}{2}(n_1+1)d^0 \rfloor \triangleq W_{n_1}(s).$$

From the last expression, it can easily be seen that the limit of $W_{n_1}(s)$ will be different along even and odd subsequences of n_1 for $d^0 = 1/2$ under consideration. As a side remark, notice that for irrational d^0 the limit may not even exist.

Consider $n_1 \to \infty$ along an even subsequence, so that $W_{n_1}(s) \to \lfloor 1/2 + |s| \rfloor$. Then, the form of the limit process whose minimizer determines the asymptotic distribution of the least squares estimator is given by $\mathbb{O}(s) = \mathbb{O}_1(s) - \mathbb{O}_2(s)$, with $\mathbb{O}_1(s) = (\sum_{0 \le i \le \lfloor 1/2+|s| \rfloor} V_i^+)1(s \ge 0)$ and $\mathbb{O}_2(s) = (\sum_{0 \le i \le \lfloor 1/2+|s| \rfloor} V_i^-)1(s \le 0)$, with the $V_i^{\pm}$ defined as before. The main difference is that the number of events up to to time s are deterministic and occur at regular intervals, and hence exhibit less variability than a Poisson process with the same number of events up to time s. Therefore, the asymptotics of $\hat{d}_{n_1}$ in this special case are described by the minimizer of a compound process, driven by a deterministic one.

3.2.2 Comparison with the One-stage estimator Subject to Prior Knowledge

We examine next the performance of the proposed procedure in the following setting. Suppose that one has some prior knowledge about the location of the change-point; this knowledge is represented by a sampling distribution for the covariate x which is more concentrated around d_0. We would like to study the performance of the two-stage procedure in the absence of such knowledge, which implies that a uniform sampling distribution is going to be employed at both stages. Notice that this setting favors the one-stage procedure. However, (3.5) indicates that the two-stage procedure more than compensates for the availability of prior knowledge. But in finite samples, the adaptive choices of design parameters may not perform well, since (3.5) is obtained through asymptotic considerations and the minimal sample sizes are required. Therefore, with small sample sizes, we select different design

γ	λ	N=50				
		k=.5	k=1	k=1.5	k=2	k=3
1/4	.2	*1.9352	0.6884	0.6594	0.6274	0.6261
	.4	*1.7263	0.5231	0.3361	0.3924	0.3862
	.5	*1.3221	0.3707	0.2178	0.2452	0.2326
	.6	0.9636	0.2476	0.1730	0.1805	0.1741
	.8	0.3121	0.0847	0.0544	0.0507	0.0531
1/3	.2	*2.7848	0.7998	0.6594	0.6274	0.6261
	.4	*2.8495	0.8647	0.3361	0.3924	0.3862
	.5	*2.1412	0.6665	0.2286	0.2452	0.2326
	.6	*1.7003	0.4369	0.1845	0.1805	0.1741
	.8	0.5758	0.1575	0.0699	0.0507	0.0531
1/2	.2	*6.2367	*1.7092	0.7342	0.6274	0.6261
	.4	*7.5633	*2.3393	0.7530	0.4858	0.3862
	.5	*6.1912	*1.7833	0.6641	0.4081	0.2326
	.6	*5.3154	*1.3570	0.5733	0.3427	0.1740
	.8	*1.9669	0.5390	0.2364	0.1260	0.0588
2/3	.2	*13.4249	*3.6592	*1.5467	0.8254	0.6261
	.4	*21.0001	*6.4081	*2.0211	*1.3194	0.5788
	.5	*18.7035	*4.9289	*2.1087	*1.3687	0.4843
	.6	*16.0174	*4.2101	*1.7890	*1.0597	0.4473
	.8	*6.7680	*1.8204	0.8253	0.4294	0.2010
3/4	.2	*13.7728	*5.4489	*2.3051	*1.2230	0.6215
	.4	*31.3548	*10.4603	*3.3751	*2.1788	0.9623
	.5	*21.4054	*8.8094	*3.6705	*2.3977	0.8453
	.6	*25.5204	*7.2652	*3.1655	*1.8754	0.7934
	.8	*12.1963	*3.3559	*1.5242	0.7999	0.3717

Table 3.1: Sampling via uniform design in the first stage, N=50, SNR=5

γ	λ	N=50				
		k=.5	k=1	k=1.5	k=2	k=3
1/4	.2	0.1552	0.1059	0.0957	0.1184	0.1116
	.4	0.6007	0.2351	0.2196	0.2166	0.1852
	.5	0.8389	0.2817	0.1910	0.2198	0.2399
	.6	0.6519	0.2220	0.1770	0.1705	0.1768
	.8	0.4559	0.1203	0.0790	0.0749	0.0800
1/3	.2	0.1569	0.1160	0.0973	0.1184	0.1116
	.4	0.5982	0.3153	0.2150	0.2137	0.1852
	.5	0.9995	0.4211	0.2005	0.2162	0.2399
	.6	0.9337	0.3756	0.1992	0.1715	0.1768
	.8	0.7523	0.2124	0.1026	0.0747	0.0800
1/2	.2	0.1508	0.1387	0.1076	0.1154	0.1116
	.4	0.5079	0.4473	0.3715	0.2427	0.1911
	.5	0.9101	0.7932	0.5004	0.3357	0.2424
	.6	*1.1269	0.8649	0.5109	0.2982	0.1793
	.8	*1.3147	0.6075	0.3158	0.1849	0.0878
2/3	.2	0.1385	0.1428	0.1222	0.1326	0.1099
	.4	0.4004	0.4580	0.5397	0.4606	0.2794
	.5	0.6985	0.8433	0.8556	0.6612	0.4669
	.6	*1.0275	*1.1844	*1.0386	0.7122	0.4231
	.8	*1.3845	*1.1224	0.7658	0.5589	0.2777
3/4	.2	0.1325	0.1428	0.1330	0.1493	0.1158
	.4	0.3637	0.4346	0.5421	0.5354	0.3884
	.5	0.6138	0.7652	0.9327	0.8624	0.6554
	.6	0.9454	*1.1844	*1.1636	0.9491	0.6542
	.8	*1.3008	*1.2972	*1.0317	0.8718	0.4782

Table 3.2: Sampling via uniform design in the first stage, N=50, SNR=2

γ	λ	N=50				
		k=.5	k=1	k=1.5	k=2	k=3
1/2	.2	0.2707	0.4548	0.1900	0.1642	0.1636
	.4	0.8135	0.5812	0.2419	0.1560	0.1171
	.5	*1.3408	0.4116	0.1928	0.1146	0.0722
	.6	*1.3442	0.3673	0.1556	0.0952	0.0474
	.8	0.4864	0.1547	0.0635	0.0358	0.0161
2/3	.2	0.1147	0.4438	0.4286	0.2351	0.1647
	.4	0.3273	0.9655	0.6890	0.4098	0.1840
	.5	0.7004	*1.0401	0.5894	0.3198	0.1335
	.6	0.9889	*1.0264	0.4644	0.3124	0.1281
	.8	0.9944	0.5137	0.2155	0.1275	0.0548

Table 3.3: N=50, SNR=5, Prior 1

parameters to see how the two-stage procedure works without prior knowledge.

We study this issue for the model given in (3.6) for SNR=5, $N = 50, 100$, $\gamma = 1/2, 2/3$ and λ, K as above. The following two choices for $f_X(x)$ were used: (i) a triangular density on $(0, 1)$ centered at $d_0 = 0.5$ denoted as Prior 1 and (ii) a step function distribution denoted as Prior 2. The densities of these distributions are shown in Figure 3.5 and the results are summarized in the following Tables (Table 3.3 - Table 3.6).

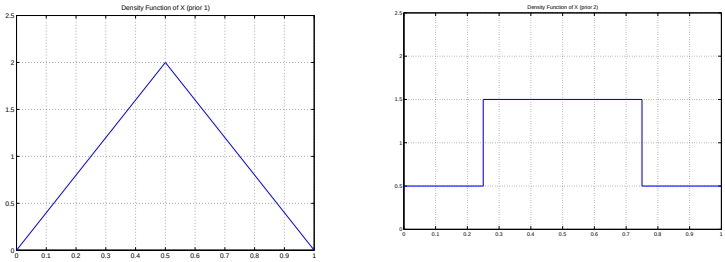


Figure 3.5: Two prior densities

γ	λ	N=100				
		k=.5	k=1	k=1.5	k=2	k=3
1/2	.2	0.3664	0.6819	0.3572	0.2077	0.1587
	.4	*2.7986	*1.0409	0.4252	0.2294	0.1040
	.5	*3.1284	0.7923	0.3627	0.1829	0.0881
	.6	*2.9556	0.7184	0.3110	0.1863	0.0775
	.8	*1.0112	0.2985	0.1113	0.0631	0.0268
2/3	.2	0.1438	0.5454	0.7504	0.5408	0.2553
	.4	0.8752	*2.9140	*1.4720	0.8098	0.4004
	.5	*2.8514	*2.7296	*1.2498	0.7028	0.3324
	.6	*2.4060	*2.5967	*1.3568	0.7241	0.3000
	.8	*2.1849	*1.2403	0.4653	0.2481	0.1104

Table 3.4: N=100, SNR=5, Prior 1

γ	λ	N=50				
		k=.5	k=1	k=1.5	k=2	k=3
1/2	.2	0.4414	0.7416	0.3099	0.2677	0.2668
	.4	*1.3206	0.9434	0.3926	0.2532	0.1900
	.5	*2.2034	0.6765	0.3169	0.1883	0.1187
	.6	*2.1956	0.5999	0.2542	0.1556	0.0774
	.8	0.8001	0.2544	0.1044	0.0589	0.0265
2/3	.2	0.1870	0.7237	0.6989	0.3834	0.2686
	.4	0.5313	*1.5673	*1.1184	0.6653	0.2986
	.5	*1.1510	*1.7093	0.9687	0.5255	0.2194
	.6	*1.6153	*1.6766	0.7586	0.5103	0.2093
	.8	*1.6359	0.8451	0.3545	0.2098	0.0902

Table 3.5: N=50, SNR=5, Prior 2

γ	λ	N=100				
		k=.5	k=1	k=1.5	k=2	k=3
1/2	.2	0.6278	*1.1683	0.6120	0.3558	0.2720
	.4	*4.7805	*1.7781	0.7263	0.3919	0.1777
	.5	*5.3403	*1.3525	0.6191	0.3122	0.1504
	.6	*5.0359	*1.2240	0.5299	0.3174	0.1321
	.8	*1.7230	0.5086	0.1896	0.1074	0.0457
2/3	.2	0.2464	0.9343	*1.2856	0.9265	0.4373
	.4	*1.4950	*4.9776	*2.5144	*1.3832	0.6839
	.5	*4.8675	*4.6596	*2.1335	*1.1998	0.5674
	.6	*4.0996	*4.4245	*2.3119	*1.2338	0.5111
	.8	*3.7228	*2.1133	0.7929	0.4227	0.1881

Table 3.6: N=100, SNR=5, Prior 2

It can be seen that even for the first prior distribution which is fairly concentrated around the change point, and for a small total budget of points ($N = 50$) with an appropriate selection of the tuning parameters the two-stage procedure outperforms the classical one; further, it proves competitive (the ratio of MSE of two-stage estimators to the one-stage estimators with prior knowledge is close to 1) for a number of other configurations. It should be noted that a small value for K, λ about 0.5 and $\gamma \geq 1/2$ are the values of the tuning parameters achieving this result. For $N = 100$, it outperforms for a larger number of configurations (in particular for $K = 1$) of the tuning parameters, while the gains in efficiency become more substantial (the ratio of MSE of two-stage estimators to the one-stage estimators with prior knowledge reaches the value 3). For the second flatter prior distribution, the results are more favorable for the two-stage procedure, since for a large number of configurations it outperforms the one-stage procedure and in the case of $N = 100$ by a fairly wide

margin in certain instances (the ratio of MSE of two-stage estimators to the one-stage estimators with prior knowledge exceeds 4 for several configurations). Obviously, the results become even more advantageous for larger budgets of points and less advantageous for more concentrated distributions around d^0 in the presence of moderate sample sizes.

3.3 Confidence Intervals for the Change-Point

We compare next the performance of exact confidence intervals based on the result established in Theorem 3.2, to those proposed in Ferger (2004). Moreover, confidence intervals for finite samples will be constructed following the discussion in Section 3.2.

For all these comparisons, simulations were run for a stump model with $\alpha_0 = 0.5,\ \beta_0 = 1.5,\ d^0 = 0.5$ and sample sizes $n = 50, 100, 200, 500, 1000$ with $N = 2000$ replicates for each n. Confidence intervals for d^0 based on the minimal minimizer $\hat{d}_{n_2,l}$, the maximal minimizer $\hat{d}_{n_2,u}$, and the average minimizer $\hat{d}_{n_2,av} = (\hat{d}_{n_2,l}+\hat{d}_{n_2,u})/2$ were constructed. Two values of $\gamma = 1/2$ and $2/3$ and two values of $K = 1$ and 2 were used together with the optimal allocation $\lambda \equiv \gamma/(1+\gamma)$ as discussed in Section 3.2. The confidence level was set at $1 - q = .95$ and the percentage of replicates for which the true change-point was included in the corresponding intervals, as well as the average length of each interval, were recorded. In what follows, the symbols d_l and d_u have the same connotations as in the proof of Theorem 3.2.

3.3.1 Conservative Intervals

Using the results of Ferger (2004), based on any two–stage estimator $\hat{d}_{n_2}$, we propose an asymptotically conservative confidence interval for d^0 at level $1 - q$:

$$I_n(q) := (\hat{d}_n - b/n_2^{1+\gamma}, \hat{d}_n - a/n_2^{1+\gamma})$$

where $a < b$ are any solutions of the inequality

$$\text{Prob}(d_u < b) - \text{Prob}(d_l \leq a) \geq 1 - q\,.$$

Based on the smallest, largest and average minimizers at Stage 2, we therefore obtain intervals

$$I_{n_2,l}(q) = (\hat{d}_{n_2,l} - b/n_2^{1+\gamma}, \hat{d}_{n_2,l} - a/n_2^{1+\gamma}),$$

$$I_{n_2,u}(q) = (\hat{d}_{n_2,u} - b/n_2^{1+\gamma}, \hat{d}_{n_2,u} - a/n_2^{1+\gamma}),$$

and

$$I_{n_2,av}(q) = (\hat{d}_{n_2,av} - b/n_2^{1+\gamma}, \hat{d}_{n_2,av} - a/n_2^{1+\gamma})$$

where a is the $q/2^{th}$ quantile of d_l and b is the $(1-q/2)^{th}$ quantile of d_u. These quantiles do not seem to be analytically determinable but can certainly be simulated to a reasonable degree of approximation.

In Table 3.7the coverage probabilities together with the length of the confidence intervals are shown for a number of combinations of sample sizes and tuning parameters and with the SNR set equal to 5. It can be seen that the recorded coverage

	n=50		n=100		n=200		n=500		n=1000	
$\gamma = \frac{1}{2}$	K=1	K=2	K=1	K=2	K=1	K=2	K=1	K=2	K=1	K=2
$\hat{I}_{n_2,l}$	97.40%	97.55%	97.40%	98.20%	97.65%	98.00%	96.55%	97.05%	97.15%	97.75%
	(.0580)	(.1208)	(.0205)	(.0427)	(.0072)	(.0151)	(.0018)	(.0038)	(.0006)	(.0014)
$\hat{I}_{n_2,u}$	97.05%	98.60%	97.65%	97.05%	97.65%	97.85%	97.40%	97.90%	97.80%	98.00%
	(.0580)	(.1208)	(.0205)	(.0427)	(.0072)	(.0151)	(.0018)	(.0038)	(.0006)	(.0014)
$\hat{I}_{n_2,av}$	99.80%	99.95%	99.80%	99.95%	100%	100%	99.80%	100%	99.70%	99.95%
	(.0580)	(.1208)	(.0205)	(.0427)	(.0072)	(.0151)	(.0018)	(.0038)	(.0006)	(.0014)
$\gamma = \frac{2}{3}$	n=50		n=100		n=200		n=500		n=1000	
$\hat{I}_{n_2,l}$	98.15%	97.70%	97.65%	98.30%	97.90%	97.70%	97.65%	97.30%	97.75%	97.70%
	(.0299)	(.0581)	(.0094)	(.0183)	(.0030)	(.0058)	(.0006)	(.0013)	(.0002)	(.0004)
$\hat{I}_{n_2,u}$	98.00%	98.20%	97.85%	98.55%	97.90%	98.10%	98.30%	97.75%	97.60%	98.50%
	(.0299)	(.0581)	(.0094)	(.0183)	(.0030)	(.0058)	(.0006)	(.0013)	(.0002)	(.0004)
$\hat{I}_{n_2,av}$	99.60%	99.95%	99.60%	99.90%	99.85%	99.95%	99.85%	99.90%	99.85%	99.95%
	(.0299)	(.0581)	(.0094)	(.0183)	(.0030)	(.0058)	(.0006)	(.0013)	(.0002)	(.0004)

Table 3.7: 95% Conservative Confidence Intervals for a combination of sample sizes and the tuning parameters γ, K and for SNR=5

exceeds the nominal level of 95% and almost approaching perfect (100%) coverage for the average minimizer.

3.3.2 Exact Confidence Intervals

On the other hand, since Theorem 3.2 provides us with the *exact* asymptotic distributions of the sample minimizers, we can construct asymptotically exact (level

$1-q$ confidence intervals) as follows:

$$\tilde{I}_{n_2,l}(q) = (\hat{d}_{n_2,l} - b_l/n_2^{1+\gamma}, \hat{d}_{n_2,l} - a_l/n_2^{1+\gamma}),$$

$$\tilde{I}_{n_2,u}(q) = (\hat{d}_{n_2,u} - b_u/n_2^{1+\gamma}, \hat{d}_{n_2,u} - a_u/n_2^{1+\gamma}),$$

$$\tilde{I}_{n_2,av}(q) = (\hat{d}_{n_2,av} - b_{av}/n_2^{1+\gamma}, \hat{d}_{n_2,av} - a_{av}/n_2^{1+\gamma})$$

where $a_l, b_l, a_u, b_u, a_{av}$ and b_{av} are the exact quantiles (a_l, a_u and a_{av} correspond to $q/2^{th}$ quantiles and b_l, b_u and b_{av} correspond to $(1-q/2)^{th}$ quantiles) of d_l, d_u and $(d_l + d_u)/2$, respectively.

In Table 3.8, the coverage probabilities together with the length of the confidence intervals are shown for a number of combinations of sample sizes and tuning parameters and with the SNR set equal to 5. It can be see that the coverage probabilities

	n=50		n=100		n=200		n=500		n=1000	
$\gamma = \frac{1}{2}$	K=1	K=2	K=1	K=2	K=1	K=2	K=1	K=2	K=1	K=2
$\tilde{I}_{n_2,l}$	95.00%	94.20%	93.80%	95.25%	95.25%	94.10%	93.40%	94.50%	95.05%	94.90%
	(.0283)	(.0599)	(.0100)	(.0212)	(.0035)	(.0075)	(.0009)	(.0019)	(.0003)	(.0007)
$\tilde{I}_{n_2,u}$	94.20%	96.50%	94.80%	94.95%	94.45%	95.85%	94.85%	95.90%	95.50%	95.85%
	(.0294)	(.0602)	(.0104)	(.0213)	(.0037)	(.0075)	(.0009)	(.0019)	(.0003)	(.0007)
$\tilde{I}_{n_2,av}$	94.30%	95.25%	94.35%	94.05%	94.40%	95.45%	93.20%	94.85%	94.55%	95.30%
	(.0236)	(.0487)	(.0083)	(.0172)	(.0029)	(.0061)	(.0007)	(.0015)	(.0003)	(.0005)
$\gamma = \frac{2}{3}$	n=50		n=100		n=200		n=500		n=1000	
$\tilde{I}_{n_2,l}$	95.65%	95.40%	95.05%	96.05%	95.35%	95.45%	95.30%	95.30%	95.05%	95.90%
	(.0148)	(.0277)	(.0047)	(.0087)	(.0015)	(.0027)	(.0003)	(.0006)	(.0001)	(.0002)
$\tilde{I}_{n_2,u}$	95.40%	96.00%	95.65%	96.60%	95.80%	96.15%	96.20%	96.60%	95.15%	96.85%
	(.0149)	(.0302)	(.0047)	(.0095)	(.0015)	(.0030)	(.0003)	(.0006)	(.0001)	(.0002)
$\tilde{I}_{n_2,av}$	95.15%	96.45%	95.20%	96.95%	94.75%	96.10%	95.30%	96.15%	94.05%	96.10%
	(.0120)	(.0253)	(.0038)	(.0080)	(.0012)	(.0025)	(.0003)	(.0005)	(.0001)	(.0002)

Table 3.8: 95% Exact Confidence Intervals for a combination of sample sizes and tuning parameters γ, K for SNR=5

are fairly close to their nominal values, especially for $\gamma = 2/3$. Further, their length is almost half of those obtained according to Ferger's (2004) method. Finally, it should be noted that analogous results were obtained for SNR=2 and 8 (not shown due to space considerations).

3.3.3 Construction of Confidence Intervals Based on Adaptive Choices of Design Parameters

Confidence intervals in finite samples can also be based on the adaptive parameter allocation strategies discussed in Section 3.2. We briefly discuss this below, adopting notation from that Section.

With tuning parameters K and γ and the optimal allocation of $\lambda \equiv \gamma/(1+\gamma)$, a level $1-q$ confidence interval for d^0 is given by

$$\left[\widehat{d_{2,av}} - n^{-(1+\gamma)}\, C_{q/2}\, \frac{2K}{\Psi(\gamma)}, \widehat{d_{2,av}} + n^{-(1+\gamma)}\, C_{q/2}\, \frac{2K}{\Psi(\gamma)}\right].$$

With the adaptive choices, $K \equiv K_{(n)} = C_\zeta\, n_1^{-(1-\gamma)}$, where $q/2 >> \zeta$ and $n_1 = n\,\gamma/(1+\gamma)$, the above confidence interval reduces to:

$$\left[\widehat{d_{2,av}} - \frac{2\,C_\zeta\, C_{q/2}}{n^2}\, \frac{(1+\gamma)^2}{\gamma},\ \ \widehat{d_{2,av}} + \frac{2\,C_\zeta\, C_{q/2}}{n^2}\, \frac{(1+\gamma)^2}{\gamma}\right].$$

To minimize the length, we let γ tend to 1, to obtain an approximate level $1-q$ confidence interval as follows:

$$\left[\widehat{d_{2,av}} - \frac{8\,C_\zeta\, C_{q/2}}{n^2},\ \ \widehat{d_{2,av}} + \frac{8\,C_\zeta\, C_{q/2}}{n^2}\right].$$

Simulations were run for the above stump model for four different sample sizes: 50, 100, 200, and 500 with 5000 replicates for each sample size and for 3 different values of SNR=2, 5, 8. Confidence intervals as defined above were constructed (with $q = 0.05$). The percentage of intervals containing the true change-point together with their length were recorded and shown in Table 3.9.

	SNR=2		SNR=5		SNR=8	
N	Coverage	Length	Coverage	Length	Coverage	Length
50	93.24%	0.2780	95.48%	0.0383	95.68%	0.0329
100	94.08%	0.0695	95.24%	0.0096	95.54%	0.0082
200	94.48%	0.0174	94.78%	0.0024	95.16%	0.0021
500	94.82%	0.0028	95.08%	0.00038	94.94%	0.00033

Table 3.9: 95% Confidence Intervals constructed using the adaptive parameter allocation strategy for different sample sizes and SNR with $\zeta = 0.0005$

We examine next the performance of confidence intervals in finite samples, but where a uniform (equispaced) design is used in the first stage (results shown in Table 3.10) and in both stages (results shown in Table 3.11). The setting is identical to that used in Table 3.9. As the discussion in Section 3.2.1 indicates, it is not obvious how the tuning parameters C_ζ should be chosen in this case; therefore, the same C_ζ value as the one used in Table 3.9 was employed. It can be seen that a uniform design used in the 1st stage does not improve performance in terms of coverage or length. However, using a uniform design in both stages and setting C_ζ and C_q to the same values as in Table 3.9 leads to rather conservative confidence intervals, especially for larger sample sizes and higher values of SNR. Notice that the lengths of the confidence intervals are identical to those in Table 3.9 due to the choice of the tuning parameters C_ζ and C_q. Nevertheless, experience shows that a uniform design used in the 1st stage gives better mean squared errors in small samples, or when d^0 is closer to the boundary of the covariate's support.

	SNR=2		SNR=5		SNR=8	
N	Coverage	Length	Coverage	Length	Coverage	Length
50	93.72%	0.2780	95.14%	0.0383	95.56%	0.0329
100	93.88%	0.0695	95.12%	0.0096	95.20%	0.0082
200	94.62%	0.0174	95.52%	0.0024	95.52%	0.0021
500	94.72%	0.0028	94.96%	0.00038	95.12%	0.00033

Table 3.10: 95% Confidence Interval constructed using the adaptive parameter allocation strategy for different sample sizes and SNR with $\zeta = 0.0005$ using a uniform design in the 1st Stage

	SNR=2		SNR=5		SNR=8	
N	Coverage	Length	Coverage	Length	Coverage	Length
50	95.06%	0.2780	100.00%	0.0383	100.00%	0.0329
100	96.66%	0.0695	99.98%	0.0096	100.00%	0.0082
200	96.94%	0.0174	100.00%	0.0024	100.00%	0.0021
500	97.32%	0.0028	99.96%	0.00038	100.00%	0.00033

Table 3.11: 95% Confidence Interval constructed using the adaptive parameter allocation strategy for different sample sizes and SNR with $\zeta = 0.0005$ using a uniform design in both stages

3.3.4 Data application

We revisit the motivating application and estimate the change-point using both the "classical" and the developed two-stage procedures. The total budget was set to $n = 70$ and the model fitted to the natural logarithm of the delays comprised two linear segments with a discontinuity. Given that the data (134 system loadings and their corresponding average delays) have been collected in advance, a sampling mechanism close in spirit to selecting covariate values from a uniform distribution was employed for both procedures. Specifically, the necessary number of points was drawn from a uniform distribution in the $[0, 1]$ interval and amongst the available 134 loadings the ones closest to the sampled points were selected, together with their corresponding responses. An analogous strategy was used when a uniform design was employed in the 1st stage of the adaptive procedure. For the two-stage procedure, we set $\lambda = 1/2$ and the remaining tuning parameters to those values provided by the adaptive strategy discussed in Section 4, with $\zeta = .0005$. The results of the "classical" procedure, the two-stage adaptive procedure with sampling from a uniform distribution in both stages and from a uniform design in the 1st stage and the uniform distribution in the 2nd stage are depicted in the left, center and right panels of Figure 3.6, respectively. The depicted fitted regression models are based on the first stage estimates for the two-stage procedure. Further, the sampled points from the two stages are shown as solid (1st stage) and open (2nd stage) circles. It can be seen that the heavier sampling in the neighborhood of the 1st stage estimate of the change-point improves the estimate given the available evidence from all 134 points shown in Figure 1.1.

The estimated change-point from the "classical" procedure is $\hat{d}_n = .737$ with a 95% confidence interval $(.682, .793)$. Using a uniform distribution in both stages gave an estimate $\hat{d}_{n_2} = .796$ with a 95% confidence interval $(.781, .811)$. On the other hand, a

combination of a uniform design in the 1st stage with that of a uniform distribution in the 2nd stage yielded an estimate $\hat{d}_{n_2} = .802$ with a 95% confidence interval $(.787, .817)$. As shown in this case and validated through other data examples, the use of uniform design in the first stage proves advantageous in practice, especially for small samples or in situations where the discontinuity lies fairly close to the boundary of the design region.

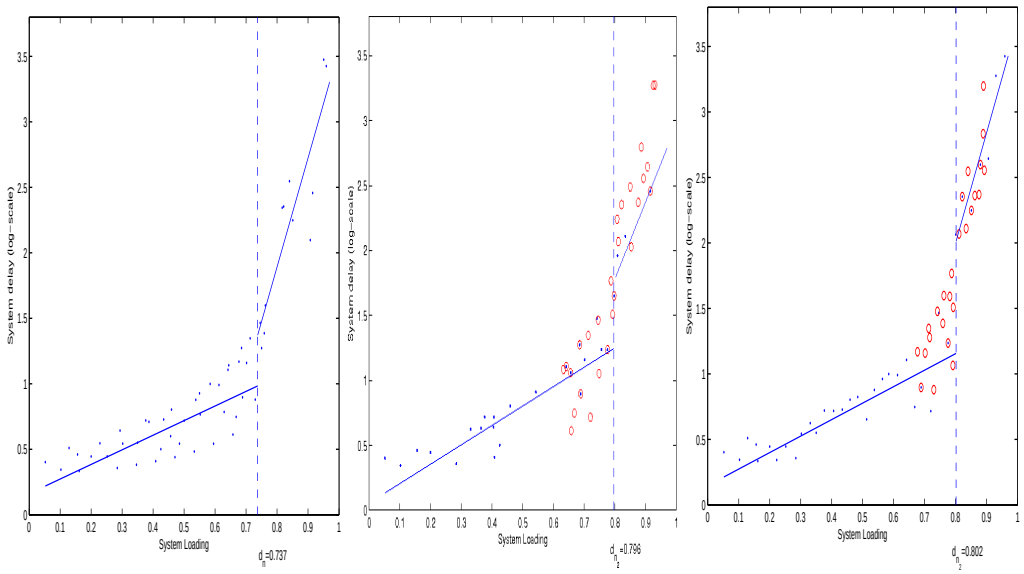


Figure 3.6: Sampled points (from 1st stage solid circles and from 2nd stage open circles) together with the fitted parametric models and estimated change point, based on a total budget of $n = 70$ points, obtained from the "classical" procedure (left panel), the two-stage adaptive procedure with sampling from a uniform distribution in both stages (center panel) and from a uniform design in the 1st stage and the uniform distribution in the 2nd stage (right panel).

3.4 Extension to A Three–stage Procedure.

As we discussed before, the two-stage procedure can be generalized to multi-stage procedures. The number of stages depends on the budget. We tried some simulation examples for a three-stage procedure and compared its performance with the classical procedure as well as the two-stage procedure using the ratio of MSE. Recall that $n_3^{2+\gamma_2}((\hat{d}_{n_3,l} - d^0), (\hat{d}_{n_3,u} - d^0))$ is $O_p(1)$ and converges in distribution to (d_l, d_u), where (d_l, d_u) is the vector of the smallest and the largest argmins of the

process $\mathbb{M}(\mid \psi_l(\beta_l, d^0) - \psi_u(\beta_u, d^0) \mid, \epsilon, C_3)$, with $C_3 = (2K_2)^{-1}(\lambda_2/\lambda_3)^{(1+\gamma_2)}$. For the three stage procedure, we take $\lambda_1 = \lambda_2 = \lambda_3 = 1/3$, $\gamma_1 = 3/5$ and $\gamma_2 = 2/5$. In the two stage procedure, we select two values for γ: 2/3 and 3/4. To compare these three procedures, the selections of K_1, K_2 and K guarantee that the sampling neighborhood in each stage has 99% coverage of the true change-point. In this example, $K_1 = 1.5$, $K_2 = 8$, $K = 1$ (when $\gamma = 2/3$) and $K = 2$ (when $\gamma = 3/4$) . We select different total sample sizes $n = 90,\ 150,\ 300$ and $n = 900$ as shown in Table 3.12, where

$$
\begin{aligned}
R_1 &= \text{MSE(one-stage)/MSE(three-stage)} \\
R_2 &= \text{MSE(two-stage with}\,\gamma = 2/3, K = 1\text{)/MSE(three-stage)} \\
R_3 &= \text{MSE(two-stage with}\,\gamma = 3/4, K = 2\text{)/MSE(three-stage)}
\end{aligned}
$$

We run simulations with 1000 replicates for each setting. The estimates from the three-stage procedure are the best ones regarding the ratios of MSE and the advantage becomes more obvious as n increases. When the sample size is small, the three-stage procedure is comparable with the two-stage procedure under some setting of design parameters. For example, we select $\gamma = 3/4$ and $K = 2$ for a two-stage procedure, and the ratio of MSE is 1.47 with $n = 90$.

n	R_1	R_2	R_3
90	2.4711	2.1939	1.4736
150	6.4581	3.2832	2.0779
300	29.245	12.902	7.7130
900	400.10	20.128	7.6691

Table 3.12: Performance of Three-stage Procedure

CHAPTER 4

Adaptive Strategies for Estimating the Regression Function

As briefly introduced in Chapter 1, we will focus on developing adaptive strategies for estimating the entire regression function in this chapter.

4.1 Optimal Allocation of Samples

Consider, for example, the general parametric model,

$$Y_i = \mu(X_i) + \epsilon_i, \quad i = 1, 2, \ldots, n$$

where $\mu(x) = \psi_l(\beta_l, x)I(x \leq d^0) + \psi_u(\beta_u, x)I(x > d^0)$, which has the same definition as in (2.1). X_i's are i.i.d. and are distributed on according to some density $f_X(x)$ on $[0, 1]$. Recall that for the multi-stage procedures of Chapter 3, we select covariate values via uniform distribution or a uniform design on $[0, 1]$. To improve the entire regression function, we use an initial fraction of the budget to estimate the parameters of the underlying model. Subsequently, the design region is partitioned into three segments, with the middle segment defined as a fixed neighborhood around the estimated change-point. A mixture of uniform densities is selected as the design density on these segments. We use λ_2, λ_1 and λ_3 to denote the uniform density on the left segment, the middle segment and the right segment, respectively.

$$f_X(x) = \begin{cases} \lambda_1 & x \in \text{ left segment} \\ \lambda_2 & x \in \text{ middle segment} \\ \lambda_3 & x \in \text{right segment} \end{cases} .$$

The objective then becomes to allocate the available samples to these three segments so as to minimize an asymptotic expected L_2 error that depends on model parameters. The estimates from the initial fraction act as surrogates. Therefore, the way of allocating the samples is related to the design density. We start with a toy problem by the piecewise-constant model, for which we assume the change point d^0 is known.

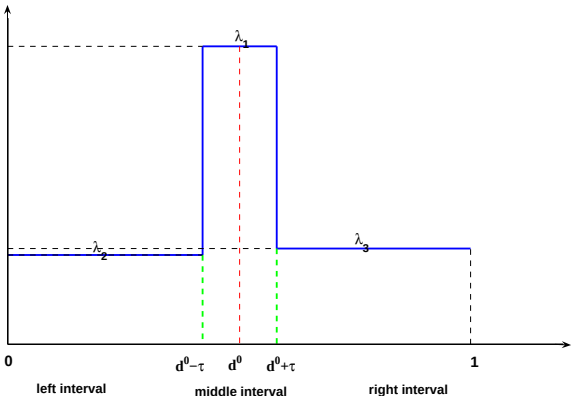


Figure 4.1: Mixture Uniform Density for Sampling

We define an appropriate interval $[d^0 - \tau, d^0 + \tau]$around the true change point. What we want is to decide how to allocate samples in the left, middle and the right intervals as shown in Figure 4.1. The height corresponds to the density in each interval, say λ_2, λ_1, and λ_3.

We adhere to the average minimizer and homoscedastic normal errors for all discussions in this chapter. By (2.2) and (2.3), we have $d_{av}(SNR, Z, 1) = (d_l(SNR, Z, 1)+ d_u(SNR, Z, 1))/2$, where Z is a standard normal random variable and the quantity $\mid \psi_l(\beta_l, d^0) - \psi_u(\beta_u, d^0) \mid /\sigma$ is denoted as SNR (signal-to-noise ratio). We define the random variable $V = \mid d_{av}(SNR, Z, 1) \mid$.

4.1.1 Piecewise–Constant Model

In this simple step function model, $\psi_l(\beta_l, x) = \alpha_0$, and $\psi_u(\beta_u, x) = \beta_0$. Let the total sample size be n and consider a L_2 error as

$$
\begin{aligned}
& L_2 \\
= & \int [[\hat{\alpha} I\{x \le \hat{d}_n\} + \hat{\beta} I\{x > \hat{d}_n\} - \alpha_0 I\{x \le d^0\} - \beta_0 I\{x > d^0\}]^2] dx \\
= & \int [I\{\hat{d}_n < d^0\}[(\hat{\alpha} - \alpha_0)^2 \hat{d}_n + (\hat{\beta} - \beta_0)^2 (1 - d^0) + (\hat{\beta} - \alpha_0)^2 (d^0 - \hat{d}_n)] \\
& + I\{\hat{d}_n > d^0\}[(\hat{\alpha} - \alpha_0)^2 d_0 + (\hat{\beta} - \beta_0)^2 (1 - \hat{d}_n) + (\hat{\beta} - \alpha_0)^2 (\hat{d}_n - d^0)]] dx \\
= & \frac{\sigma^2}{n} \left[I\{\hat{d}_n < d^0\} \left[\frac{n(\hat{\alpha} - \alpha_0)^2}{\sigma^2} \hat{d}_n + \frac{n(\hat{\beta} - \beta_0)^2}{\sigma^2} (1 - d^0) + (\frac{\hat{\beta} - \alpha_0}{\sigma})^2 n(d^0 - \hat{d}_n) \right] \right. \\
& \left. + I\{\hat{d}_n > d^0\} \left[\frac{n(\hat{\alpha} - \alpha_0)^2}{\sigma^2} d^0 + \frac{n(\hat{\beta} - \beta_0)^2}{\sigma^2} (1 - \hat{d}_n) + (\frac{\beta_0 - \hat{\alpha}}{\sigma})^2 n(\hat{d}_n - d^0) \right] \right] \\
= & \frac{\sigma^2}{n} \left[I\{\hat{d}_n < d^0\} \left[\frac{n(\hat{\alpha} - \alpha_0)^2}{\sigma^2} d^0 + \frac{n(\hat{\beta} - \beta_0)^2}{\sigma^2} (1 - d^0) + (\frac{\hat{\beta} - \alpha_0}{\sigma})^2 n(d^0 - \hat{d}_n) \right. \right. \\
& \qquad \left. + \frac{n(\hat{\alpha} - \alpha_0)^2}{\sigma^2} (\hat{d}_n - d^0) \right] \\
& \quad + I\{\hat{d}_n > d^0\} \left[\frac{n(\hat{\alpha} - \alpha_0)^2}{\sigma^2} d^0 + \frac{n(\hat{\beta} - \beta_0)^2}{\sigma^2} (1 - d^0) + (\frac{\beta_0 - \hat{\alpha}}{\sigma})^2 n(\hat{d}_n - d^0) \right. \\
& \qquad \left. \left. + \frac{n(\hat{\beta} - \beta_0)^2}{\sigma^2} (d^0 - \hat{d}_n) \right] \right]
\end{aligned}
$$

From the discussions in Chapter 2, we know that $\hat{\alpha}$, $\hat{\beta}$ and $\hat{d}_n$ are asymptotically independent, $\sqrt{n}(\hat{\alpha} - \alpha_0)$ and $\sqrt{n}(\hat{\beta} - \beta_0)$ are mean zero Gaussian with respective variances $\sigma^2 / F_X(d^0)$ and $\sigma^2/(1 - F_X(d^0))$, and $n(\hat{d}_n - d^0)$ converges weakly to the average minimizer of $\mathbb{M}_{|\alpha_0 - \beta_0|, \epsilon_1, f_X(d^0)}$. Therefore, $\frac{n(\hat{\alpha} - \alpha_0)^2}{\sigma^2}(\hat{d}_n - d^0)$, $\frac{n(\hat{\beta} - \beta_0)^2}{\sigma^2}(d^0 - \hat{d}_n)$, $(\frac{\hat{\beta} - \alpha_0}{\sigma})^2 n(d^0 - \hat{d}_n)$ and $(\frac{\beta_0 - \hat{\alpha}}{\sigma})^2 n(\hat{d}_n - d^0)$ in the expression above have the rate of convergence $1/n$, then we define the asymptotic expected L_2 error as

$$
\begin{aligned}
EL_2 & \triangleq \frac{\sigma^2}{n} E \left[\frac{n(\hat{\alpha} - \alpha_0)^2}{\sigma^2} d^0 + \frac{n(\hat{\beta} - \beta_0)^2}{\sigma^2} (1 - d^0) + (\frac{\beta_0 - \alpha_0}{\sigma})^2 n |\hat{d}_n - d^0| \right] \\
& = \frac{\sigma^2}{n} \left[\frac{d^0}{F_X(d^0)} + \frac{1 - d^0}{1 - F_X(d^0)} + (SNR)^2 \frac{1}{f_X(d^0)} E(V) \right]
\end{aligned}
$$

where the last component is obtained by the scaling relationship between the minimizer of $\mathbb{M}_{|\alpha_0-\beta_0|,\epsilon_1,f_X(d^0)}$ and the minimizer of $\mathbb{M}_{SNR,Z,1}$.

Set $L = d^0 - \tau$, $U = d^0 + \tau$. We use a mixture of uniform densities as a design density for X on $[0,1]$:

$$f_X(x) = \begin{cases} \lambda_1 & x \in [0, L) \\ \lambda_2 & x \in [L, U] \\ \lambda_3 & x \in (U, 1) \end{cases} .$$

However, in practice, we do not know d^0. We use a fraction,say $1/2$, of the sample size to obtain an estimate of the change point , $\hat{d}_1$ and an estimate of SNR. We substitute $\hat{d}_1$ for d^0 and take $\tau = C_\zeta/n \triangleq c_n$ (C_ζ is the upper ζ^{th} quantile of the Argmin $\mathbb{M}_{SNR,Z,1}$ for a very small fraction ζ) to obtain the following surrogate expression:

$$\begin{aligned} EL_2 &= \frac{\sigma^2}{n}\left[\frac{\hat{d}_1}{F_X(\hat{d}_1)} + \frac{1-\hat{d}_1}{1-F_X(\hat{d}_1)} + \frac{\hat{B}}{\lambda_1}\right] \\ &= \frac{\sigma^2}{n}\left[\frac{\hat{d}_1}{\lambda_2 L_n + \lambda_1 c_n} + \frac{1-\hat{d}_1}{1-(\lambda_2 L_n + \lambda_1 c_n)} + \frac{\hat{B}}{\lambda_1}\right] \end{aligned}$$

where $L_n = \hat{d}^1 - c_n$ and $\hat{B} = \widehat{(SNR)}^2\widehat{E(V)}$.

Our objective is to minimize EL_2 above under the following constraints:

$$2\lambda_1 c_n + \lambda_2 L_n + \lambda_3(1-U_n) = 1, \lambda_1 > 0, \lambda_2 \geq 0, \lambda_3 \geq 0$$

where $U_n = \hat{d}_1 + c_n$.

To simplify notation, we define $\lambda^* = \lambda_1 c_n$ and $D = \lambda_1 c_n + \lambda_2 L_n$. We seek to minimize

$$g(D, \lambda^*) = \frac{\hat{d}_1}{D} + \frac{1-\hat{d}_1}{1-D} + \frac{\hat{B}c_n}{\lambda^*} \tag{4.1}$$

such that

$$\lambda^* > 0, D \geq \lambda^*, D + \lambda^* \leq 1.$$

The optimal solution is lies the shaded part in Figure 4.2.

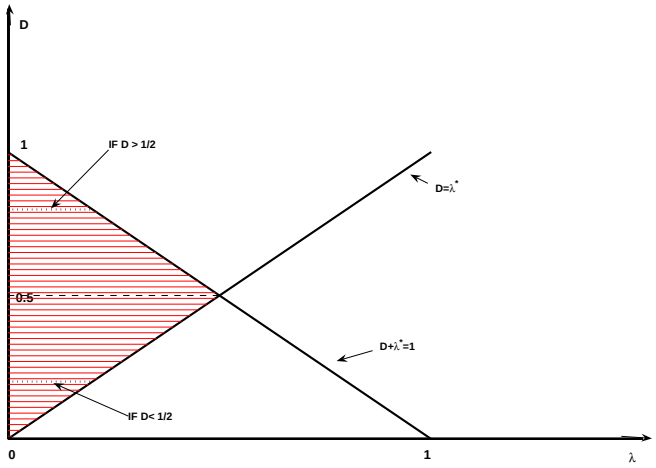


Figure 4.2: Optimal problem for the piecewise-constant model.

Theorem 4.1. *Define*

$$\begin{aligned}
\tilde{g}_1(D) &= \frac{\hat{d}_1}{D} + \frac{1-\hat{d}_1}{1-D} + \frac{\hat{B}c_n}{D} \\
\tilde{g}_2(D) &= \frac{\hat{d}_1}{D} + \frac{1-\hat{d}_1}{1-D} + \frac{\hat{B}c_n}{1-D} \\
D_1^* &= \frac{\sqrt{\frac{\hat{d}_1+\hat{B}c_n}{1-\hat{d}_1}}}{1+\sqrt{\frac{\hat{d}_1+\hat{B}c_n}{1-\hat{d}_1}}} \\
D_2^* &= \frac{\sqrt{\frac{\hat{d}_1}{1-\hat{d}_1+\hat{B}c_n}}}{1+\sqrt{\frac{\hat{d}_1}{1-\hat{d}_1+\hat{B}c_n}}}.
\end{aligned}$$

The optimal solutions to (4.1) are:

1. *When* $\hat{d}_n^{(1)} \le \frac{1-Bc_n}{2}$,

   If $\tilde{g}_1(D_1^*) < \tilde{g}_2(1/2)$, *then* $D_{opt}^* = D_1^*$ *and* $\lambda_{opt}^* = D_1^*$; *otherwise,* $D_{opt}^* = 1/2$ *and* $\lambda_{opt}^* = 1/2$.

2. *When* $\hat{d}_n^{(1)} \ge \frac{1+Bc_n}{2}$,

   If $\tilde{g}_2(D_2^*) < \tilde{g}_1(1/2)$, *then* $D_{opt}^* = D_2^*$ *and* $\lambda_{opt}^* = 1 - D_2^*$; *otherwise,* $D_{opt}^* = 1/2$ *and* $\lambda_{opt}^* = 1/2$.

3. *When* $\frac{1-Bc_n}{2} < \hat{d}_n^{(1)} < \frac{1+Bc_n}{2}$,

   $D_{opt}^* = 1/2$ *and* $\lambda_{opt}^* = 1/2$.

Proof of Theorem 4.1: For a fixed $D \in [0,1]$, define $g_D(\lambda^*) = \frac{\hat{d}_1}{D} + \frac{1-\hat{d}_1}{1-D} + \frac{\hat{B}c_n}{\lambda^*}$ and find the optimal value of λ^*, denoted as λ^*_D, s.t.

$$\lambda^*_D = \begin{cases} D & D \leq 1/2 \\ 1-D & D \geq 1/2 \end{cases}$$

Then,

$$\min_{D,\lambda^*} g(D,\lambda^*) = \min_D \min_{\lambda^*} g_D(\lambda^*)$$

If $D \leq 1/2$,

$$\min_{\lambda^*} g(D,\lambda^*) = \frac{\hat{d}_1}{D} + \frac{1-\hat{d}_1}{1-D} + \frac{\hat{B}c_n}{D} \triangleq \tilde{g}_1(D)$$

and if $D \geq 1/2$,

$$\min_{\lambda^*} g(D,\lambda^*) = \frac{\hat{d}_1}{D} + \frac{1-\hat{d}_1}{1-D} + \frac{\hat{B}c_n}{1-D} \triangleq \tilde{g}_2(D).$$

We seek to minimize $\tilde{g}_1(D)$ on $[0,1/2]$ and minimize $\tilde{g}_2(D)$ on $[1/2,1]$ separately, then find the global optimal solution. Check that

$$\tilde{g}_1'(D) = -\frac{\hat{d}_1 + \hat{B}c_n}{D^2} + \frac{1-\hat{d}_1}{(1-D)^2}.$$

The solution to $\tilde{g}_1'(D) = 0$ in $(0,1)$ is

$$D_1^* = \frac{\sqrt{\frac{\hat{d}_1+\hat{B}c_n}{1-\hat{d}_1}}}{1+\sqrt{\frac{\hat{d}_1+\hat{B}c_n}{1-\hat{d}_1}}}$$

Since

$$g_1''(D) = 2(\hat{d}_1 + \hat{B}c_n)D^{-3} + 2(1-\hat{d}_1)(1-D)^{-3} > 0,$$

then D_1^* is the minimizer of $\tilde{g}_1(D)$. If $\hat{d}_1 \leq \frac{1-Bc_n}{2}$, which means that $D_1^* \leq 1/2$, then the optimal solution on $[0,\ 1/2]$ is $D^*_{opt,1} = D_1^*$ and $\lambda^*_{opt,1} = D_1^*$; otherwise, $D^*_{opt,1} = 1/2$ and $\lambda^*_{opt,1} = 1/2$.

Similarly,

$$\tilde{g}_2'(D) = -\frac{\hat{d}_1}{D^2} + \frac{1-\hat{d}_1+\hat{B}c_n}{(1-D)^2}.$$

The solution to $\tilde{g}_2'(D) = 0$ in $[0, 1]$ is

$$D_2^* = \frac{\sqrt{\frac{\hat{d}_1}{1-\hat{d}_1+\hat{B}c_n}}}{1+\sqrt{\frac{\hat{d}_1}{1-\hat{d}_1+\hat{B}c_n}}}$$

which is the minimizer of $\tilde{g}_2(D)$ since

$$g_2''(D) = 2\hat{d}_1 D^{-3} + 2(1-\hat{d}_1+\hat{B}c_n)(1-D)^{-3} > 0.$$

Again, if $\hat{d}_1 \geq \frac{1+\hat{B}c_n}{2}$, which means that $D_2^* \geq 1/2$, then the optimal solution on $[1/2, 1]$ is $D_{opt,2}^* = D_2^*$ and $\lambda_{opt,2}^* = 1 - D_2^*$; otherwise, $D_{opt,2}^* = 1/2$ and $\lambda_{opt,2}^* = 1/2$.

Hence, to obtain the global optimal solution, there are three cases.

- **Case I:** When $\hat{d}_1 \leq \frac{1-\hat{B}c_n}{2}$ (see Figure 4.3),

  D_2^* is the optimal solution when $D \in [0, 1/2]$ and $1/2$ is the optimal solution when $D \in [1/2, 1]$. If $\tilde{g}_1(D_1^*) < \tilde{g}_2(1/2)$, then the global optimal solutions are $D_{opt}^* = D_1^*$ and $\lambda_{opt}^* = D_1^*$; otherwise, $D_{opt}^* = 1/2$ and $\lambda_{opt}^* = 1/2$.

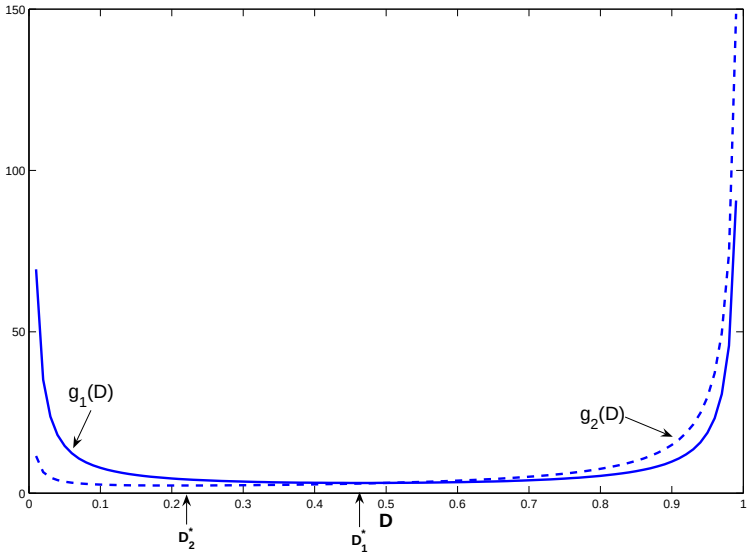


Figure 4.3: Optimal solution for the piecewise-constant model, when $\hat{d}_1 \leq \frac{1-\hat{B}c_n}{2}$.

- **Case II:** When $\hat{d}_1 \geq \frac{1+\hat{B}c_n}{2}$ (see Figure 4.4),

  $1/2$ is the optimal solution when $D \in [0, 1/2]$ and D_1^* is the optimal solution

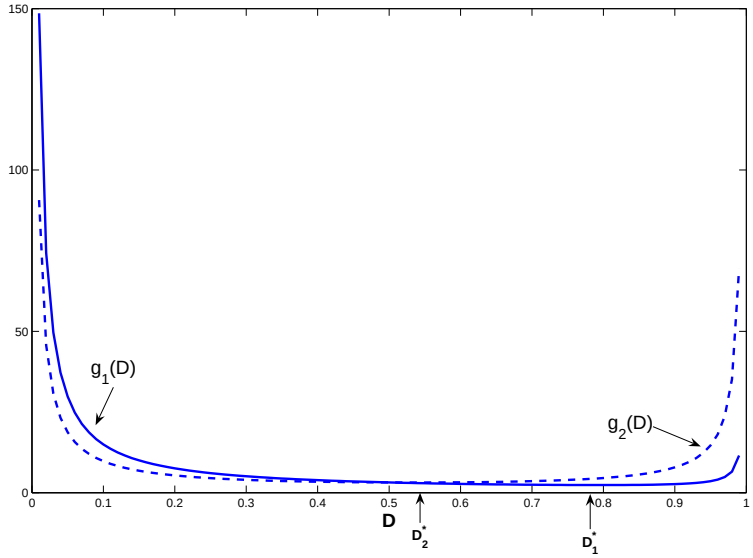


Figure 4.4: Optimal solution for the piecewise-constant model, $\hat{d}_1 \geq \frac{1+\hat{B}c_n}{2}$.

when $D \in [1/2, 1]$. If $\tilde{g}_2(D_2^*) < \tilde{g}_1(1/2)$, then $D_{opt}^* = D_2^*$ and $\lambda_{opt}^* = 1 - D_2^*$; otherwise, $D_{opt}^* = 1/2$ and $\lambda_{opt}^* = 1/2$.

- **Case III:** When $\frac{1-\hat{B}c_n}{2} < \hat{d}_1 < \frac{1+\hat{B}c_n}{2}$ (see Figure 4.5),

  Note that, $1/2$ is the optimal solution when $D \in [0, 1/2]$ as well as when $D \in [1/2, 1]$, so $D_{opt}^* = 1/2$ and $\lambda_{opt}^* = 1/2$.

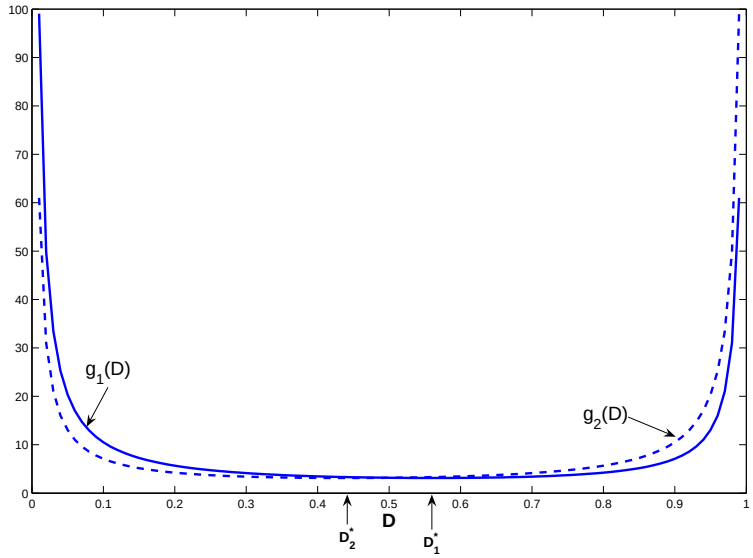


Figure 4.5: Optimal solution for the piecewise-constant model, $\frac{1-\hat{B}c_n}{2} < \hat{d}_1 < \frac{1+\hat{B}c_n}{2}$.

Therefore, the optimal selections of the design density are $\lambda_{1,opt} = \lambda^*_{opt}/c_n$, $\lambda_{2,opt} = (D^*_{opt} - \lambda^*_{opt})/L_n$, and $\lambda_{3,opt} = (1 - D^*_{opt} - \lambda^*_{opt})/U_n$.

4.1.2 General Parametric Model.

We are more interested in general parametric model, where

$$\mu(x) = \psi_l(\beta_l^0, x)I(x \le d^0) + \psi_u(\beta_u^0, x)I(x > d^0).$$

The definition of L_2 error becomes

$$\begin{aligned}
L_{2par,n} &= \int [\psi_l(\hat{\beta}_l, x)I\{x \le \hat{d}_n\} + \psi_u(\hat{\beta}_u, x)I\{x > \hat{d}_n\} \\
&\quad -\psi_l(\beta_l^0, x)I\{x \le d^0\} - \psi_u(\beta_u^0, x)I\{x > d^0\}]^2 dx \\
&= \left[I\{\hat{d}_n < d^0\} \left[\int_0^{\hat{d}_n} (\psi_l(\hat{\beta}_l, x) - \psi_l(\beta_l^0, x))^2 dx \right.\right. \\
&\qquad + \int_{d^0}^1 (\psi_u(\hat{\beta}_u, x) - \psi_u(\beta_u^0, x))^2 dx \\
&\qquad \left.\left. + \int_{\hat{d}_n}^{d^0} (\psi_u(\hat{\beta}_u, x) - \psi_l(\beta_l^0, x))^2 dx \right] \right. \\
&\quad \left. + I\{\hat{d}_n \ge d^0\} \left[\int_0^{d^0} (\psi_l(\hat{\beta}_l, x) - \psi_l(\beta_l^0, x))^2 dx \right.\right. \\
&\qquad + \int_{\hat{d}_n}^1 (\psi_u(\hat{\beta}_u, x) - \psi_u(\beta_u^0, x))^2 dx \\
&\qquad \left.\left. + \int_{d^0}^{\hat{d}_n} (\psi_u(\beta_u^0, x) - \psi_l(\hat{\beta}_l, x))^2 dx \right] \right]
\end{aligned}$$

If $\hat{d}_n < d^0$, by Taylor expansion, we derive the expression as

$$\begin{aligned}
&\int_0^{\hat{d}_n} (\psi_l(\hat{\beta}_l, x) - \psi_l(\beta_l^0, x))^2 dx \\
&= \int_0^{\hat{d}_n} \left[\sum_{k=1}^p (\hat{\beta}_{l,k} - \beta_{l,k}^0) \frac{\partial}{\partial \beta_{l,k}} \psi_l(\beta_l^0, x) + \psi_l(\beta_l^0, x) - \psi_l(\beta_l^0, x) \right. \\
&\qquad \left. + \frac{1}{2} \sum_{i,j} (\hat{\beta}_{l,i} - \beta_{l,i}^0)(\hat{\beta}_{l,j} - \beta_{l,j}^0) \frac{\partial^2 \psi_l(\beta_l^*, x)}{\partial \beta_{l,i} \partial \beta_{l,j}} \right]^2 dx \\
&= \int_0^{\hat{d}_n} \left[\sum_{k=1}^p (\hat{\beta}_{l,k} - \beta_{l,k}^0) \frac{\partial}{\partial \beta_{l,k}} \psi_l(\beta_l^0, x) \right. \\
&\qquad \left. + \frac{1}{2} \sum_{i,j} (\hat{\beta}_{l,i} - \beta_{l,i}^0)(\hat{\beta}_{l,j} - \beta_{l,j}^0) \frac{\partial^2 \psi_l(\beta_l^*, x)}{\partial \beta_{l,i} \partial \beta_{l,j}} \right]^2 dx
\end{aligned}$$

Define

$$\Psi^l_{i,j}(d) = \int_0^d \frac{\partial}{\partial\beta_{l,i}}\psi_l(\beta^0_l, x)\frac{\partial}{\partial\beta_{l,j}}\psi_l(\beta^0_l, x)dx$$

and

$$R^*_n = \frac{1}{2}\sum_{i,j}(\hat{\beta}_{l,i} - \beta^0_{l,i})(\hat{\beta}_{l,j} - \beta^0_{l,j})\frac{\partial^2\psi_l(\beta^*_l, x)}{\partial\beta_{l,i}\partial\beta_{l,j}}$$

Then

$$\begin{aligned}
&\int_0^{\hat{d}_n}(\psi_l(\hat{\beta}_l, x) - \psi_l(\beta^0_l, x))^2dx \\
= &\sum_{k=1}^p[(\hat{\beta}_{l,k} - \beta^0_{l,k})^2\Psi^l_{k,k}(\hat{d}_n)] + 2\sum_{i<j}[(\hat{\beta}_{l,i} - \beta^0_{l,i})(\hat{\beta}_{l,j} - \beta^0_{l,j})\Psi^l_{i,j}(\hat{d}_n)] \\
&+\int_0^{\hat{d}_n}\left[2\sum_{k=1}^p(\hat{\beta}_{l,k} - \beta^0_{l,k})\frac{\partial}{\partial\beta_{l,k}}\psi_l(\beta^0_l, x)R^*_n + R^{*2}_n\right]dx \\
= &\sum_{k=1}^p[(\hat{\beta}_{l,k} - \beta^0_{l,k})^2\Psi^l_{k,k}(d^0)] + 2\sum_{i<j}[(\hat{\beta}_{l,i} - \beta^0_{l,i})(\hat{\beta}_{l,j} - \beta^0_{l,j})\Psi^l_{i,j}(d^0)] \\
&+\sum_{k=1}^p[(\hat{\beta}_{l,k} - \beta^0_{l,k})^2(\Psi^l_{k,k}(\hat{d}_n) - \Psi^l_{k,k}(d^0))] \\
&+2\sum_{i<j}[(\hat{\beta}_{l,i} - \beta^0_{l,i})(\hat{\beta}_{l,j} - \beta^0_{l,j})(\Psi^l_{i,j}(\hat{d}_n) - \Psi^l_{i,j}(d^0))] \\
&+\int_0^{\hat{d}_n}\left[2\sum_{k=1}^p(\hat{\beta}_{l,k} - \beta^0_{l,k})\frac{\partial}{\partial\beta_{l,k}}\psi_l(\beta^0_l, x)R^*_n + R^{*2}_n\right]dx
\end{aligned}$$

Since $\int_0^{\hat{d}_n}\left[2\sum_{k=1}^p(\hat{\beta}_{l,k} - \beta^0_{l,k})\frac{\partial}{\partial\beta_{l,k}}\psi_l(\beta^0_l, x)R^*_n + R^{*2}_n\right]dx$, $\sum_{k=1}^p[(\hat{\beta}_{l,k}-\beta^0_{l,k})^2(\Psi^l_{k,k}(\hat{d}_n)-\Psi^l_{k,k}(d^0))]$ and $\sum_{i<j}[(\hat{\beta}_{l,i}-\beta^0_{l,i})(\hat{\beta}_{l,j}-\beta^0_{l,j})(\Psi^l_{i,j}(\hat{d}_n)-\Psi^l_{i,j}(d^0))]$ have faster rate of convergence than $1/n$, we can write

$$\begin{aligned}
&\int_0^{\hat{d}_n}(\psi_l(\hat{\beta}_l, x) - \psi_l(\beta^0_l, x))^2dx \\
= &\frac{\sigma^2}{n}\left[\sum_{k=1}^p\frac{n(\hat{\beta}_{l,k} - \beta^0_{l,k})^2}{\sigma^2}\Psi^l_{k,k}(d^0)\right. \\
&\left.+2\sum_{i<j}\frac{\sqrt{n}(\hat{\beta}_{l,i} - \beta^0_{l,i})\sqrt{n}(\hat{\beta}_{l,j} - \beta^0_{l,j})}{\sigma^2}\Psi^l_{i,j}(d^0)\right] + o_p(1/n)
\end{aligned}$$

In the same way of derivation, by defining $\Psi^u_{i,j}(d) = \int_d^1 \frac{\partial}{\partial\beta_{u,i}}\psi_u(\beta^0_u, x)\frac{\partial}{\partial\beta_{u,j}}\psi_u(\beta^0_u, x)dx$, we can obtain that

$$\begin{aligned}
&\int_{d^0}^1 (\psi_u(\hat{\beta}_u, x) - \psi_u(\beta^0_u, x))^2 dx \\
= \;& \frac{\sigma^2}{n}\left[\sum_{k=1}^q \frac{n(\hat{\beta}_{u,k} - \beta^0_{u,k})}{\sigma^2}\Psi^u_{k,k}(d^0)\right. \\
& \left. +2\sum_{i<j}\frac{\sqrt{n}(\hat{\beta}_{u,i} - \beta^0_{u,i})\sqrt{n}(\hat{\beta}_{u,j} - \beta_{u,j})}{\sigma^2}\Psi^u_{i,j}(d^0)\right] + o_p(1/n)
\end{aligned}$$

$\int_{\hat{d}_n}^{d^0}(\psi_u(\hat{\beta}_u, x) - \psi_l(\beta^0_l, x))^2 dx$ can be expressed as

$$\begin{aligned}
&\int_{\hat{d}_n}^{d^0} (\psi_u(\hat{\beta}_l, x) - \psi_l(\beta^0_l, x))^2 dx \\
= \;& \frac{\sigma^2}{n} n\int_{\hat{d}_n}^{d^0}\left(\sum_{k=1}^q(\hat{\beta}_{u,k} - \beta^0_{u,k})\frac{\partial\psi_u(\beta^0_u, x)}{\partial\beta_{u,k}} + \psi(\beta^0_u, x) - \psi_l(\beta^0_l, x)\right. \\
& \left. +\frac{1}{2}\sum_{i,j}(\hat{\beta}_{u,i} - \beta^0_{u,i})(\hat{\beta}_{u,j} - \beta^0_{u,j})\frac{\partial^2\psi_u(\beta^*_u, x)}{\partial\beta_{u,i}\partial\beta_{u,j}}\right)^2 dx/\sigma^2 \\
= \;& \frac{\sigma^2}{n}(n(d^0 - \hat{d}_n))(\psi(\beta^0_u, d^0) - \psi_l(\beta^0_l, d^0))^2/\sigma^2 + o_p(1/n) \\
= \;& \frac{\sigma^2}{n}(n(d^0 - \hat{d}_n))(SNR)^2 + o_p(1/n)
\end{aligned}$$

If $\hat{d}_n > d^0$, the similar expressions can be derived:

$$\begin{aligned}
&\int_0^{d^0} (\psi_l(\hat{\beta}_l, x) - \psi_l(\beta^0_l, x))^2 dx \\
= \;& \frac{\sigma^2}{n}\left[\sum_{k=1}^p \frac{n(\hat{\beta}_{l,k} - \beta^0_{l,k})^2}{\sigma^2}\Psi^u_{k,k}(d^0)\right. \\
& \left. +2\sum_{i<j}\frac{\sqrt{n}(\hat{\beta}_{l,i} - \beta^0_{l,i})\sqrt{n}(\hat{\beta}_{l,j} - \beta^0_{l,j})}{\sigma^2}\Psi^l_{i,j}(d^0)\right] + o_p(1/n) \\
&\int_{\hat{d}_n}^1 (\psi_u(\hat{\beta}_u, x) - \psi_u(\beta^0_u, x))^2 dx \\
= \;& \frac{\sigma^2}{n}\left[\sum_{k=1}^q \frac{n(\hat{\beta}_{u,k} - \beta^0_{u,k})}{\sigma^2}\Psi^u_{k,k}(d^0)\right. \\
& \left. +2\sum_{i<j}\frac{\sqrt{n}(\hat{\beta}_{u,i} - \beta^0_{u,i})\sqrt{n}(\hat{\beta}_{u,j} - \beta^0_{u,j})}{\sigma^2}\Psi^u_{i,j}(d^0)\right] + o_p(1/n) \\
&\int_{d^0}^{\hat{d}_n} (\psi_u(\beta^0_u, x) - \psi_l(\hat{\beta}_l, x))^2 dx \\
= \;& \frac{\sigma^2}{n}(n(\hat{d}_n - d^0))(SNR)^2 + o_p(1/n)
\end{aligned}$$

Since asymptotic normalities of β_l and β_u are the same as with d^0 known,

$$\begin{aligned}\sqrt{n}(\hat{\beta}_l - \beta_l^0) &\rightarrow N(0, \Sigma_l)\\ \sqrt{n}(\hat{\beta}_u - \beta_u^0) &\rightarrow N(0, \Sigma_u)\end{aligned}$$

where Σ_l and Σ_u are the corresponding dispersion matrices of asymptotic normal distributions and we will specify them later. Therefore, we have the asymptotic expected L_2 error:

$$\begin{aligned}EL_{2par,n} &= \frac{\sigma^2}{n}\left[\sum_{k=1}^{p}\frac{Var(\sqrt{n}(\hat{\beta}_{l,k}-\beta_{l,k}^0))}{\sigma^2}\Psi_{k,k}^l(d^0)\right.\\ &\quad +2\sum_{i<j}\frac{Cov(\sqrt{n}(\hat{\beta}_{l,i}-\beta_{l,i}^0),\sqrt{n}(\hat{\beta}_{l,j}-\beta_{l,j}^0))}{\sigma^2}\Psi_{i,j}^l(d^0)\\ &\quad +\sum_{k=1}^{q}\frac{Var(\sqrt{n}(\hat{\beta}_{u,k}-\beta_{u,k}^0))}{\sigma^2}\Psi_{k,k}^u(d^0)\\ &\quad +2\sum_{i<j}\frac{Cov(\sqrt{n}(\hat{\beta}_{u,i}-\beta_{u,i}^0),\sqrt{n}(\hat{\beta}_{u,j}-\beta_{u,j}^0))}{\sigma^2}\Psi_{i,j}^u(d^0)\\ &\quad \left.+E(n|\hat{d}-d^0|)SNR^2\right]\\ &= \frac{\sigma^2}{n}\left[\sum_{k=1}^{p}\frac{Var(\sqrt{n}(\hat{\beta}_{l,k}-\beta_{l,k}^0))}{\sigma^2}\Psi_{k,k}^l(d^0)\right.\\ &\quad +2\sum_{i<j}\frac{Cov(\sqrt{n}(\hat{\beta}_{l,i}-\beta_{l,i}^0),\sqrt{n}(\hat{\beta}_{l,j}-\beta_{l,j}^0))}{\sigma^2}\Psi_{i,j}^l(d^0)\\ &\quad +\sum_{k=1}^{q}\frac{Var(\sqrt{n}(\hat{\beta}_{u,k}-\beta_{u,k}^0))}{\sigma^2}\Psi_{k,k}^u(d^0)\\ &\quad +2\sum_{i<j}\frac{Cov(\sqrt{n}(\hat{\beta}_{u,i}-\beta_{u,i}^0),\sqrt{n}(\hat{\beta}_{u,j}-\beta_{u,j}^0))}{\sigma^2}\Psi_{i,j}^u(d^0)\\ &\quad \left.+E(V)(SNR)^2/f_X(d^0)\right]\end{aligned}$$

The asymptotic dispersion matrices for $\sqrt{n}(\hat{\beta}_l - \beta_l)$ and $\sqrt{n}(\hat{\beta}_u - \beta_u)$ depend on the density of X , which leads to complexities of optimal solution arguments. According to the asymptotic property of Z-estimator, we know that

$$\begin{aligned}Var(\sqrt{n}(\hat{\beta}_l - \beta_l^0)/\sigma) &\rightarrow V_{\beta_l}^{-1} \equiv \Sigma_l/\sigma^2\\ Var(\sqrt{n}(\hat{\beta}_u - \beta_u^0)/\sigma) &\rightarrow V_{\beta_u}^{-1} \equiv \Sigma_u/\sigma^2\end{aligned}$$

where

$$\begin{aligned} V_{\beta_l} &= \int_0^{d^0} \nabla\psi_l(\beta_l, x) \cdot (\nabla\psi_l(\beta_l, x))^T f_X(x)dx \\ V_{\beta_u} &= \int_{d^0}^1 \nabla\psi_u(\beta_u, x) \cdot (\nabla\psi_u(\beta_u, x))^T f_X(x)dx \end{aligned}$$

where $\nabla\psi_l(\beta_l, x)$ and $\nabla\psi_u(\beta_u, x)$ are the gradients in β_l and β_u, respectively. Now, the optimization problem is defined as following:

$$(\lambda_1^*, \lambda_2^*, \lambda_3^*) = \text{ Argmin } \tilde{h}(\lambda_1, \lambda_2, \lambda_3, SNR, d^0, n)$$

such that

$$\lambda_1 \geq 0,\ \lambda_2 \geq 0,\ \lambda_3 \geq 0,\ 2\lambda_1\tau + \lambda_2 L \leq 1$$

where

$$\begin{aligned} \tilde{h}(\cdot) &= \sum_{k=1}^{p}(V_{\beta_l}^{-1})_{kk}\Psi_{k,k}^l(d^0) + 2\sum_{i<j}(V_{\beta_l}^{-1})_{ij}\Psi_{i,j}^l(d^0) \\ &\quad + \sum_{k=1}^{q}(V_{\beta_u}^{-1})_{kk}\Psi_{k,k}^u(d^0) + 2\sum_{i<j}(V_{\beta_u}^{-1})_{ij}\Psi_{i,j}^u(d^0) \\ &\quad + E(V)(SNR)^2/\lambda_1 \end{aligned}$$

Again, in practice, we do not know d^0 and SNR. We need substitute d^0 by $\hat{d}_1$ from initial samples and take $\tau = C_\zeta/n \triangleq c_n$, $L_n = \hat{d}_1 - c_n$, where C_ζ is upper ζ^{th} quantile of Argmin $\mathbb{M}_{S\hat{N}R,Z,1}$. Hence the surrogate expression is

$$\begin{aligned} h(\cdot) &= \sum_{k=1}^{p}(V_{\beta_l}^{-1})_{kk}\Psi_{k,k}^l(\hat{d}_1) + 2\sum_{i<j}(V_{\beta_l}^{-1})_{ij}\Psi_{i,j}^l(\hat{d}_1) \\ &\quad + \sum_{k=1}^{q}(V_{\beta_u}^{-1})_{kk}\Psi_{k,k}^u(\hat{d}_1) + 2\sum_{i<j}(V_{\beta_u}^{-1})_{ij}\Psi_{i,j}^u(\hat{d}_1) \\ &\quad + \widehat{E(V)}\widehat{(SNR)}^2/\lambda_1 \end{aligned}$$

The optimal solutions to $h(\cdot)$ will provide suggest on how to allocate samples in $[0, 1]$.

4.1.3 Linear–Linear Model

Since we can obtain the general optimal formula for parametric models, the polynomial model is one specific case. For a linear-linear model, we can define the

expected L_2 error as

$$\begin{aligned} EL_{2ll,n} &= E\int[(\hat{\alpha}_1+\hat{\alpha}_2 x)I\{x\le \hat{d}_n^{(1)}\}+(\hat{\beta}_1+\hat{\beta}_2 x)I\{x>\hat{d}_n^{(1)}\} \\ &\quad -(\alpha_1^0+\alpha_2^0 x)I\{x\le d^0\}-(\beta_1^0+\beta_2^0 x)I\{x>d^0\}]^2dx \end{aligned}$$

Using the results from Section 4.1.2, we seek to minimize:

$$\begin{aligned} \tilde{h}(\cdot) &= \sum_{k=1}^{p}(V_{\beta_l}^{-1})_{kk}\Psi_{k,k}^{l}(d^0)+2\sum_{i<j}(V_{\beta_l}^{-1})_{ij}\Psi_{i,j}^{l}(d^0) \\ &\quad +\sum_{k=1}^{q}(V_{\beta_u}^{-1})_{kk}\Psi_{k,k}^{u}(d^0)+2\sum_{i<j}(V_{\beta_u}^{-1})_{ij}\Psi_{i,j}^{u}(d^0) \\ &\quad +E(V)(SNR)^2/\lambda_1 \end{aligned}$$

We derive the exact forms of $V_{\beta_l}^{-1}$ and $V_{\beta_u}^{-1}$ for this specific model next.

Define $\hat{\alpha}=(\hat{\alpha}_1,\hat{\alpha}_2)'$ and $\hat{\beta}=(\hat{\beta}_1,\hat{\beta}_2)'$, to derive the asymptotic dispersion matrices of $\sqrt{n}(\hat{\alpha}-\alpha^0)/\sigma$ and $\sqrt{n}(\hat{\beta}-\beta^0)/\sigma$, suppose $\sum I(x_i\le d^0)=m$.

$$\begin{aligned} Var(\hat{\alpha}/\sigma|X) &= (X'X)^{-1}=\begin{bmatrix} m & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}^{-1} \\ &= \frac{1}{m\sum X_i^2-(\sum X_i)^2}\begin{bmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & m \end{bmatrix} \end{aligned}$$

where

$$\begin{aligned} \frac{1}{m\sum X_i^2-(\sum X_i)^2} &= \frac{1}{m^2}\cdot\frac{1}{1/m\sum X_i^2-(1/m\sum X_i)^2} \\ 1/m\sum X_i^2-(1/m\sum X_i)^2 &= \frac{\mathbb{P}_n X^2 I(X\le d^0)}{\mathbb{P}_n I(X\le d^0)}-\left(\frac{\mathbb{P}_n X I(X\le d^0)}{\mathbb{P}_n I(X\le d^0)}\right)^2 \\ &\longrightarrow \frac{\int_0^{d^0}x^2 f_X(x)dx}{F_X(d^0)}-\left(\frac{\int_0^{d^0}x f_X(x)dx}{F_X(d^0)}\right)^2 \\ &\triangleq \delta_1 \end{aligned}$$

Then

$$\begin{aligned} Var(\sqrt{n}(\hat{\alpha}-\alpha^0)/\sigma|X) &= \frac{n}{m}Var(\sqrt{m}(\hat{\alpha}-\alpha^0)|X) \\ &= \frac{n}{m}\cdot\frac{1}{m^2\delta_1}\begin{bmatrix} m\sum X_i^2 & -m\sum X_i \\ -m\sum X_i & m^2 \end{bmatrix} \\ &= \frac{n}{m}\cdot\frac{1}{\delta_1}\begin{bmatrix} 1/m\sum X_i^2 & -1/m\sum X_i \\ -1/m\sum X_i & 1 \end{bmatrix} \\ Var(\sqrt{n}(\hat{\alpha}-\alpha^0)/\sigma) &\to \frac{1}{F_X^2(d^0)\delta_1}\begin{bmatrix} \int_0^{d^0} x^2 f_X(x)dx & -\int_0^{d^0} x f_X(x)dx \\ -\int_0^{d^0} x f_X(x)dx & F(d^0) \end{bmatrix} \end{aligned}$$

Similarly, we can obtain

$$Var(\sqrt{n}(\hat{\beta}-\beta^0)/\sigma) \to \frac{1}{(1-F_X(d^0))^2\delta_2}\begin{bmatrix} \int_{d^0}^1 x^2 f_X(x)dx & -\int_{d^0}^1 x f_X(x)dx \\ -\int_{d^0}^1 x f_X(x)dx & 1-F(d^0) \end{bmatrix}$$

where

$$\delta_2 = \frac{\int_{d^0}^1 x^2 f_X(x)dx}{1-F_X(d^0)} - \left(\frac{\int_{d^0}^1 x f_X(x)dx}{1-F_X(d^0)}\right)^2$$

For simplicity, we define

$$\begin{aligned} H_1(d) &= \textstyle\int_0^d x f_X(x)dx \quad H_2 = \int_0^d x^2 f_X(x)dx \\ G_1(d) &= \textstyle\int_d^1 x f_X(x)dx \quad G_2 = \int_d^1 x^2 f_X(x)dx \end{aligned}$$

then

$$Var(\sqrt{n}(\hat{\alpha}-\alpha^0)/\sigma) \to \frac{1}{H_2(d^0)F_X(d^0)-H_1^2(d^0)}\begin{bmatrix} H_2(d^0) & -H_1(d^0) \\ -H_1(d^0) & F_X(d^0) \end{bmatrix} \tag{4.2}$$

$$Var(\sqrt{n}(\hat{\beta}-\beta^0)/\sigma) \to \frac{1}{G_2(d^0)(1-F_X(d^0))-G_1^2(d^0)}\begin{bmatrix} G_2(d^0) & -G_1(d^0) \\ -G_1(d^0) & 1-F_X(d^0) \end{bmatrix} \tag{4.3}$$

Using the results from the discussion on general parametric models in Section 4.1.2, we have

$$V_{\beta_l} \equiv V_\alpha = \begin{bmatrix} F_X(d^0) & H_1(d^0) \\ H_1(d^0) & H_2(d^0) \end{bmatrix} \tag{4.4}$$

$$V_{\beta_u} \equiv V_\alpha = \begin{bmatrix} 1-F_X(d^0) & G_1(d^0) \\ G_1(d^0) & G_2(d^0) \end{bmatrix} \tag{4.5}$$

Then

$$
\begin{aligned}
Var(\sqrt{n}(\hat{\alpha}-\alpha^0)/\sigma) &\rightarrow \begin{bmatrix} F_X(d^0) & H_1(d^0) \\ H_1(d^0) & H_2(d^0) \end{bmatrix}^{-1} \\
&= \frac{1}{H_2(d^0)F_X(d^0)-H_1^2(d^0)} \begin{bmatrix} H_2(d^0) & -H_1(d^0) \\ -H_1(d^0) & F_X(d^0) \end{bmatrix} \\
Var(\sqrt{n}(\hat{\beta}-\beta^0)/\sigma) &\rightarrow \begin{bmatrix} 1-F_X(d^0) & G_1(d^0) \\ G_1(d^0) & G_2(d^0) \end{bmatrix}^{-1} \\
&= \frac{1}{G_2(d^0)(1-F_X(d^0))-G_1^2(d^0)} \begin{bmatrix} G_2(d^0) & -G_1(d^0) \\ -G_1(d^0) & F_X(d^0) \end{bmatrix} \sigma^2
\end{aligned}
$$

which are exactly the same as in (4.2) and (4.3), respectively. It is not hard to verify that $\Psi^l_{1,1}(d^0) = d^0$, $\Psi^l_{2,2}(d^0) = {d^0}^3/3$, $\Psi^l_{1,2}(d^0) = {d^0}^2/2$, $\Psi^u_{1,1}(d^0) = 1-d^0$, $\Psi^u_{2,2}(d^0) = (1-{d^0}^3)/3$ and $\Psi^l_{1,2}(d^0) = (1-{d^0}^2)/2$.

Therefore, we can obtain the surrogate expression of our objective function as following by substituting d^0 by $\hat{d}_1$ from initial samples, taking $\tau = c_n = C_\zeta/n$, $L_n = \hat{d}_1 - c_n$ and $U_n = \hat{d}_1 + c_n$ and using the estimated SNR.

$$
\begin{aligned}
h(\cdot) &= \frac{\hat{d}_1}{F_X^2(\hat{d}_1)\delta_1}\int_0^{\hat{d}_1} x^2 f_X(x)dx + \frac{\hat{d}_1^3}{3F_X(\hat{d}_1)\delta_1} - \frac{\hat{d}_1^2}{F_X^2(\hat{d}_1)\delta_1}\int_0^{\hat{d}_1} x f_X(x)dx \\
&+\frac{(1-\hat{d}_1)}{(1-F_X(\hat{d}_1))^2\delta_2}\int_{\hat{d}_1}^1 x^2 f_X(x)dx + \frac{(1-\hat{d}_1^3)}{3(1-F_X(\hat{d}_1))\delta_2} \\
&-\frac{(1-\hat{d}_1^2)}{(1-F_X(\hat{d}_1))^2\delta_2}\int_{\hat{d}_1}^1 x f_X(x)dx \\
&+\widehat{(SNR)}^2 \frac{1}{f_X(\hat{d}_1)}\widehat{EV}
\end{aligned}
$$

such that

$$\lambda_1 \geq 0,\ \lambda_2 \geq 0,\ \lambda_3 \geq 0,\ 2\lambda_1 c_n + \lambda_2 L_n \leq 1$$

where

$$
\begin{aligned}
f_X(\hat{d}_1) &= \lambda_1 \\
F_X(\hat{d}_1) &= \lambda_2 L_n + \lambda_1 c_n \\
1 - F_x(\hat{d}_1) &= 1 - (\lambda_2 L_n + \lambda_1 c_n) \\
H_1(\hat{d}_1) &= 1/2(\lambda_2 L_n^2 + \lambda_1(\hat{d}_1^2 - L_n^2)) \\
H_2(\hat{d}_1) &= 1/3(\lambda_2 L_n^3 + \lambda_1(\hat{d}_1^3 - L_n^3)) \\
G_1(\hat{d}_1) &= 1/2(\lambda_1(U_n^2 - \hat{d}_1^2) + \lambda_3(1 - U_n^2)) \\
&= 1/2(\lambda_1(U_n^2 - \hat{d}_1^2) + (1 - \lambda_2 L_n - 2\lambda_1 c_n)(1 + U_n)) \\
G_2(\hat{d}_1) &= 1/3(\lambda_1(U_n^3 - \hat{d}_1^3) + \lambda_3(1 - U_n^3)) \\
&= 1/3(\lambda_1(U_n^3 - \hat{d}_1^3) + (1 - \lambda_2 L_n - 2\lambda_1 c_n)(1 + U_n + U_n^2)) \\
\frac{1}{\delta_1} &= \frac{(\lambda_2 L_n + \lambda_1 c_n)^2}{1/3(\lambda_2 L_n^3 + \lambda_1(\hat{d}_1^3 - L_n^3))(\lambda_2 L_n + \lambda_1 c_n) - 1/4(\lambda_2 L_n^2 + \lambda_1(\hat{d}_1^2 - L_n^2))^2} \\
\frac{1}{\delta_2} &= \frac{1 - (\lambda_2 L_n + \lambda_1 c_n)^2}{B_1 - B_2} \\
B_1 &= 1/3(\lambda_1(U_n^3 - \hat{d}_1^3) + (1 - \lambda_2 L_n - 2\lambda_1 c_n)(1 + U_n + U_n^2))(1 - \lambda_2 L_n + \lambda_1 c_n) \\
B_2 &= 1/4(\lambda_1(U_n^2 - \hat{d}_1^2) + (1 - \lambda_2 L_n - 2\lambda_1 c_n)(1 + U_n))^2
\end{aligned}
$$

It is not easy to derive the optimal solutions from the defined optimization problem directly (see Figure 4.6). We use the optimization function "fmincon" in Matlab to search for the optimal solutions and obtain Table 4.1. We list some optimal

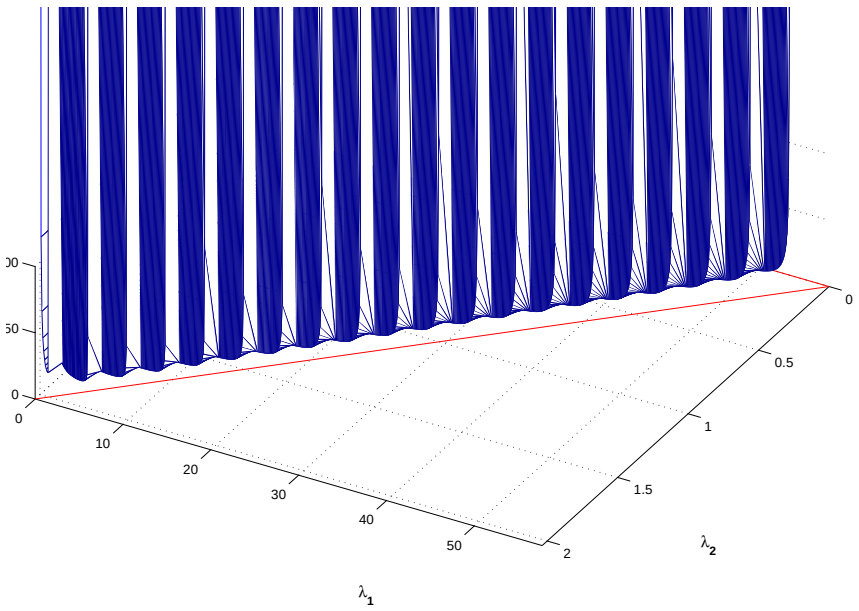


Figure 4.6: $h(\lambda_1, \lambda_2)$ function, linear-linear model

allocations for SNR=2 and SNR=5 by using five true values of change point. There are symmetric patterns as we expected. For example, when SNR=5, $d^0 = 0.2$, the

optimal allocation is $(.2188, .3126, .4686)$, while when $d^0 = 0.8$, the optimal allocation is $(.4686, .3126, .2188)$. These results are reasonable, since those change points are very close to the covariate's boundary, we should not put much allocation in those small areas.

d^0	SNR=5			SNR=2		
	left	middle	right	left	middle	right
0.2	.2188	.3126	.4686	.2079	.3295	.4626
0.4	.3080	.3072	.3848	.3003	.3216	.3781
0.5	.3467	.3067	.3466	.3395	.3210	.3396
0.6	.3848	.3072	.3080	.3781	.3216	.3003
0.8	.4686	.3126	.2188	.4632	.3287	.2081

Table 4.1: Optimal allocation of samples in the second stage when using d^0, $n_1 = n_2 = 500$, linear-linear model

Define

$$
\begin{aligned}
h_1 &= \frac{\hat{d}_1}{F_X^2(\hat{d}_1)\delta_1}\int_0^{\hat{d}_1} x^2 f_X(x)dx + \frac{\hat{d}_1^3}{3F_X(\hat{d}_1)\delta_1} - \frac{\hat{d}_1^2}{F_X^2(\hat{d}_1)\delta_1}\int_0^{\hat{d}_1} x f_X(x)dx \\
h_2 &= +\frac{(1-\hat{d}_1)}{(1-F_X(\hat{d}_1))^2\delta_2}\int_{\hat{d}_1}^1 x^2 f_X(x)dx + \frac{(1-\hat{d}_1^3)}{3(1-F_X(\hat{d}_1))\delta_2} \\
&\quad -\frac{(1-\hat{d}_1^2)}{(1-F_X(\hat{d}_1))^2\delta_2}\int_{\hat{d}_1}^1 x f_X(x)dx \\
h_3 &= \widehat{(SNR)}^2\frac{1}{f_X(\hat{d}_1)}\widehat{EV}
\end{aligned}
$$

To understand how h_1, h_2, and h_3 affect the value of $h(\cdot)$, we draw two plots as shown in Figure 4.7. The upper one is $\log(h_1 + h_2)$ vs. the allocation in the middle interval. And the lower one is $\log(h_3)$ vs. the allocation in the middle interval. From these pictures, the optimal allocation is around 0.3, which is consistent with the numerical results.

4.1.4 Quadratic–Quadratic Model

In real world we may be more interested in considering a higher degree polynomial regression function. For example, a quadratic-quadratic model, the definition of EL_2 becomes

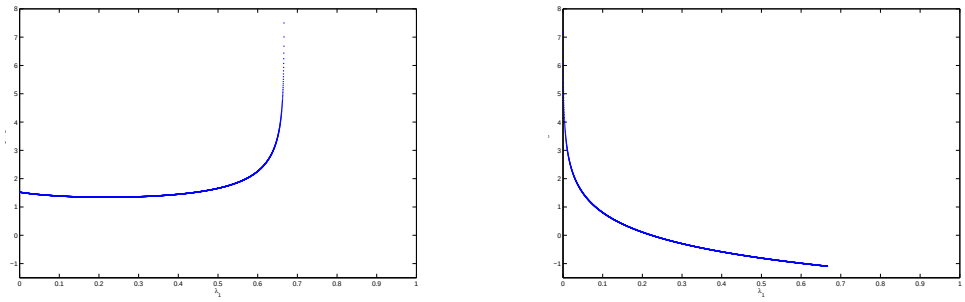


Figure 4.7: Optimal allocation in the middle interval, linear-linear model

$$EL_{2qq,n} = E\int[(\hat{\alpha}_1+\hat{\alpha}_2 x+\hat{\alpha}_3 x^2)I\{x\le \hat{d}_n\}+(\hat{\beta}_1+\hat{\beta}_2 x+\hat{\beta}_3 x^2)I\{x>\hat{d}_n\}$$

Again, we can obtain the following expression for $EL_{2qq,n}$ by using the similar manner as in Section 4.1.2:

$$\begin{aligned}\tilde{h}(\cdot) &= \sum_{k=1}^{p}(V_{\beta_l}^{-1})_{kk}\Psi_{k,k}^{l}(d^0)+2\sum_{i<j}(V_{\beta_l}^{-1})_{ij}\Psi_{i,j}^{l}(d^0)\\ &\quad+\sum_{k=1}^{q}(V_{\beta_u}^{-1})_{kk}\Psi_{k,k}^{u}(d^0)+2\sum_{i<j}(V_{\beta_u}^{-1})_{ij}\Psi_{i,j}^{u}(d^0)\\ &\quad+E(V)(SNR)^2/\lambda_1\end{aligned}$$

To specify the dispersion matrices of $\sqrt{n}(\hat{\alpha}-\alpha^0)/\sigma$ and $\sqrt{n}(\hat{\beta}-\beta^0)/\sigma$, we can derive them as following or by using the results directly from the discussion of general parametric model. For $\hat{\alpha}=(\hat{\alpha}_1,\hat{\alpha}_2,\hat{\alpha}_3)'$, suppose $\sum I(x_i\le d^0)=m$,

$$\begin{aligned}&Var(\hat{\alpha}|X)=(X'X)^{-1}\sigma^2\\ =\;&\frac{1}{\Delta_l}\begin{bmatrix}\sum X_i^2\sum X_i^4-(\sum X_i^3)^2 & \sum X_i^2\sum X_i^3-\sum X_i\sum X_i^4 & \sum X_i\sum X_i^3-(\sum X_i^2)^2\\ \sum X_i^2\sum X_i^3-\sum X_i\sum X_i^4 & -(\sum X_i^2)^2+m\sum X_i^4 & \sum X_i\sum X_i^2-m\sum X_i^3\\ \sum X_i\sum X_i^3-(\sum X_i^2)^2 & \sum X_i\sum X_i^2-m\sum X_i^3 & m\sum X_i^2-(\sum X_i)^2\end{bmatrix}\sigma^2\\ &Var(\sqrt{n}(\hat{\alpha}-\alpha^0)|X)\\ =\;&\frac{1}{F_X(d^0)\Delta_l}\cdot m\\ &\times\begin{bmatrix}\sum X_i^2\sum X_i^4-(\sum X_i^3)^2 & \sum X_i^2\sum X_i^3-\sum X_i\sum X_i^4 & \sum X_i\sum X_i^3-(\sum X_i^2)^2\\ \sum X_i^2\sum X_i^3-\sum X_i\sum X_i^4 & -(\sum X_i^2)^2+m\sum X_i^4 & \sum X_i\sum X_i^2-m\sum X_i^3\\ \sum X_i\sum X_i^3-(\sum X_i^2)^2 & \sum X_i\sum X_i^2-m\sum X_i^3 & m\sum X_i^2-(\sum X_i)^2\end{bmatrix}\sigma^2\end{aligned}$$

where

$$\Delta_l=-(\sum X_i^2)^3+2\sum X_i\sum X_i^2\sum X_i^3-(\sum X_i)^2\sum X_i^4-m(\sum X_i^3)^2+m\sum X_i^2\sum X_i^4$$

Define

$$\begin{aligned} H_1(d) &= \int_0^d x f_X(x)dx, \qquad H_2(d) = \int_0^d x^2 f_X(x)dx, \\ H_3(d) &= \int_0^d x^3 f_X(x)dx, \qquad H_4(d) = \int_0^d x^4 f_X(x)dx \end{aligned}$$

We would obtain

$$\begin{aligned} & \frac{1}{m^3}\Delta_l \\ \longrightarrow \quad & \frac{-H_2^3(d^0)}{F_X(d^0)^3} + \frac{2H_1(d^0)H_2(d^0)H_3(d^0)}{F_X(d^0)^3} - \frac{H_1^2(d^0)H_4(d^0)}{F_X(d^0)^3} - \frac{H_3^2(d^0)}{F_X(d^0)^2} + \frac{H_2(d^0)H_4(d^0)}{F_X(d^0)^2} \\ \triangleq \quad & \delta_{ql}(d^0) \end{aligned}$$

Then

$$Var(\sqrt{n}(\hat{\alpha} - \alpha^0)) = \frac{1}{F_X(d^0)^3 \delta_{ql}(d^0)} W(d^0)\sigma^2$$

The elements of $W(d^0)$ are:

$$\begin{aligned} w_{11} &= H_2(d^0)H_4(d^0) - H_3^2(d^0) \qquad & w_{12} = w_{21} &= H_2(d^0)H_3(d^0) - H_1(d^0)H_4(d^0) \\ w_{22} &= -H_2^2(d^0) + H_4(d^0)F_X(d^0) & w_{13} = w_{31} &= H_1(d^0)H_3(d^0) - H_2^2(d^0) \\ w_{33} &= H_2(d^0)F_X(d^0) - H_1^2(d^0) \qquad & w_{23} = w_{32} &= H_1(d^0)H_2(d^0) - H_3(d^0)F_X(d^0) \end{aligned}$$

For $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)'$, suppose $\sum I(x_i > d^0) = n - m$,

$$\begin{aligned} & Var(\hat{\beta}|X) = (X'X)^{-1}\sigma^2 \\ = \quad & \frac{\sigma^2}{\Delta_r} \begin{bmatrix} \sum X_i^2 \sum X_i^4 - (\sum X_i^3)^2 & \sum X_i^2 \sum X_i^3 - \sum X_i \sum X_i^4 & \sum X_i \sum X_i^3 - (\sum X_i^2)^2 \\ \sum X_i^2 \sum X_i^3 - \sum X_i \sum X_i^4 & -(\sum X_i^2)^2 + (n-m)\sum X_i^4 & \sum X_i \sum X_i^2 - (n-m)\sum X_i^3 \\ \sum X_i \sum X_i^3 - (\sum X_i^2)^2 & \sum X_i \sum X_i^2 - (n-m)\sum X_i^3 & (n-m)\sum X_i^2 - (\sum X_i)^2 \end{bmatrix} \\ & Var(\sqrt{n}(\hat{\beta} - \beta^0)|X) \\ = \quad & \frac{\sigma^2}{(1 - F_X(d^0))\Delta_r} \cdot (n-m) \\ & \times \begin{bmatrix} \sum X_i^2 \sum X_i^4 - (\sum X_i^3)^2 & \sum X_i^2 \sum X_i^3 - \sum X_i \sum X_i^4 & \sum X_i \sum X_i^3 - (\sum X_i^2)^2 \\ \sum X_i^2 \sum X_i^3 - \sum X_i \sum X_i^4 & -(\sum X_i^2)^2 + (n-m)\sum X_i^4 & \sum X_i \sum X_i^2 - (n-m)\sum X_i^3 \\ \sum X_i \sum X_i^3 - (\sum X_i^2)^2 & \sum X_i \sum X_i^2 - (n-m)\sum X_i^3 & (n-m)\sum X_i^2 - (\sum X_i)^2 \end{bmatrix} \end{aligned}$$

where

$$\Delta_r = -(\sum X_i^2)^3 + 2\sum X_i \sum X_i^2 \sum X_i^3 - (\sum X_i)^2 \sum X_i^4 - (n-m)(\sum X_i^3)^2 + (n-m)\sum X_i^2 \sum X_i^4$$

Define

$$\begin{aligned} G_1(d) &= \int_d^1 x f_X(x)dx, \qquad G_2(d) = \int_d^1 x^2 f_X(x)dx, \\ G_3(d) &= \int_d^1 x^3 f_X(x)dx, \qquad G_4(d) = \int_d^1 x^4 f_X(x)dx \end{aligned}$$

and

$$\begin{aligned} & \frac{1}{(n-m)^3}\Delta_r \\ \longrightarrow \quad & \frac{-G_2^3(d^0)}{(1-F_X(d^0))^3} + \frac{2G_1(d^0)G_2(d^0)G_3(d^0)}{(1-F_X(d^0))^3} - \frac{G_1^2(d^0)G_4(d^0)}{(1-F_X(d^0))^3} - \frac{G_3^2(d^0)}{(1-F_X(d^0))^2} + \frac{G_2(d^0)G_4(d^0)}{(1-F_X(d^0))^2} \\ \triangleq \quad & \delta_{qr}(d^0) \end{aligned}$$

Then

$$Var(\sqrt{n}(\hat{\beta}-\beta^0)) = \frac{1}{(1-F_X(d^0))^3\delta_{qr}(d^0)}V(d^0)\sigma^2$$

The elements of $V(d^0)$ are:

$$\begin{aligned} v_{11} &= G_2(d^0)G_4(d^0) - G_3^2(d^0) \\ v_{22} &= -G_2^2(d^0) + G_4(d^0)(1-F_X(d^0)) \\ v_{33} &= G_2(d^0)(1-F_X(d^0)) - G_1^2(d^0) \\ v_{12} &= v_{21} = G_2(d^0)G_3(d^0) - G_1(d^0)G_4(d^0) \\ v_{13} &= v_{31} = G_1(d^0)G_3(d^0) - G_2^2(d^0) \\ v_{23} &= v_{32} = G_1(d^0)G_2(d^0) - G_3(d^0)(1-F_X(d^0)) \end{aligned}$$

Then

$$\begin{aligned} h(\cdot) \quad = \quad & \frac{1}{F_X(d^0)\delta_{ql}(d^0)}\Big[w_{11}(d^0)d^0 + w_{12}(d^0)d^{0^2} \\ & +1/3(w_{22}(d^0)+2w_{13}(d^0))d^{0^3} + 1/2w_{23}(d^0)d^{0^4} + 1/5w_{33}(d^0)d^{0^5}\Big] \\ & +\frac{1}{(1-F_X(d^0))\delta_{qr}(d^0)}\Big[v_{11}(d^0)(1-d^0) + v_{12}(d^0)(1-d^{0^2}) \\ & +1/3(v_{22}(d^0)+2v_{13}(d^0))(1-d^{0^3}) + 1/2v_{23}(d^0)(1-d^{0^4}) \\ & +1/5v_{33}(d^0)(1-d^{0^5})\Big] \\ & +(SNR)^2\frac{1}{f_X(d^0)}EV \end{aligned}$$

where $w_{ij}(d^0)$ and $v_{ij}(d^0)$ are the (i,j)th element in $W(d^0)$ and $V(d^0)$, respectively.

By substituting d^0 for $\hat{d}_1$ from the initial samples, we have

$$
\begin{aligned}
H_1(\hat{d}_1) &= 1/2(\lambda_2 L_n^2 + \lambda_1(\hat{d}_1^2 - L_n^2)) \\
H_2(\hat{d}_1) &= 1/3(\lambda_2 L_n^3 + \lambda_1(\hat{d}_1^3 - L_n^3)) \\
H_3(\hat{d}_1) &= 1/4(\lambda_2 L_n^4 + \lambda_1(\hat{d}_1^4 - L_n^4)) \\
H_4(\hat{d}_1) &= 1/5(\lambda_2 L_n^5 + \lambda_1(\hat{d}_1^5 - L_n^5)) \\
G_1(\hat{d}_1) &= 1/2(\lambda_1(U_n^2 - \hat{d}_1^2) + (1 - \lambda_2 L_n - 2\lambda_1 c_n)(1 + U_n)) \\
G_2(\hat{d}_1) &= 1/3(\lambda_1(U_n^3 - \hat{d}_1^3) + (1 - \lambda_2 L_n - 2\lambda_1 c_n)(1 + U_n + U_n^2)) \\
G_3(\hat{d}_1) &= 1/4(\lambda_1(U_n^4 - \hat{d}_1^4) + (1 - \lambda_2 L_n - 2\lambda_1 c_n)(1 + U_n + U_n^2 + U_n^3)) \\
G_4(\hat{d}_1) &= 1/5(\lambda_1(U_n^5 - \hat{d}_1^5) + (1 - \lambda_2 L_n - 2\lambda_1 c_n)(1 + U_n + U_n^2 + U_n^3 + U_n^4))
\end{aligned}
$$

We seek to minimize the surrogate expression

$$
\begin{aligned}
h_{qq}(\cdot) &= \frac{1}{F_X(\hat{d}_1)\delta_{ql}(\hat{d}_1)} \Big[w_{11}(\hat{d}_1)\hat{d}_1 + w_{12}(\hat{d}_1)\hat{d}_1^2 \\
&\quad + 1/3(w_{22}(\hat{d}_1) + 2w_{13}(\hat{d}_1))\hat{d}_1^3 + 1/2 w_{23}(\hat{d}_1)\hat{d}_1^4 + 1/5 w_{33}(\hat{d}_1)\hat{d}_1^5 \Big] \\
&\quad + \frac{1}{(1 - F_X(\hat{d}_1))\delta_{qr}(\hat{d}_1)} \Big[v_{11}(\hat{d}_1)(1 - \hat{d}_1) + v_{12}(\hat{d}_1)(1 - \hat{d}_1^2) \\
&\quad + 1/3(v_{22}(\hat{d}_1) + 2v_{13}(\hat{d}_1))(1 - \hat{d}_1^3) + 1/2 v_{23}(\hat{d}_1)(1 - \hat{d}_1^4) \\
&\quad + 1/5 v_{33}(\hat{d}_1)(1 - \hat{d}_1^5) \Big] \\
&\quad + \widehat{(SNR)}^2 \frac{1}{f_X(\hat{d}_1)} \widehat{EV}
\end{aligned}
$$

4.1.5 Other Models

In some practical problems, we need to consider models with polynomials of different degrees on either side of the change point. The corresponding optimization problems follow as special cases of the general parametric model. In this section, we provide expressions for the objective functions corresponding to the following models: constant-linear, linear-constant, linear-quadratic, and quadratic-linear. v_{ij}'s and w_{ij}'s have the same connotations as before.

1. Constant–Linear Model

$$
\begin{aligned}
h_{cl}(\cdot) &= \frac{\hat{d}_1}{F_X(\hat{d}_1)} + \frac{(1 - \hat{d}_1)}{(1 - F_X(\hat{d}_1))^2 \delta_2} G_2(\hat{d}_1) + \frac{(1 - \hat{d}_1^3)}{3(1 - F_X(\hat{d}_1))\delta_2} \\
&\quad - \frac{(1 - \hat{d}_1^2)}{(1 - F_X(\hat{d}_1))^2 \delta_2} G_1(\hat{d}_1) + \widehat{(SNR)}^2 \frac{1}{f_X(\hat{d}_1)} \widehat{EV}
\end{aligned}
$$

2. Constant–Quadratic Model

$$\begin{aligned} h_{cq}(\cdot) &= \frac{\hat{d}_1}{F_X(\hat{d}_1)} \\ &+\frac{1}{(1-F_X(\hat{d}_1))\delta_1(\hat{d}_1)}\Big[v_{11}(\hat{d}_1)(1-\hat{d}_1)+v_{12}(\hat{d}_1)(1-\hat{d}_1^2) \\ &+1/3(v_{22}(\hat{d}_1)+2v_{13}(\hat{d}_1))(1-\hat{d}_1^3)+1/2v_{23}(\hat{d}_1)(1-\hat{d}_1^4)+1/5v_{33}(\hat{d}_1)(1-\hat{d}_1^5)\Big] \\ &+\widehat{(SNR)}^2\frac{1}{f_X(\hat{d}_1)}\widehat{EV} \end{aligned}$$

3. Linear–Quadratic Model

$$\begin{aligned} h_{lq}(\cdot) &= \frac{\hat{d}_1}{F_X^2(\hat{d}_1)\delta_1}H_2(\hat{d}_1)+\frac{\hat{d}_1^3}{3F_X(\hat{d}_1)\delta_1}-\frac{\hat{d}_1^2}{F_X(\hat{d}_1)^2\delta_1}H_1(\hat{d}_1) \\ &+\frac{1}{(1-F_X(\hat{d}_1))\delta_{qr}(\hat{d}_1)}\Big[v_{11}(\hat{d}_1)(1-\hat{d}_1)+v_{12}(\hat{d}_1)(1-\hat{d}_1^2) \\ &+1/3(v_{22}(\hat{d}_1)+2v_{13}(\hat{d}_1))(1-\hat{d}_1^3)+1/2v_{23}(\hat{d}_1)(1-\hat{d}_1^4)+1/5v_{33}(\hat{d}_1)(1-\hat{d}_1^5)\Big] \\ &+\widehat{(SNR)}^2\frac{1}{f_X(\hat{d}_1)}\widehat{EV} \end{aligned}$$

4. Quadratic–Linear Model

$$\begin{aligned} h_{ql}(\cdot) &= \frac{1}{F_X(\hat{d}_1)\delta_{ql}(\hat{d}_1)}\Big[w_{11}(\hat{d}_1)\hat{d}_1+w_{12}(\hat{d}_1)\hat{d}_1^2 \\ &+1/3(w_{22}(\hat{d}_1)+2w_{13}(\hat{d}_1))\hat{d}_1^3+1/2w_{23}(\hat{d}_1)\hat{d}_1^4+1/5w_{33}(\hat{d}_1)\hat{d}_1^5\Big] \\ &+\frac{(1-\hat{d}_1)}{(1-F_X(\hat{d}_1))^2\delta_2}G_2(\hat{d}_1)+\frac{(1-\hat{d}_1^3)}{3(1-F_X(\hat{d}_1))\delta_2}-\frac{(1-\hat{d}_1^2)}{(1-F_X(\hat{d}_1))^2\delta_2}G_1(\hat{d}_1) \\ &+\widehat{(SNR)}^2\frac{1}{f_X(\hat{d}_1)}\widehat{EV} \end{aligned}$$

4.1.6 Numerical Results of Optimal Allocations

In Table 4.2-4.3, we provide optimal allocations in the left and middle parts for five models, five locations of the change point and four values of SNR with $n_1 = n_2 = 500$ and $n_1 = n_2 = 100$. The allocation is not monotonically increasing or decreasing with SNR. If we fix SNR and the location of the change point, the allocation in the middle part is decreasing as the complexity of the model increases. It is reasonable, since we expect to have more samples from the outside of the neighborhood to improve the performance of the regression function.

SNR	d^0	C-C		C-L		L-L		L-Q		Q-Q	
		left	mid.	left	mid.	left	mid.	left	mid.	left	mid.
2	0.2	0	.7725	0	.4934	.2079	.3295	.0739	.2599	.2342	.2551
	0.4	0	.9628	.1010	.4740	.3003	.3216	.1159	.2610	.3363	.2440
	0.5	0	1	.1471	.4660	.3395	.3210	.1372	.2664	.3785	.2429
	0.6	.0372	.9628	.1918	.4613	.3781	.3216	.1612	.2755	.4198	.2440
	0.8	.2274	.7726	.2884	.4638	.4632	.3287	.2323	.3265	.5096	.2567
5	0.2	0	.7685	.0001	.4868	.2188	.3126	.0770	.2468	.2490	.2351
	0.4	0	.9602	.1093	.4581	.3080	.3072	.1168	.2478	.3452	.2281
	0.5	0	1	.1556	.4496	.3467	.3067	.1373	.2523	.3863	.2274
	0.6	.0398	.9602	.2008	.4436	.3848	.3072	.1603	.2598	.4268	.2281
	0.8	.2315	.7685	.2984	.4410	.4686	.3126	.2216	.3310	.5167	.2341
8	0.2	0	.8560	0	.5285	.1991	.3808	.0800	.3203	.2304	.2957
	0.4	0	1	.0274	.5761	.2796	.3727	.1210	.3231	.3194	.2866
	0.5	0	1	.0697	.5686	.3141	.3718	.1415	.3287	.3572	.2857
	0.6	0	1	.1099	.5648	.3478	.3727	.1642	.3376	.3940	.2865
	0.8	.1440	.8560	.1924	.5700	.4201	.3808	.2246	.3378	.4844	.2861
15	0.2	0	1	0	.6284	.1583	.5137	.0759	.4615	.1893	.4233
	0.4	0	1	0	.6788	.2222	.5031	.1137	.4663	.2638	.4114
	0.5	0	1	0	.7058	.2492	.5015	.1318	.4732	.2949	.4102
	0.6	0	1	0	.7351	.2750	.5026	.1509	.4833	.3248	.4114
	0.8	0	1	0	.8077	.3281	.5136	.1994	.5182	.3877	.4217

Table 4.2: Optimal allocations, $n_1 = n_2 = 500$.

4.2 Optimal λ for Estimating Regression Function

Recall the adaptive two-stage procedure we introduced in Chapter 3. The updated estimate of the change–point is computed by minimizing

$$\mathbb{P}_{n_2}[\{(w - \psi_l(\hat{\beta}_{l,n_1}, u))^2 I(u \leq d) + (w - \psi_u(\hat{\beta}_{u,n_1}, u))^2 I(u > d)]$$

We are interested in selecting the allocation parameter λ to improve the model's performance as a whole. We naturally want to focus attention on the quality of the estimated regression parameters from the first stage as well as the change point estimator from the second stage. Therefore, another optimal selection criterion is defined. Moreover, optimization problems are derived to the piecewise–constant model, the linear–linear model and the linear–quadratic model and are extended to the general parametric model.

SNR	d^0	C-C		C-L		L-L		L-Q		Q-Q	
		left	mid.	left	mid.	left	mid.	left	mid.	left	mid.
2	0.2	0	.9951	0	.6220	.0852	.5676	.0237	.4982	.0952	.5036
	0.4	0	1	0	.6868	.2010	.5253	.0988	.4894	.2190	.4698
	0.5	0	1	0	.7231	.2390	.5219	.1260	.5017	.2672	.4656
	0.6	0	1	0	.7653	.2738	.5253	.1542	.5231	.3110	.4700
	0.8	.0049	.9951	0	.9089	.3471	.5677	.2262	.6187	.4012	.5036
5	0.2	0	.9864	0	.5966	.1567	.4838	.0697	.4248	.1757	.4045
	0.4	0	1	0	.6571	.2347	.4691	.1145	.4279	.2724	.3833
	0.5	0	1	0	.6889	.2662	.4677	.1354	.4356	.3093	.3813
	0.6	0	1	0	.7234	.2962	.4691	.1578	.4478	.3443	.3833
	0.8	.0136	.9864	.0500	.7622	.3595	.4838	.2204	.4957	.4168	.3991
8	0.2	0	1	0	.6663	.1328	.5686	.0653	.5171	.1512	.4889
	0.4	0	1	0	.7116	.1979	.5537	.1055	.5209	.2355	.4674
	0.5	0	1	0	.7367	.2239	.5523	.1237	.5289	.2673	.4654
	0.6	0	1	0	.7647	.2484	.5537	.1426	.5409	.2971	.4674
	0.8	0	1	0	.8390	.2986	.5686	.1931	.5835	.3601	.4884
15	0.2	0	1	0	.7777	.0901	.7056	.0497	.6659	.1026	.6390
	0.4	0	1	0	.8062	.1369	.6915	.0812	.6694	.1685	.6171
	0.5	0	1	0	.8226	.1549	.6901	.0943	.6767	.1925	.6150
	0.6	0	1	0	.8414	.1716	.6915	.1074	.6878	.2145	.6170
	0.8	0	1	0	.8934	.2043	.7056	.1387	.7282	.2578	.6385

Table 4.3: Optimal allocations, $n_1 = n_2 = 100$.

4.2.1 Piecewise–Constant Model

We start with a simple step regression model and define the expected L_2 error $E\tilde{L}_2$ as

$$
\begin{aligned}
& E\int[[\hat{\alpha}_{n_1}I\{x \le \hat{d}_{n_2}\} + \hat{\beta}_{n_1}I\{x > \hat{d}_{n_2}\} - \alpha_0 I\{x \le d^0\} - \beta_0 I\{x > d^0\}]^2]dx \\
= \; & E\int[I\{\hat{d}_{n_2} \le d^0\}[(\hat{\alpha}_{n_1} - \alpha_0)^2\hat{d}_{n_2} + (\hat{\beta}_{n_1} - \beta_0)^2(1 - d^0) + (\hat{\beta}_{n_1} - \alpha_0)^2(d^0 - \hat{d}_{n_2})] \\
& + I\{\hat{d}_{n_2} > d^0\}[(\hat{\alpha}_{n_1} - \alpha_0)^2 d_0 + (\hat{\beta}_{n_1} - \beta_0)^2(1 - \hat{d}_{n_2}) + (\beta_0 - \hat{\alpha}_{n_1})^2(\hat{d}_{n_2} - d^0)]]dx \\
= \; & \frac{\sigma^2}{n^{1+\gamma}}E\Bigg[I\{\hat{d}_{n_2} \le d^0\}\Bigg[\frac{n(\hat{\alpha}_{n_1} - \alpha_0)^2}{\sigma^2}n^\gamma\hat{d}_{n_2} + \frac{n(\hat{\beta}_{n_1} - \beta_0)^2}{\sigma^2}n^\gamma(1 - d^0) \\
& \qquad + (\frac{\hat{\beta}_{n_1} - \alpha_0}{\sigma})^2 n^{1+\gamma}(d^0 - \hat{d}_{n_2})\Bigg] \\
& \quad + I\{\hat{d}_{n_2} \le d^0\}\Bigg[\frac{n(\hat{\alpha}_{n_1} - \alpha_0)^2}{\sigma^2}n^\gamma d^0 + \frac{n(\hat{\beta}_{n_1} - \beta_0)^2}{\sigma^2}n^\gamma(1 - \hat{d}_{n_2}) \\
& \qquad + (\frac{\beta_0 - \hat{\alpha}_{n_1}}{\sigma})^2 n^{1+\gamma}(\hat{d}_{n_2} - d^0)\Bigg]\Bigg]
\end{aligned}
$$

Note that $(\hat{\alpha}_{n_1} - \alpha_0)^2(\hat{d}_{n_2} - d^0)$, $(\hat{\beta}_{n_1} - \alpha_0)^2(\hat{d}_{n_2} - d^0)$, $(\hat{\beta}_{n_1} - \alpha_0)^2(\hat{d}_{n_2} - d^0)$ and $(\beta_0 - \hat{\alpha}_{n_1})^2(\hat{d}_{n_2} - d^0)$ have faster rate of convergence than $1/n^2$, since the estimate

from the second stage has an accelerated rate of convergence compared with the classical procedure, which is one of our main results in Chapter 3. Then

$$\begin{aligned} E\tilde{L}_2 &= \frac{\sigma^2}{n^{1+\gamma}} E\left[\frac{n(\hat{\alpha}-\alpha_0)^2}{\sigma^2} n^{\gamma} d^0 + \frac{n(\hat{\beta}-\beta_0)^2}{\sigma^2} n^{\gamma}(1-d^0)\right. \\ &\qquad \left. +(\frac{\beta_0-\alpha_0}{\sigma})^2 n^{1+\gamma} |\hat{d}_{n_2} - d^0|\right] + o(1/n^{1+\gamma}) \end{aligned}$$

Again, by asymptotic normalities of $\sqrt{n}(\hat{\alpha}-\alpha_0)/\sigma$ and $\sqrt{n}(\hat{\beta}-\beta_0)/\sigma$, we obtain the asymptotic expression of the expected L_2 error as following:

$$\begin{aligned} E\tilde{L}_2 &= \frac{\sigma^2}{n^{1+\gamma}} \left[\frac{1}{\lambda}\left(\frac{n^{\gamma} d^0}{F_X(d^0)} + \frac{n^{\gamma}(1-d^0)}{1-F_X(d^0)}\right) + (SNR)^2 (\frac{1-\lambda}{\lambda})^{\gamma} 2K \frac{1}{(1-\lambda)^{1+\gamma}} E(V)\right] \\ &\quad +o(1/n^{1+\gamma}) \\ &= \frac{\sigma^2}{n^{1+\gamma}} \left[\frac{1}{\lambda}\left(\frac{n^{\gamma} d^0}{F_X(d^0)} + \frac{n^{\gamma}(1-d^0)}{1-F_X(d^0)}\right) + (SNR)^2 \frac{2K}{\lambda^{\gamma}(1-\lambda)} E(V)\right] + o(1/n^{1+\gamma}) \end{aligned}$$

Recalling the adaptive strategies we developed in Section 3.2, take $K = C_{\zeta} n_1^{-(1-\gamma)}$, where C_{ζ} is the ζ^{th} upper quantile of $\mathbb{M}_{SNR,Z,1}$. We usually select very small C_{ζ} to make the 'zoom-in' neighborhood contain d^0 with probability close to 1 with increasing n. Then, we have

$$\begin{aligned} E\tilde{L}_2 &= \frac{\sigma^2}{n^{1+\gamma}} \left[\frac{1}{\lambda}\left(\frac{n^{\gamma} d^0}{F_X(d^0)} + \frac{n^{\gamma}(1-d^0)}{1-F_X(d^0)}\right) + (SNR)^2 \frac{2C_{\zeta} n_1^{-(1-\gamma)}}{\lambda^{\gamma}(1-\lambda)} E(V)\right] + o(1/n^{1+\gamma}) \\ &= \frac{\sigma^2}{n^{1+\gamma}} \left[\frac{1}{\lambda}\left(\frac{n^{\gamma} d^0}{F_X(d^0)} + \frac{n^{\gamma}(1-d^0)}{1-F_X(d^0)}\right) + (SNR)^2 \frac{2C_{\zeta} n^{\gamma}}{\lambda(1-\lambda)n} E(V)\right] + o(1/n^{1+\gamma}) \\ &= \frac{\sigma^2}{n} \left[\frac{1}{\lambda}\left(\frac{d^0}{F_X(d^0)} + \frac{(1-d^0)}{1-F_X(d^0)}\right) + (SNR)^2 \frac{2C_{\zeta}}{\lambda(1-\lambda)n} E(V)\right] + o(1/n^{1+\gamma}) \\ &= \frac{\sigma^2}{n} \left[\frac{2}{\lambda} + (SNR)^2 \frac{2C_{\zeta}}{\lambda(1-\lambda)n} E(V)\right] + o(1/n^{1+\gamma}), \quad (\text{since } F_X(d^0) = d^0) \end{aligned}$$

where V follows the same definition as at the beginning of this chapter. What we want is to minimize

$$\left[\frac{2}{\lambda} + (SNR)^2 \frac{2C_{\zeta}}{\lambda(1-\lambda)n} E(V)\right]$$

under constraint:

$$0 < \lambda < 1$$

Note that the optimization problem above only depends on SNR and sample size n. It is not related to the location of change point.

SNR	n=500	n=250	n=100
2	.7555	.6967	.6221
5	.7594	.7007	.6257
15	.5944	.5565	.5258

Table 4.4: Optimal selection for allocation parameter λ, piecewise-constant model

4.2.2 Linear-Linear Model

Define

$$\begin{aligned} E\tilde{L}_{2ll,n} &= E\int[(\hat{\alpha}_{n_1,1}+\hat{\alpha}_{n_1,2}x)I\{x\le\hat{d}_{n_2}\}+(\hat{\beta}_{n_1,1}+\hat{\beta}_{n_1,2}x)I\{x>\hat{d}_{n_2}\} \\ &\quad -(\alpha_1^0+\alpha_2^0x)I\{x\le d^0\}-(\beta_1^0+\beta_2^0x)I\{x>d^0\}]^2dx \end{aligned}$$

By the similar derivation as in Section 4.2.1 and the asymptotic normalities of regression parameters, we have the asymptotic expression of the expected L_2 error as following

$$\begin{aligned} E\tilde{L}_{2ll,n} &= \left[\sum_{k=1}^{p}(V_{\beta_l}^{-1})_{kk}\Psi_{k,k}^l(d^0)+2\sum_{i<j}(V_{\beta_l}^{-1})_{ij}\Psi_{i,j}^l(d^0)\right. \\ &\quad \left.+\sum_{k=1}^{q}(V_{\beta_u}^{-1})_{kk}\Psi_{k,k}^u(d^0)+2\sum_{i<j}(V_{\beta_u}^{-1})_{kk}\Psi_{i,j}^u(d^0)\right] \\ &\quad +\frac{\sigma^2}{n}(SNR)^2\frac{2C_\zeta}{\lambda(1-\lambda)n}E(V) \end{aligned}$$

where V_{β_l} and V_{β_u} are equal to (4.4) and (4.5), respectively. And $\Psi_{ij}^l(d^0)$ and $\Psi_{ij}^u(d^0)$ are the same as in Section 4.1.3. Therefore, we have

$$E\tilde{L}_{2ll,n}\equiv M_{ll,n,1}^\star+M_{ll,n,2}^\star+M_{ll,n,3}^\star+o(1/n^{1+\gamma})$$

where

$$\begin{aligned} M_{ll,n,1}^\star &= \frac{\sigma^2}{n^{1+\gamma}}\frac{1}{\lambda}\left[\frac{n^\gamma d^0}{F_X(d^0)^2\delta_1}\int_0^{d^0}x^2f_X(x)dx+\frac{n^\gamma {d^0}^3}{3F_X(d^0)\delta_1}\right. \\ &\quad \left.-\frac{n^\gamma {d^0}^2}{F_X(d^0)^2\delta_1}\int_0^{d^0}xf_X(x)dx\right] \end{aligned}$$

$$
\begin{aligned}
M^{\star}_{ll,n,2} &= \frac{\sigma^2}{n^{1+\gamma}}\frac{1}{\lambda}\left[\frac{n^{\gamma}(1-d^0)}{(1-F_X(d^0))^2\delta_2}\int_{d^0}^{1}x^2 f_X(x)dx + \frac{n^{\gamma}(1-d^{0^3})}{3(1-F_X(d^0))^2\delta_2}\right. \\
&\quad \left. -\frac{n^{\gamma}(1-d^{0^2})}{(1-F_X(d^0))^2\delta_2}\int_{d^0}^{1}x f_X(x)dx\right] \\
M^{\star}_{ll,n,3} &= \frac{\sigma^2}{n^{1+\gamma}}(SNR)^2\frac{2K}{\lambda^{\gamma}(1-\lambda)}E(V)
\end{aligned}
$$

Then we have

$$
\begin{aligned}
\tilde{EL}_{2ll,n} &= \frac{\sigma^2}{n^{1+\gamma}}\frac{1}{\lambda}\left[\frac{n^{\gamma}d^0}{F_X(d^0)^2\delta_1}H_2(d^0) + \frac{n^{\gamma}d^{0^3}}{3\delta_1 F_X(d^0)} - \frac{n^{\gamma}d^{0^2}}{F_X(d^0)^2\delta_1}H_1(d^0)\right] \\
&\quad + \frac{\sigma^2}{n^{1+\gamma}}\frac{1}{\lambda}\left[\frac{n^{\gamma}(1-d^0)}{(1-F_X(d^0))^2\delta_2}G_2(d^0) + \frac{n^{\gamma}(1-d^{0^3})}{3\delta_2(1-F_X(d^0))}\right. \\
&\quad \left. -\frac{n^{\gamma}(1-d^{0^2})}{(1-F_X(d^0))^2\delta_2}G_1(d^0)\right] \\
&\quad + \frac{\sigma^2}{n^{1+\gamma}}(SNR)^2\frac{2K}{\lambda^{\gamma}(1-\lambda)}E(V)
\end{aligned}
$$

Since we know that

$$F_X(d^0) = d^0,$$

according to our two-stage procedure in which we select covariate values uniformly, we can obtain

$$
\begin{aligned}
&\tilde{EL}_{2ll,n} \\
&= \frac{\sigma^2}{n^{1+\gamma}}\frac{1}{\lambda}\left[\frac{n^{\gamma}d^{0^2}}{6\delta_1} + \frac{n^{\gamma}(1-d^0)^2}{6\delta_2}\right] + \frac{\sigma^2}{n^{1+\gamma}}(SNR)^2\frac{2K}{\lambda^{\gamma}(1-\lambda)}E(V) + o(\frac{1}{n^{1+\gamma}})
\end{aligned}
$$

Plugging in $K = 2C_{\zeta}n_1^{-(1-\gamma)}$ and $F_X(d^0) = d^0$, this expected L_2 error simplifies to

$$\frac{\sigma^2}{n}\left[\frac{4}{\lambda} + (SNR)^2\frac{2C_{\zeta}}{\lambda(1-\lambda)n}E(V)\right] \triangleq h^{\star}(\lambda).$$

From this expression, note that it only depends on SNR and sample size n, as shown through the results of our simulated studies (in Tabel 4.5). We draw a picture of the objective function for $n = 250$ and SNR=5 (Figure 4.8), the optimal solution is around .76, which is consistent with the result is Table 4.5.

SNR	n=500	n=250	n=100
2	.8097	.7555	.6777
5	.8131	.7594	.6817
15	.6448	.5944	.5471

Table 4.5: Optimal selection for allocation parameter λ, linear-linear model

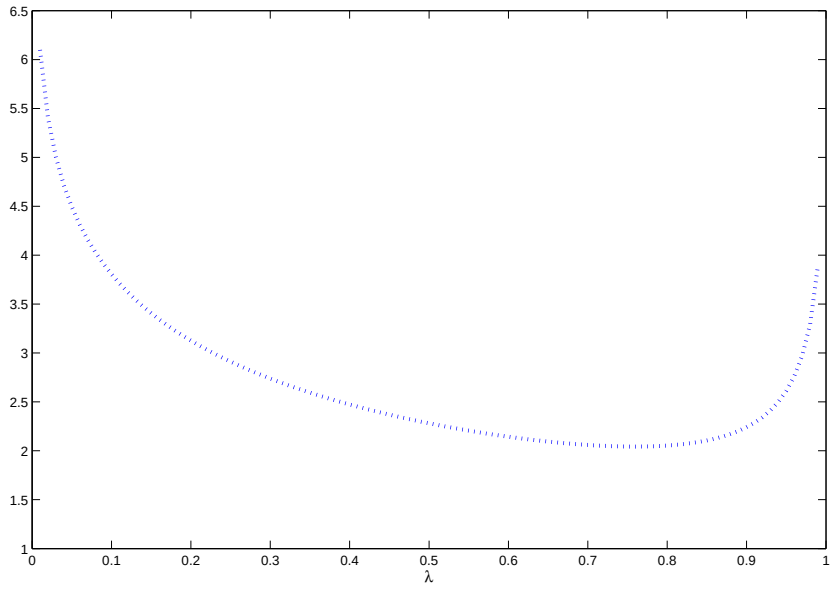


Figure 4.8: Linear-linear model, $log(h^\star) \sim \lambda$, SNR=5, n=250

4.2.3 Linear-Quadratic Model.

In a special case, consider, for example, a linear-quadratic model,

$$\begin{aligned} E\tilde{L}_{2lq,n}^{\star} &= \int [(\hat{\alpha}_{n_1,1} + \hat{\alpha}_{n_2,2}x)I\{x \leq \hat{d}_{n_2}\} + (\hat{\beta}_{n_1,1} + \hat{\beta}_{n_1,2}x + \hat{\beta}_{n_1,3}x^2)I\{x > \hat{d}_{n_2}\} \\ &\quad -(\alpha_1^0 + \alpha_2^0 x)I\{x \leq d^0\} - (\beta_1^0 + \beta_2^0 x + \beta_3^0 x^2)I\{x > d^0\}]^2 dx \\ &= h^\star(\lambda, n, d^0, SNR) + o(1/n^{1+\gamma}) \end{aligned}$$

where

$$h^\star(\lambda, n, d^0, SNR) = M_{lq,n,1}^\star + M_{lq,n,2}^\star + M_{lq,n,3}^\star$$

with

$$
\begin{aligned}
M^{\star}_{lq,n,1} &= M_{ll,1} = \frac{\sigma^2}{n^{1+\gamma}} \frac{1}{\lambda} \left[\frac{n^{\gamma} {d^0}^2}{6\delta_1} \right] = \frac{\sigma^2}{n} \frac{2}{\lambda} \\
M^{\star}_{lq,n,2} &= \frac{\sigma^2}{n^{1+\gamma}} \cdot \frac{n^{\gamma}}{(1 - F_X(d^0))\delta_{qr}(d^0)\lambda} \Big[v_{11}(d^0)(1-d^0) + v_{12}(d^0)(1-{d^0}^2) \\
&\quad +1/3(v_{22}(d^0) + 2v_{13}(d^0))(1-{d^0}^3) + 1/2 v_{23}(d^0)(1-{d^0}^4) + 1/5 v_{33}(d^0)(1-{d^0}^5) \Big] \\
&= \frac{\sigma^2}{n} \cdot \frac{1}{(1 - F_X(d^0))\delta_{qr}(d^0)\lambda} \Big[v_{11}(d^0)(1-d^0) + v_{12}(d^0)(1-{d^0}^2) \\
&\quad +1/3(v_{22}(d^0) + 2v_{13}(d^0))(1-{d^0}^3) + 1/2 v_{23}(d^0)(1-{d^0}^4) + 1/5 v_{33}(d^0)(1-{d^0}^5) \Big] \\
M_{lq,n,3\star} &= \frac{\sigma^2}{n^{1+\gamma}} (SNR)^2 \frac{2K}{\lambda^{\gamma}(1-\lambda)} E(V) \\
&= \frac{\sigma^2}{n} (SNR)^2 \frac{2C_{\zeta}}{\lambda(1-\lambda)n} E(V)
\end{aligned}
$$

where $v_{ij}(d^0)$ is the (i,j)th element in $V(d^0)$ from Section 4.1. The optimal selections for the allocation parameter λ are shown in the following tables (Table 4.6 and Table 4.7). The optimal solutions are depending on SNR, sample size n and the location of change point.

SNR	$d^0 = .2$	$d^0 = .4$	$d^0 = .5$	$d^0 = .6$	$d^0 = .8$
2	.8060	.8197	.8255	.8307	.8399
5	.8094	.8230	.8287	.8339	.8429
8	.7464	.7622	.7690	.7752	.7861
100	.5061	.5074	.5080	.5086	.5098

Table 4.6: Optimal selection of allocation parameter λ, n=500, linear-quadratic model

SNR	$d^0 = .2$	$d^0 = .4$	$d^0 = .5$	$d^0 = .6$	$d^0 = .8$
2	.6734	.6896	.6967	.7033	.7151
5	.6773	.6936	.7007	.7073	.7192
8	.6141	.6282	.6346	.6406	.6517
100	.5013	.5015	.5016	.5018	.5020

Table 4.7: Optimal selection of allocation parameter λ, n=100, linear-quadratic model

Tables 4.5- 4.7 show that the optimal λ's are usually bigger than half, which is reasonable, since we put more attention on estimating the regression function from the first stage. Therefore, we would like to put a relatively bigger amount of budget in the first stage.

4.2.4 General Parametric Model.

Now, let us think about extending to the general parametric model. Using the notation in Section 4.1.2, we seek to minimize

$$\begin{aligned} h^{\star}_{par} &= \frac{\sigma^2}{n\lambda}\left[\sum_{k=1}^{p}(V_{\beta_l}^{-1})_{kk}\Psi^{l}_{k,k}(d^0) + 2\sum_{i<j}(V_{\beta_l}^{-1})_{ij}\Psi^{l}_{i,j}(d^0)\right. \\ &\quad \left. + \sum_{k=1}^{q}(V_{\beta_u}^{-1})_{kk}\Psi^{u}_{k,k}(d^0) + 2\sum_{i<j}(V_{\beta_u}^{-1})_{kk}\Psi^{u}_{i,j}(d^0)\right] \\ &\quad + \frac{\sigma^2}{n}(SNR)^2\frac{2C_{\zeta}}{\lambda(1-\lambda)n}E(V) \end{aligned}$$

Note that our objective functions usually depend on d^0 and SNR, however, in practice, both of them are unknown. Therefore, we need use partial budget to conduct a one-stage procedure to obtain the estimate of change point as well as the estimate of SNR before searching for the optimal λ for estimating regression function.

4.3 Comparing the Two Allocation Strategies

To compare these two criterions, we select same total sample sizes $n = 100$ and $n = 500$ on the piecewise constant model, the linear-linear model, the linear-quadratic model and the quadratic-quadratic model. We use true d^0 and SNR in all simulation studies for comparison. In the first criterion, $n/2$ samples are used in each stage. We calculate those scaled expected L_2 errors of optimal selections from two versions as shown in Table 4.8-4.15. In general, the first criterion is better than the second one. But the second one is connected with our proposed two-stage procedure in Chapter 3, which gives us insight on how to allocate samples in each stage for estimating the regression function.

Remark: We observe that the optimal λ does not depends on the location of d^0 when the polynomial on either side of the change point has the same degree. It is not hard to obtain this conclusion on the piecewise-constant model and the linear-linear model as we derived above. Our simulations suggest that the quadratic-quadratic

model displays the same feature, but it does not seem easy to verify it algebraically.

SNR	$\frac{n}{\sigma^2}EL_2$ of 1st criterion	$\frac{n}{\sigma^2}EL_2$ of 2nd criterion
2	4.3385	8.1885
5	4.2311	7.9596
8	6.9606	13.6276
15	19.4233	38.7434
100	758.1298	1516.3

Table 4.8: Comparison of optimal selections, $n = 100$, piecewise-constant.

SNR	d^0	$\frac{n}{\sigma^2}EL_2$ of 1st criterion	$\frac{n}{\sigma^2}EL_2$ of 2nd criterion
2	.2	2.4121	3.9132
	.4	2.4677	
	.5	2.4677	
	.6	2.4677	
	.8	2.4121	
5	.2	2.3867	3.8547
	.4	2.4462	
	.5	2.4462	
	.6	2.4462	
	.8	2.3867	
8	.2	2.9885	5.2176
	.4	2.9921	
	.5	2.9921	
	.6	2.9921	
	.8	2.9885	
15	.2	5.4847	10.5917
	.4	5.4847	
	.5	5.4847	
	.6	5.4847	
	.8	5.4847	
100	.2	153.2260	306.4389
	.4	153.2260	
	.5	153.2260	
	.6	153.2260	
	.8	153.2260	

Table 4.9: Comparison of optimal selections, $n = 500$, piecewise-constant.

4.4 Data Application

Both criterions are used on the motivating application. The total budget was still set to $n = 70$. For the first criterion, we set $\lambda = .5$. In the first stage, 35 samples were obtained from a uniform distribution. SNR is estimated from the first stage, $\hat{SNR} = 5.23$. Using "fmincon" in Matlab, we found that the optimal allocation of samples for the second stage is $(.2032, .4946, .3022)$. The estimated change-point

SNR	d^0	$\frac{n}{\sigma^2}EL_2$ of 1st criterion	$\frac{n}{\sigma^2}EL_2$ of 2nd criterion
2	.2	6.7473	11.2555
	.4	7.6728	
	.5	7.7772	
	.6	7.6728	
	.8	6.7473	
5	.2	8.0664	711.0087
	.4	8.6654	
	.5	8.7257	
	.6	8.6654	
	.8	8.0664	
8	.2	12.4800	16.9786
	.4	13.2222	
	.5	13.2957	
	.6	13.2222	
	.8	12.4800	
15	.2	29.1884	42.4699
	.4	30.4860	
	.5	30.6068	
	.6	30.4860	
	.8	29.1884	
100	.2	809.8291	1520.25
	.4	824.1190	
	.5	825.3000	
	.6	824.1190	
	.8	809.8291	

Table 4.10: Comparison of optimal selections, $n = 100$, linear-linear.

from the second stage is $\hat{d}_{n_2} = .796$ with a 95% confidence interval $(.7889, .8027)$(see left panel of Figure 4.9). The depicted fitted regression models are based on the second stage estimates.

For the second criterion, we use the estimated SNR from the first criterion and set $\zeta = .0005$. The optimal allocation for λ is .6219. Therefore, 43 samples are obtained from a uniform distribution in the first stage. The remaining 27 samples from a uniform distribution on the "zoom-in" neighborhood gave an estimate $\hat{d}_{n_2} = .796$ with a 95% confidence interval $(.7890, .8037)$ (see right panel of Figure 4.9). The depicted fitted regression models are based on the first stage estimates.

SNR	d^0	$\frac{n}{\sigma^2}EL_2$ of 1st criterion	$\frac{n}{\sigma^2}EL_2$ of 2nd criterion
2	.2	4.8328	3.9132
	.4	5.2182	
	.5	5.2599	
	.6	5.2182	
	.8	4.8328	
5	.2	4.9020	3.8547
	.4	5.2461	
	.5	5.2844	
	.6	5.2461	
	.8	4.9020	
8	.2	6.2137	5.2176
	.4	6.5873	
	.5	6.6286	
	.6	6.5873	
	.8	6.2137	
15	.2	10.8609	10.5917
	.4	11.3384	
	.5	11.3908	
	.6	11.3384	
	.8	10.8609	
100	.2	185.7516	306.4389
	.4	187.7338	
	.5	187.9448	
	.6	187.7338	
	.8	187.7516	

Table 4.11: Comparison of optimal selections, $n = 500$, linear-linear.

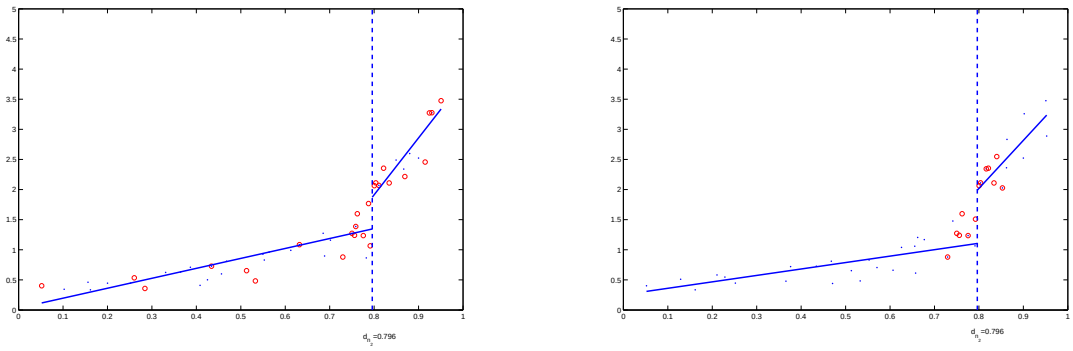


Figure 4.9: Sampled points (from 1st stage solid circles and from 2nd stage open circles) together with the fitted parametric models and estimated change point, based on a total budget of $n = 70$ points, obtained from the two-stage adaptive procedure with sampling from a non-uniform distribution in the second stages (left panel) and from optimal allocation for λ (right panel).

SNR	d^0	$\frac{n}{\sigma^2}EL_2$ of 1st criterion	$\frac{n}{\sigma^2}EL_2$ of 2nd criterion
2	.2	8.0919	10.9594
	.4	8.8326	12.1330
	.5	8.8035	12.7101
	.6	8.5686	13.2815
	.8	7.3047	14.4093
5	.2	9.8618	10.7144
	.4	10.2252	11.8811
	.5	10.1099	12.4549
	.6	9.8301	13.0230
	.8	8.6572	14.1444
8	.2	14.7650	16.6541
	.4	15.1831	17.9417
	.5	15.0232	18.5752
	.6	14.6583	19.2026
	.8	13.1672	20.4405
15	.2	32.6919	42.1037
	.4	33.3915	43.5609
	.5	33.0948	44.2821
	.6	32.4626	44.9986
	.8	30.1652	46.4185
100	.2	824.9265	1519.9
	.4	834.7703	1521.4
	.5	833.8909	1522.2
	.6	831.5467	1523.0
	.8	829.7139	1524.6

Table 4.12: Comparison of optimal selections, $n = 100$, linear-quadratic.

SNR	d^0	$\frac{n}{\sigma^2}EL_2$ of 1st criterion	$\frac{n}{\sigma^2}EL_2$ of 2nd criterion
2	.2	6.1935	6.2102
	.4	6.4377	7.1941
	.5	6.3727	7.0803
	.6	6.1994	8.1633
	.8	5.4374	9.1209
5	.2	6.2815	6.1410
	.4	6.4910	7.1209
	.5	6.4299	7.6052
	.6	6.2714	8.0863
	.8	5.6030	9.0404
8	.2	7.8225	7.7118
	.4	8.0313	8.7720
	.5	7.9560	9.2944
	.6	7.7748	9.8125
	.8	7.0225	10.8372
15	.2	13.0989	13.4997
	.4	13.3248	14.7331
	.5	13.2102	15.3396
	.6	12.9612	15.9400
	.8	11.9539	17.1246
100	.2	195.5964	310.01
	.4	196.2392	311.58
	.5	195.6492	312.37
	.6	194.5099	313.16
	.8	189.9790	314.73

Table 4.13: Comparison of optimal selections, $n = 500$, linear-quadratic.

SNR	d^0	$\frac{n}{\sigma^2}EL_2$ of 1st criterion	$\frac{n}{\sigma^2}EL_2$ of 2nd criterion
2	.2	9.0405	14.1291
	.4	9.7763	
	.5	9.8758	
	.6	9.7763	
	.8	9.0405	
5	.2	10.5164	13.8658
	.4	11.4877	
	.5	11.5959	
	.6	11.4877	
	.8	10.5164	
8	.2	15.5094	20.1330
	.4	16.7195	
	.5	16.8560	
	.6	16.7195	
	.8	15.5094	
15	.2	33.7251	46.0651
	.4	35.4574	
	.5	35.6783	
	.6	35.4574	
	.8	33.7251	
100	.2	838.9074	1524.2
	.4	842.2476	
	.5	842.5248	
	.6	842.2476	
	.8	838.9149	

Table 4.14: Comparison of optimal selections, $n = 100$, quadratic-quadratic.

SNR	d^0	$\frac{n}{\sigma^2}EL_2$ of 1st criterion	$\frac{n}{\sigma^2}EL_2$ of 2nd criterion
2	.2	6.8773	8.8825
	.4	7.5212	
	.5	7.5900	
	.6	7.5212	
	.8	6.8773	
5	.2	7.0656	8.8028
	.4	7.6207	
	.5	7.6819	
	.6	7.6207	
	.8	7.0664	
8	.2	8.7246	10.5823
	.4	9.3332	
	.5	9.4000	
	.6	9.3332	
	.8	8.7271	
15	.2	14.2936	16.8303
	.4	15.0776	
	.5	15.1625	
	.6	15.0776	
	.8	14.2947	
100	.2	199.7810	314.3374
	.4	203.1658	
	.5	203.5084	
	.6	203.1658	
	.8	199.7828	

Table 4.15: Comparison of optimal selections, $n = 500$, quadratic-quadratic.

CHAPTER 5

Application to Jump Boundary Curve Detection

5.1 Jump Boundary Curve Detection Problem Formulation and Literature Review

In this chapter, we turn our focus on estimating boundaries in high dimensional data. Specially, the underlying response surface arises from data (y_i, x_{1i}, x_{2i}), with y being the response variable. The model is given by

$$y_i = f_1(x_{1i}, x_{2i})I_{x_{2i} \geq g(x_{1i})} + f_2(x_{1i}, x_{2i})I_{x_{2i} < g(x_{1i})} + \varepsilon_i$$

where y is the response surface defined by 2-dimensional step functions, $f_1(x_1, x_2)$ and $f_2(x_1, x_2)$, respectively, with $g(\cdot)$ the jump boundary curve we are interested in identifying and estimating and ε a mean zero, homoscedastic error term. The main motivation of the proposed approach is to identify and *model* the jump boundary curve in an as computationally efficient manner as possible.

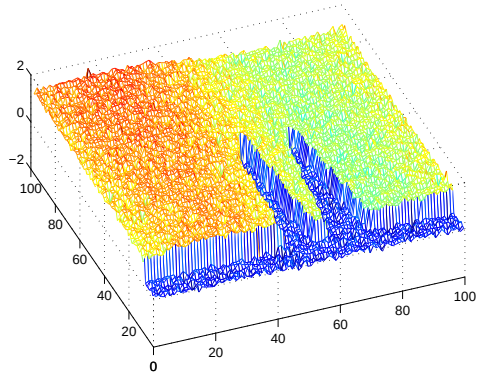


Figure 5.1: Three-dimension plot of Example 1

Budget constraints dictate that a total of $N = n_1 \times n_2$ samples can be obtained, where n_1 is the number of points to be allocated along the X_1 direction and n_2 along the X_2 direction for each selected value of X_1. We start by examining the case where all the budget is used by adopting appropriately the multi-stage procedures to the present setting. We then briefly discuss a strategy that economizes on points to be allocated.

5.1.1 Literature Review

Qiu (2002) presented a jump detection procedure based on local smoothing techniques. We will review more about it later when comparing it with our methodology.

Chu et al. (1998) discuss an approach to a problem called "sigma filter" and proposed an improvement based on running M estimation. It is about edge-preserving smoothers problem in image processing. Classical smoothing is inappropriate when the image has jumps or edges between regions, because smoothing tends to blur the edge. Therefore, an M estimator is defined and is used to recover the target curve. The idea is developed in the context of one-dimensional nonparametric regression with jumps. Consider data of the form

$$Y_i = m(x_i) + \varepsilon_i, \quad i = 1, \ldots, n$$

where $m(x)$ is the target curve which is smooth except for some jump continuities, x_i are assumed equally spaced on the unit interval, and ε_i are independent, identically distributed random variables with mean zero. To get an edge-preserving smoother, Chu et al. (1998) defined $\hat{m}_M(x_i)$ to be the local minimizer of the M function $S(\theta)$. The similarity of this problem to the boundary jump detection one allows one to use this methodology for the latter problem as well.

Hall and Molchanov (2003) study the problem under consideration and propose a methodology based on "sequential refinement with reassessment"; the main idea is that at every step of the procedure there is an assesment of the correctness of

each sequential result, together with a reappraisal of previous results in case the proposed test statistic comes up significant. They focus on a univariate problem, but also discuss how the 2-dimensional case can be handled through a combination of univariate results. Specifically, they assume that the function f is defined on an interval l, and has a jump discountinuity at a point γ in the interior of l. For differentiable functions f_1 and f_2 we have that

$$y = f(x) = f(x|\gamma) = f_1(x) + f_2(x)I_{\{x>\gamma\}},$$

where f can be observed at points $x = x_i \in l$, subject to error: $y_i = f(x_i) + \varepsilon_i$, where the design points x_i are open to sequential choice and the errors ε_i are independent and identically distributed with zero mean. A recursive method for estimating γ was introduced in this paper.

In practice this technique would be applied only after a "pilot" estimator, $\tilde{\gamma}$, had been constructed using a portion of the permitted sample size, n. This would lead to a preliminary interval l_1, a strict subset of l, in which the first estimator in the recursion would be constructed, using m design points $x_1 < \ldots < x_m$ equally spaced on l_1. l_1 is the first of a sequence of confidence sets for the true value of γ. At the kth stage of the algorithm l_k shall be determined. Assume $n = lm$, where l, m are positive integers. Each sequential sample will be of size m, and there will be l stages in the algorithm. In the first stage, distribute m equally spaced points on the first interval l_1 and sample f at those spaces. Under the temporary assumption that the data are Normally distributed with known variance, compute the statistic $T(\gamma)$ associated with a likelihood ratio test of the null hypothesis that f is constant on l_1, against the alternative that f takes different but constant values on either side of γ. Take $\hat{\gamma_1}$ to be that value of γ, chosen from among the m design points, that gives an extremum for the test. For the univariate problem, the splitting point (γ) corresponds to the the jump boundary. In the spatial case, they strike an arc

of some radius (related to the property of the jump boundary curve)and centered at the current point, across the tangent approximation to the curve in the direction of travel. Then the next estimate was found by applying the sequential refinement with reassessment (SRR) method to the one-dimensional problem on the arc. From the first two estimates of points on the curve one may obtain an approximation to the tangent. Each subsequential estimate was computed by striking an arc across the most recent tangent estimate and solving the one-dimensional splitting-point problem on the arc, using SRR method.

5.2 Description of Multi-stage Procedure

It is assumed that due to prior information the boundary curve can be parameterized as a function of x_1, i.e. the boundary is given by $x_2 = g(x_1) + \mu$, where μ is a homoscedastic error term. It is further assume that $g(\cdot)$ is a smooth function of x_1. The proposed procedure is summarized next. (i) Select $(x_{1,1}, x_{1,2}, \cdots, x_{1,n_1})$ via uniform design on X_1 direction. (ii) For each fixed $x_{1,i}$, we estimate the change point via our two-stage procedure on X_2 direction with respect to the surface response values, and denote it as b_i. (iii) A non-parametric model (or spline) $\hat{g}(x_1)$ will be fitted by use those $(x_{1,i}, b_i)$. Pointwise confidence intervals can be constructed as in Section 3.3.

Given the results established so far following holds

Proposition 5.1. *$\hat{g}(x_{1,i})$ converges pointwisely to $g(x_1)$ as $n_1 \to \infty$ and $n_2 \to \infty$.*

We illustrate next the performance of the procedure through a number of examples.

- Example 1:

$$g(x_1) = 0.1 + 0.5e^{10(14x_1 - 7.2)^2} + 0.5e^{-4(14x_1 - 9.4)^2}$$

- Example 2:

$$g(x_1) = \begin{cases} 0.05 + 1.2x_1 & \text{if } x_1 \in [0, 0.25) \\ 4.1 - 15x_1 & \text{if } x_1 \in [0.25, 0.26) \\ -0.73 + 3.6x_1 & \text{if } x_1 \in [0.26, 0.4) \\ 0.9 - 0.5x_1 & \text{if } x_1 \in [0.4, 0.6) \\ 0.075 + 0.87x_1 & \text{if } x_1 \in [0/6, 1] \end{cases}$$

- Example 3:

$$g(x_1) = 0.25\sin(10\pi x_1) + 0.5$$

- Example 4 (same as in Qiu (2002)):

$$g(x_1) = 0.6\sin(\pi x_1) + 0.2$$

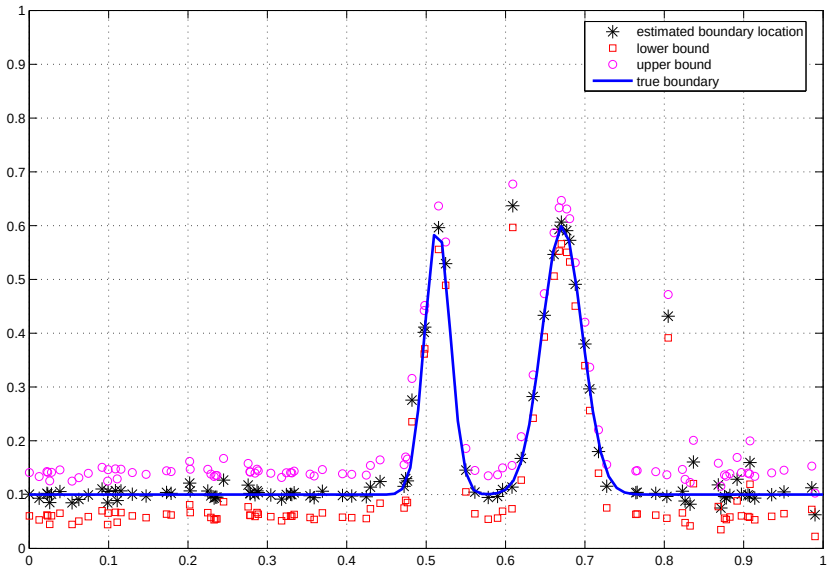


Figure 5.2: Jump boundary detection from One-stage procedure, Example 1

The results including estimated boundary locations and pointwise confidence intervals are given in Figure 5.2-5.5.Suppose we can obtain 100×100 samples for each example. When we estimate the location of change point on X_2 direction, we use

uniform design to obtain the first stage samples. The reason is: when the sample size is small, sampling from uniform distribution may not guarantee that there exist samples on both sides of the true change point. Otherwise, it would lead to a big bias for estimation. In example 1 as shown in Figure 5.2, for this single simulation, $\frac{1}{100}\sum_{i=1}^{100}(g(x_{1,i}) - \hat{g}(x_{1,i}))^2 = .00064$

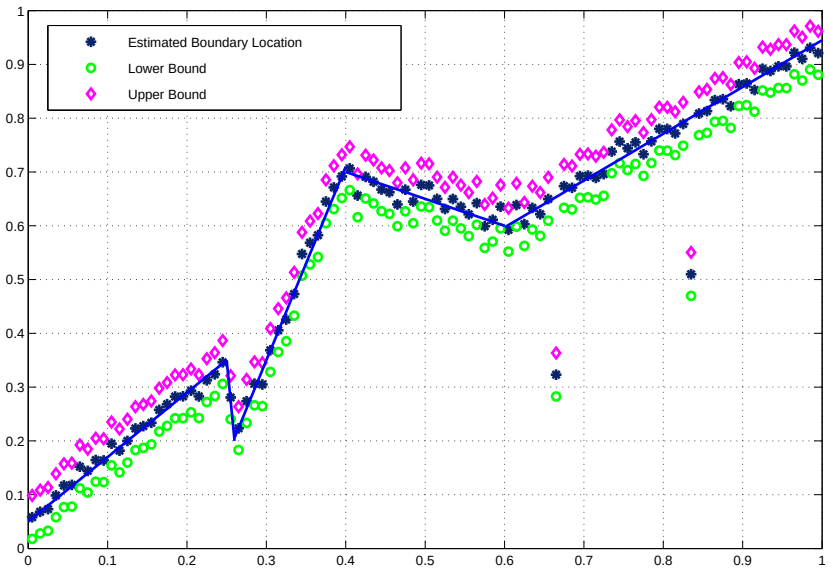


Figure 5.3: Jump boundary detection from One-stage procedure, Example 2

5.3 Adaptive Procedure for X_1

The assumed smoothness of the boundary curve $g(\cdot)$ suggests that one could save points in the budget by reducing the sampling density along the X_1 axis. We discuss next a qualitative methodology for selection of x_1 values. The adaptive procedure will be illustrated by using the first simulation example discussed in Section 5.1.

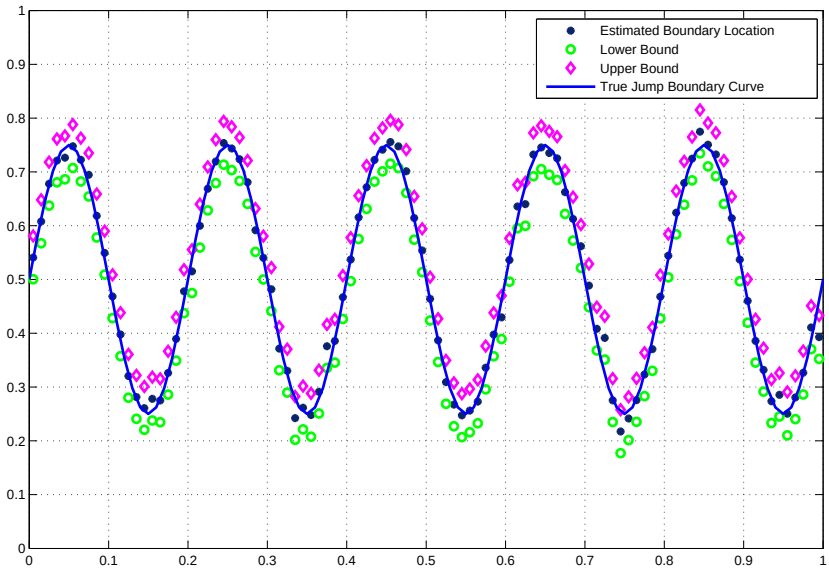


Figure 5.4: Jump boundary detection from One-stage procedure, Example 3

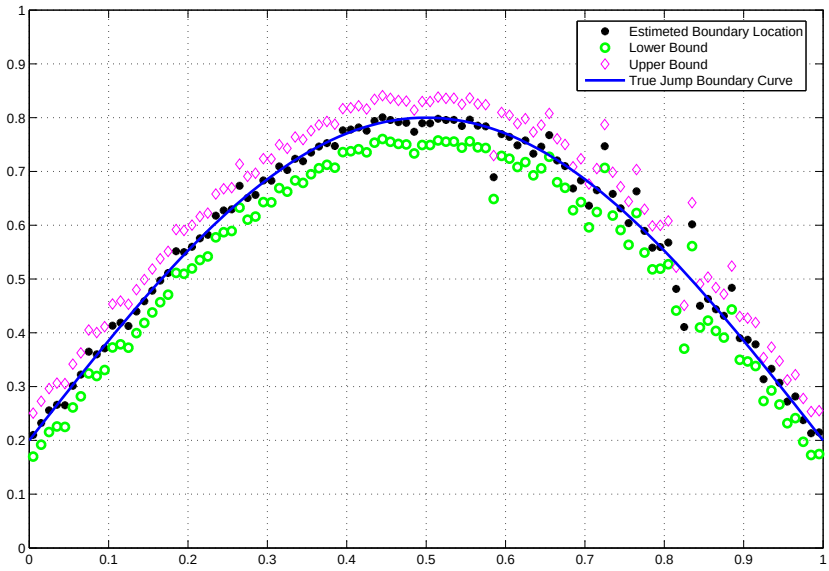


Figure 5.5: Jump boundary detection from One-stage procedure, Model from Qiu (2002)

1. Start by allocating 20 samples $(x_{1,1}, x_{1,2}, \cdots, x_{1,20})$ via a uniform design on [0, 1].

2. For each $x_{1,i}$, estimate the change point via the proposed two-stage procedure along the X_2 direction with respect to the surface response and denote it as b_i.

3. Check for possible outliers by adding more samples on selected locations along the X_1 direction. The new points are selected at locations that deviate from the remaining ones.

4. Use all $(x_{1,i}, b_i)$ we have so far to fit a structural change model $b(x_1)$ and calculate its MSE. Note that we only have small number of estimated boundary points. In general, a simple polynomial model does not work well for $g(z)$, due to the potential complicated nature of the boundary curve. Hence, structural change models are considered here which can provide a better approximation of the form

$$||g(z) - \sum_{i=1}^{k} h_i(z) I_{\{z \in \mathbf{S_i}\}}|| < \varepsilon$$

where $\{h_i(z)\}$ are polynomials defined on subregions $\{\mathbf{S_i}\}$ of the design space as follows $\{\mathbf{S_i} : z \in (\gamma_{i-1}, \gamma_i]\}$.

5. Augment new locations on X_1 direction;

6. For each new $x_{1,i}$, estimate the change point on X_2 direction with respect to the surface response.

7. Repeat Step 4, 5 and 6, till the change in the MSE is small enough.

5.3.1 Simulation Examples

1. Example 1

   In this example, we find three peaks (see Figure 5.6) after selecting the initial 20 sample along the X_1 direction via a uniform design on $[0, 1]$ and estimating the change point along the X_2 direction. . To check if some of them are outliers, along X_1 direction, we select the location around the middle of peak and valley, add six more samples and estimate corresponding change points. As shown in Figure 5.7, we doubt that the left one is an outlier. Go on to add two more

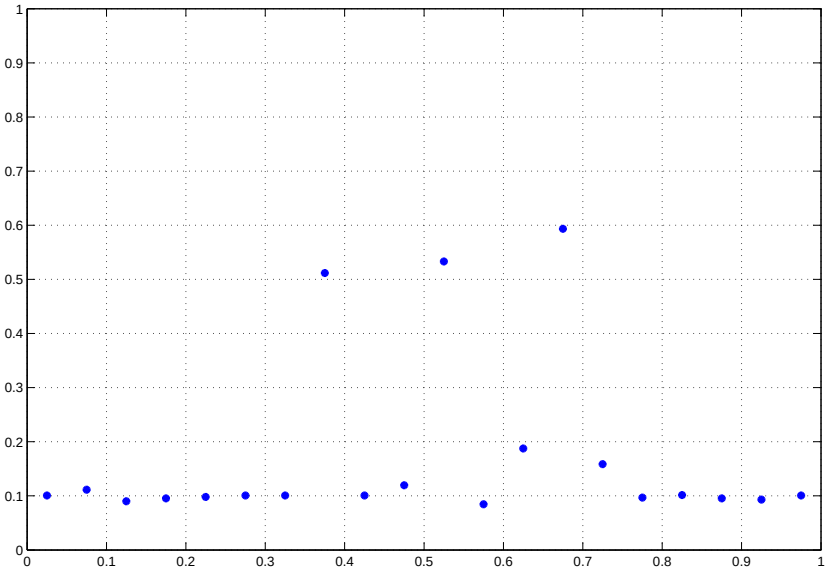


Figure 5.6: Initial 20 samples and change-point estimates, Example 1

new samples on X_1 direction and estimate those change–points on X_2 direction. From the Figure 5.8, we decide to treat the left one as an outlier.

We are more interested in the middle part of the boundary,since other parts are both so flat that we do not want to waste budge on them. A linear–linear–linear–linear model is selected as our first fitting model for the middle part of the jump boundary and the MSE is .0016 (as shown in Figure 5.9). The MSE of the new fitted linear–linear–linear–linear model is .0022 (see Figure 5.10) after adding four new points.

Eventually, 35 points are used along the X_1 direction. A quadratic–quadratic–quadratic model is fitted with $MSE = .0009$ (as shown in Figure 5.11). Again, since it is a simulated example, we know the true boundary curve, we can calculate the MSE of our fitted boundary curve: $\frac{1}{100}\sum_{i=1}^{100}(g(x_{1,i})-\hat{g}^*(x_{1,i}))^2 = .00056$. Compared with the result by using up all budget, our estimated boundary curve achieves the same performance but only one third of budget is used.

2. Example 2

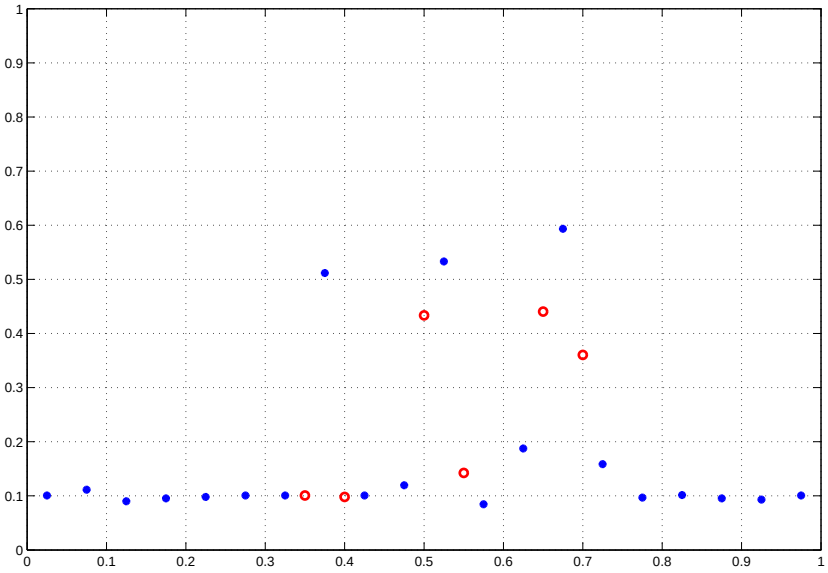


Figure 5.7: Add six more samples on X_1 direction and obtain estimated boundary locations, Example 1

As shown in Figure 5.12 and Figure 5.3, the angle (around .26 on X_1 direction) between the second and the third linear sub-models are so narrow. It is hard to obtain samples nearby that area if we only spread limited samples uniformly on the whole data range. We need adaptive sampling strategy to obtain good samples. We start with 20 samples on X_1 direction and estimate the location of change-point on X_2 direction, for each $x_{1,i}$. Follow our proposed adaptive procedure, we finally obtain perfectly fitted jump boundary curve by using less than one-third budget, as shown in Figure 5.13-5.19.

3. Example 3

   It is not difficult to detect those type of jump boundary curves as in Figure 5.20. Look at Figure 5.21, we find that there exists obvious period, which informs us to add nine new samples each of which is located between a peak and a valley (see Figure 5.22). Then we have enough boundary points to fit a smooth curve by cubic-splines (Figure 5.23).

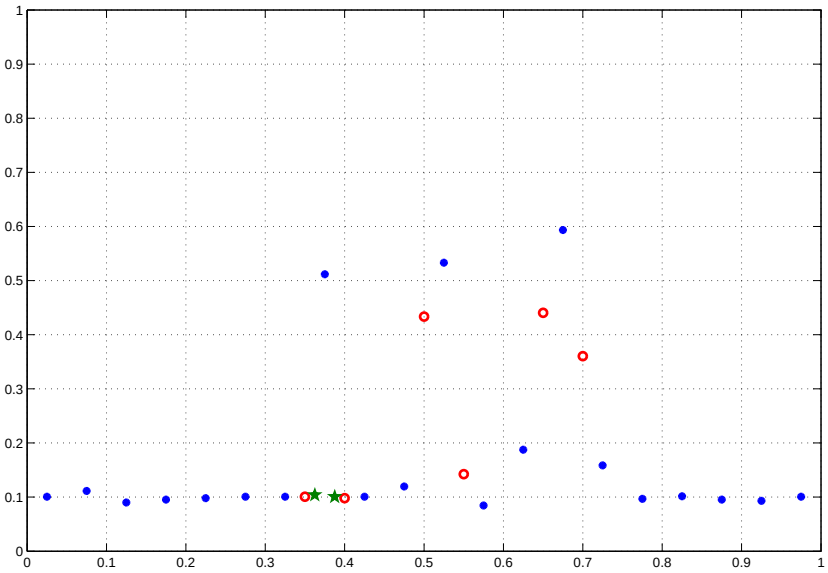


Figure 5.8: Add two more samples on X_1 direction and obtain estimated boundary locations to check outliers, Example 1

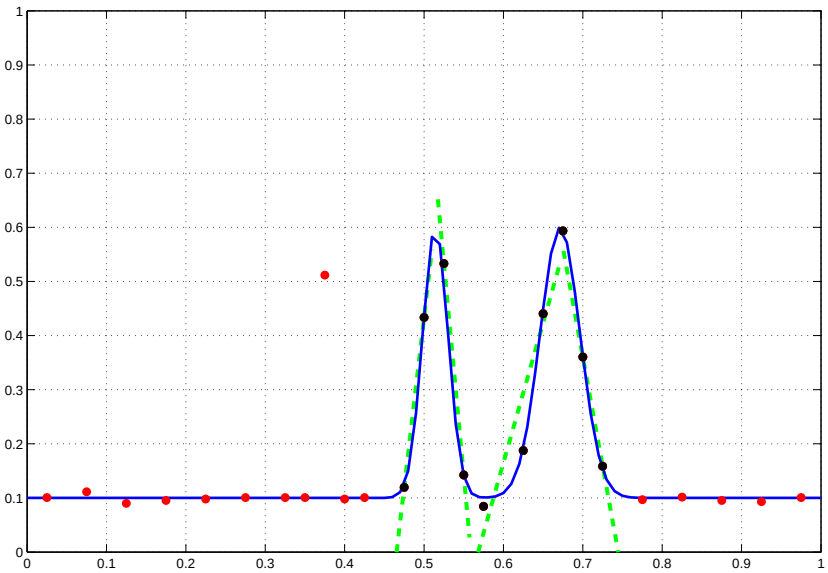


Figure 5.9: First fitted linear-linear-linear-linear model with MSE=.0016, Example 1

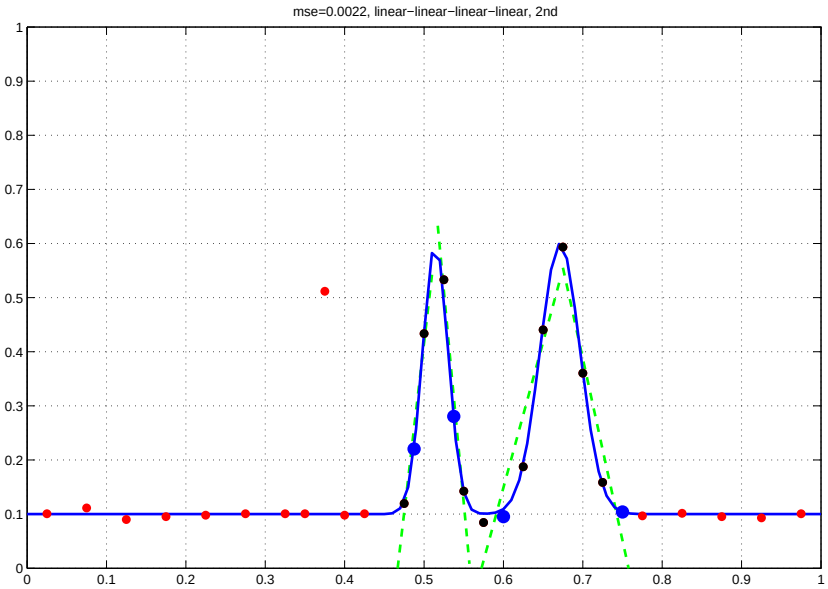


Figure 5.10: Add four new points and updated linear-linear-linear-linear model with MSE=.0022, Example 1

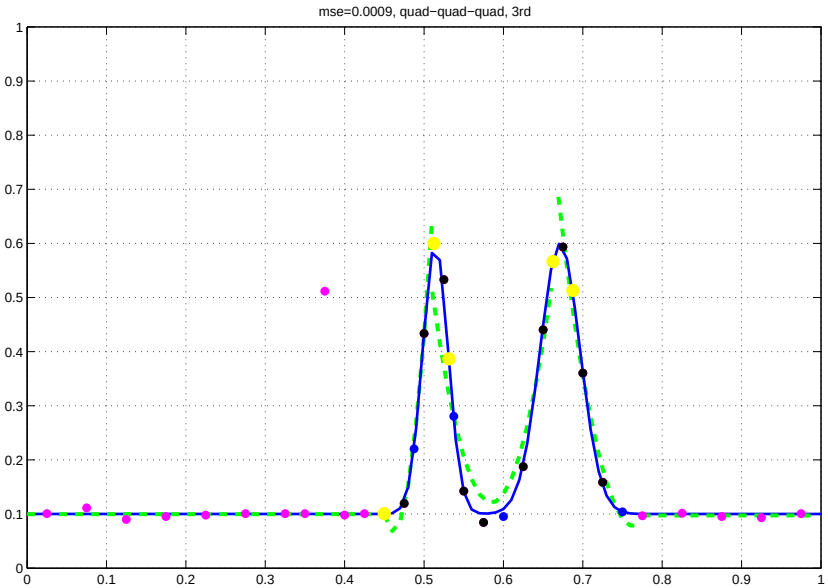


Figure 5.11: Add five more points and fit quadratic-quadratic-quadratic model with MSE=.0009, Example 1

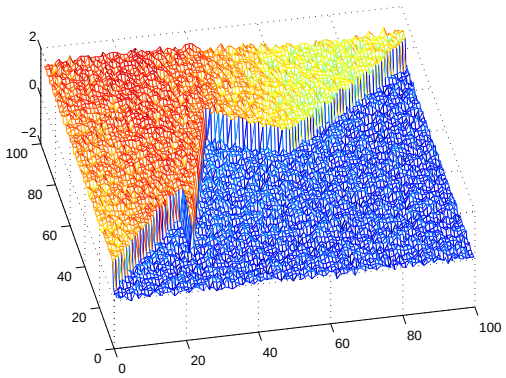


Figure 5.12: Three-dimension plot of Example 2

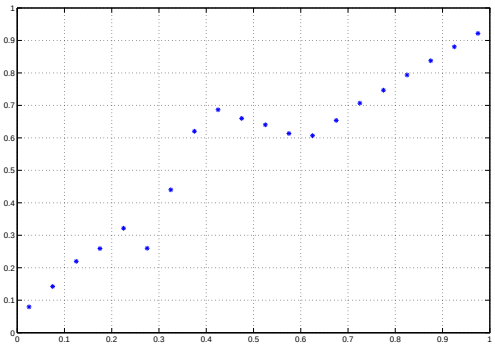


Figure 5.13: Initial 20 samples, Example 2

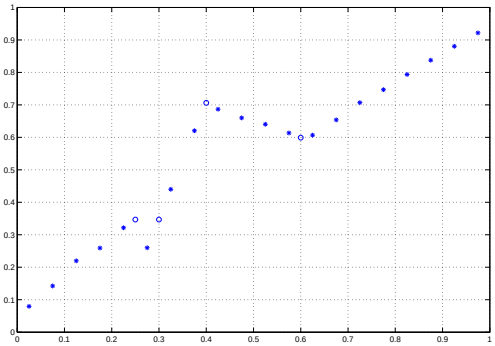


Figure 5.14: Four points are added, Example 2

5.3.2 Comparison with Kernel Smoothing Techniques

Qiu (2002) presented a procedure which simplifies the computation of some existing kernel-type methods in the statistical literature and adopts more flexible modeling

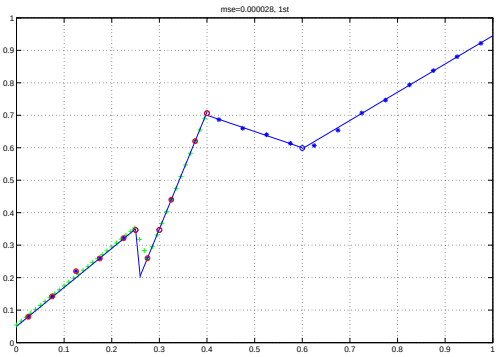


Figure 5.15: First fitted model, MSE=0.000028, Example 2

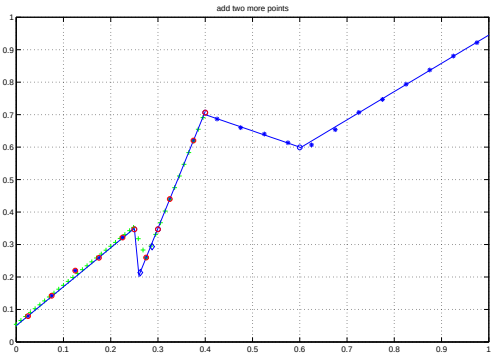


Figure 5.16: Two more points are added, Example 2

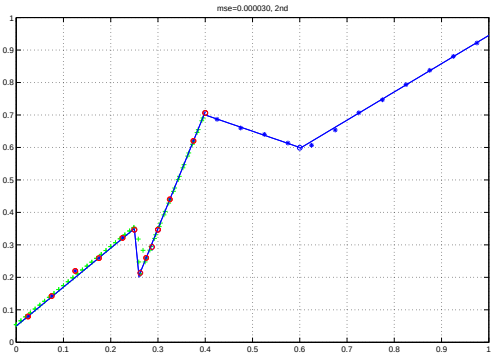


Figure 5.17: Updated model, MSE=0.000030, Example 2

assumptions.

The underlying regression model is given by

$$y_i = f(x_{1i}, x_{2i}) + \varepsilon_i, \qquad i = 1, 2, \ldots, n,$$

where $\{y_i\}$ are observations, $\{(x_{1i}, x_{2i})\}$ are design points in design space Ω which

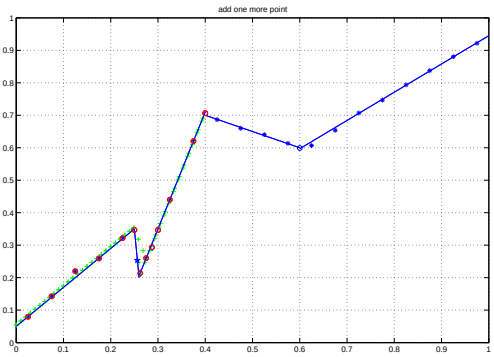


Figure 5.18: One new point is added, Example 2

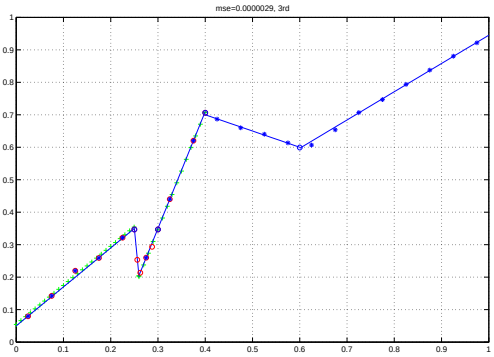


Figure 5.19: Final fitted model, MSE=0.0000029, Example 2

is a connected region in R^2, $f(x_1, x_2)$ is a bivariate regression function which is continuous in Ω except at the jump locations, and $\{\varepsilon_i\}$ are iid errors with mean 0 and variance σ^2.

For each $(x_1, x_2) \in \Omega \backslash O(\partial\Omega, b_n)$, where $\partial\Omega$ denotes the boundary point set of Ω, $O(\partial\Omega, b_n)$ is the border region of Ω, a difference of two weighted averages of the observations in the upper and lower sided neighborhoods of (x_1, x_2) is defined as $M_n^{(1)}(x_1, x_2)$. To detect jumps that are perpendicular to the x_1-axis, Qiu(1997) defined another quantity $RM_n(\theta, x_1, x_2)$ as a difference of two weighted averages of the observations located in two neighborhood on two different sides of the point (x_1, x_2) along the direction of $(\cos(\theta), \sin(\theta))$. To simplify its computation, a difference of two weighted average of the observations in the left and right sided neighborhoods

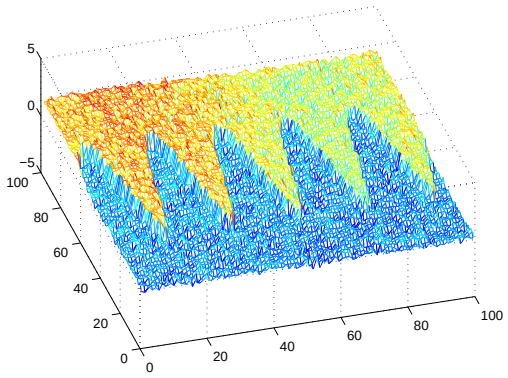


Figure 5.20: Three dimension plot of Example 3

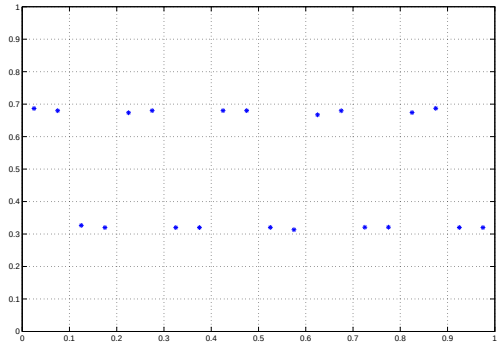


Figure 5.21: Initial 20 samples, Example 3

of (x_1, x_2) is defined as $M_n^{(2)}(x_1, x_2)$. Then, the following quantity is defined

$$M_n(x_1, x_2) = \max\{|M_n^{(1)}(x_1, x_2)|, |M_n^{(2)}(x_1, x_2)|\}$$

as a jump detection criterion; large values of $M_n(x_1, x_2)$ indicate a possible jump at (x_1, x_2). The point set

$$\hat{D_n} := \{(x_{1i}, x_{2i}) : M_n(x_{1i}, x_{2i}) \geq u_n\}$$

corresponds to the estimate of the true underlying jump locations $D := \{(x_1, x_2) : (x_1, x_2)$ is a point on the jump boundaries$\}$, where u_n is a positive threshold value. The jump detection procedure searched the x_1-axis and x_2-axis directions only at each design point for a possible jump. The computation complexity is increasing when the number of design points is larger. This idea can be generalized by search-

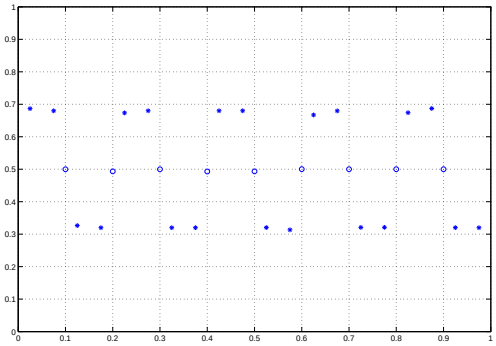


Figure 5.22: Nine new points are added, Example 3

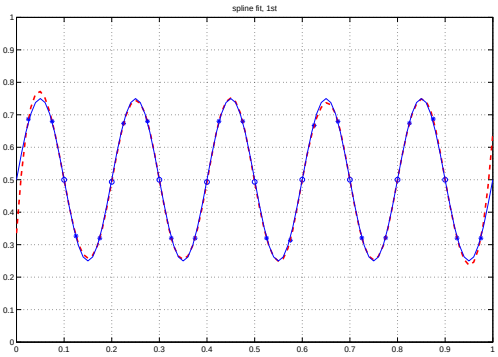


Figure 5.23: Cubic-spline fitting, Example 3

ing more than two directions as follows. Let $0 \leq \theta_1 \leq \theta_2 \leq \ldots \leq \theta_m < \pi$ be m directions in $[0, \pi)$, where $m \geq 2$ is an integer. At point (x_1, x_2), $M_n(x_1, x_2) := \max\{|RM_n(\theta_i, x_1, x_2)|, \quad i = 1, 2, \ldots, m\}$. Jump detection improves as more directions are searched at each design point, however, the procedure spends more computation time at the same time.

We try the same example from Qiu(2002) and consider the regression function $f(x_1, x_2) = \frac{1}{4}(1 - x_1)x_2 + [1 + 0.2sin(2\pi x_1)]I_{x_2 \geq 0.6\sin(\pi x_1)+0.2}$, for $(x_1, x_2) \in [0, 1] \times [0, 1]$ which has a unique jump location curve $g(x_1) = 0.6\sin(\pi x_1) + 0.2$ with jump magnitude $1 + 0.2\sin(2\pi x_1)$. The true regression surface and jump location curve are shown in figure 5.24. Observations are generated from this model with $\varepsilon \sim N(0, \sigma^2)$ at design points $(x_{1i}, x_{2j}) = (i/n_1, j/n_1)$, for $i, j = 1, 2, \ldots, n_1$. The sample size

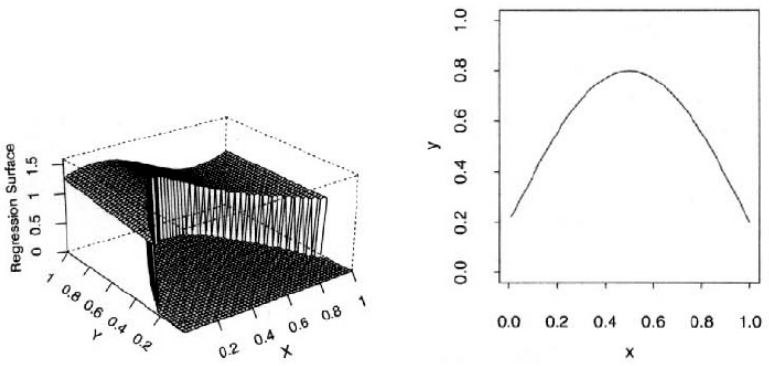


Figure 5.24: The True Regression Surface and Jump Boundary Curve of the Model from Qiu(2002)

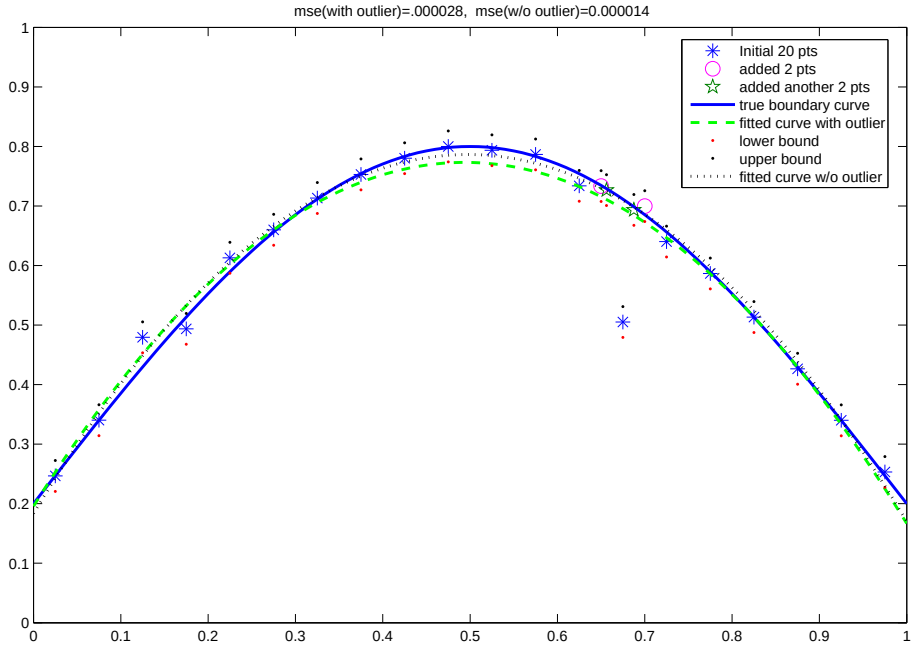
is $n = n_1 \times n_1$. We take $n_1 = 100$ and $\sigma = 0.5$. Figure 5.25 provide the fitting results. The upper one is the fitted boundary curve by the procedure we propose. And the lower one is got from Qiu(2002). After four repeats, we obtain the fitted curve. Compared to the kernel smoothing method, our procedure is much cheaper and faster.

5.3.3 Discussion

When the jump boundary curve (Figure 5.26) exists only in the middle of (X_1, X_2) plane, it is not hard to see the random pattern around those areas where there is no jumps in reality. Hence, we can ignore those areas and apply our procedure on what we are interested in.

5.3.4 Future Work

In this section a qualitative methodology was introduced for detection of a jump boundary curve. In our procedure, a uniform design is used to generate samples along the X_1 direction. For simplicity, we focus on one-stage procedure to describe possible problems. Suppose we use grid points $\{i/n_1\}$ on the X_1 direction, obtain the change point estimates $\{b(i/n_1)\}$ on the X_2 direction for each fixed x_1, and obtain a fitted curve. There are many issues that need to be examined in depth: such as consistency of the estimates, rate of convergence, asymptotics, confidence bounds,



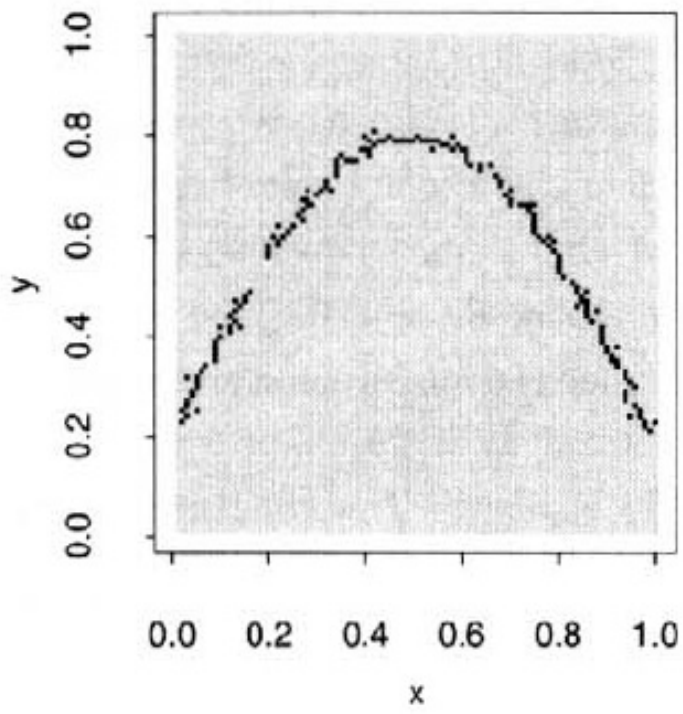


Figure 5.25: Comparison of the Fitted Jump Boundary Curves of the Model from Qiu(2002)

and pointwise confidence sets for the 2-dimensional case.

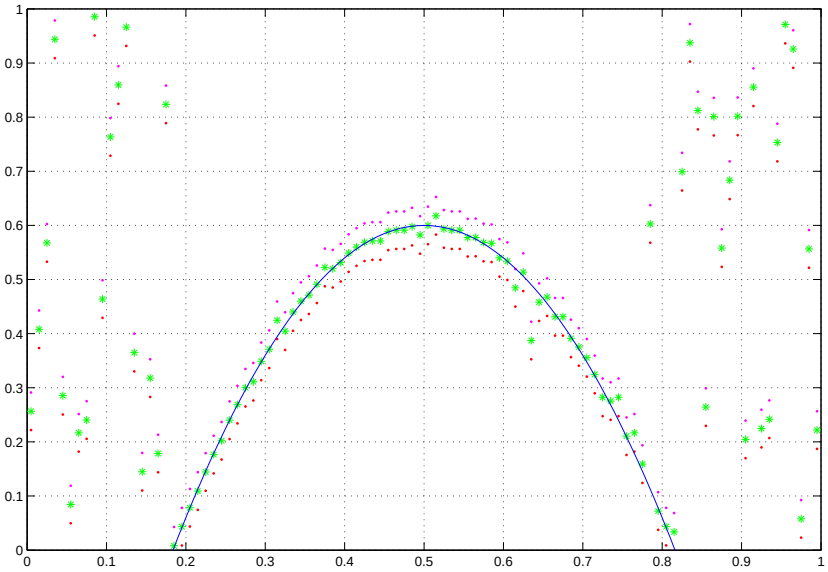


Figure 5.26: Special Example

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Basseville, M. and Nikiforov, I. (1993), *Detection of Abrupt Changes: Theory and Application*, Prentice Hall, Englewood Cliffs, NJ.

[2] Billingsley, P. (1999), *Convergence of Probability Measures*, Wiley Series in Probability and Statistics, New York, NY

[3] Chu, C.K., Glad, I.K., Godtliebsen, F. and Marron, J.S.(1998), Edge-Preserving Smoothers for Image Processing, *Journal of the American Statistical Association*, Vol. 93, No. 442.

[4] Dempfle, A. and Stute, W. (2002), Nonparametric estimation of a discontinuity in regression, *Statistica Neerlandica*, 56, 233-242.

[5] Ferger, D. (2004), A Continuous mapping theorem for the argmax-functional in the non-unique case, *Statistica Neerlandica*, 58, 83-96.

[6] Ferger, D.(2005), On the minimizing point of the incorrectly centered empirical process and its limit distribution in nonregular experiments, *ESAIM: Probability and Statistics*, 9, 307-322.

[7] Gijbels, I., Hall, P. and Kneip, A. (1999), On the estimation of jump points in smooth curves, *Annals of the Institute of Statistical Mathematics*, 51, 231-251.

[8] Gregoire, G. and Hamrouni, Z. (2002), Change point estimation by local linear smoothing, *Journal of Multivariate Analysis*, 83, 56-83.

[9] Hall, P. and Molchanov, I.S. (2003), Sequential methods for design-adaptive estimation of discontinuities in regression curves and surfaces, *Annals of Statistics*, 31, 921-941.

[10] Hung, Y.C. and Michailidis, G. (2007), Modeling, Analysis and Simulation of Switched Processing Systems, to appear in *ACM Transactions on Modeling, Analysis and Simulation*

[11] Kosorok, M.R. (2006), *Introduction to Empirical Processes and Semiparametric Inference*, to appear in Springer Series in Statistics.

[12] Kosorok, M.R. and Song, R. (2007), Inference under right censoring for transformation models with a change-point based on a covariate threshold, *Annals of Statistics*, 35, 957-989.

[13] Lai, T.L. (2001), Sequential analysis: some classical problems and new challenges, *Statistica Sinica*, 11, 303-408

[14] Lan, Y., Banerjee, M., and Michailidis, G. (2007), Change–point Estimation Under Adaptive Sampling, *Submitted*.

[15] Loader, C.R. (1996), Change point estimation using nonparametric regression, *Annals of Statistics*, 24, 1667-1678.

[16] Müller, H.G. (1992), Change-points in nonparametric regression analysis, *Annals of Statistics*, 20, 737-761.

[17] Müller, H.G. and Song, K.S. (1997), Two-stage change-point estimators in smooth regression models, *Statistics and Probability Letters*, 34, 323-335.

[18] Pons, O. (2003), Estimation in a Cox regression model with a change-point according to a threshold in a covariate, *Annals of Statistics*, 31, 442-463.

[19] Qiu, Peihua (1997) "Nonparametric Estimation of Jump Surface", *Sankhya*, Series A, **59**, 268-294.

[20] Qiu, P. (2002),A Nonparametric Procedure to Detect Jumps in Regression Surfaces, *Journal of Computational and Graphical Statistics*, Vol. 11, No. 4, 799-822.

[21] Qiu, Peihua (2005) Image Processing and Jump Regression Analysis, *Wiley*.

[22] Ritov, Y. (1990), Asymptotic efficient estimation of the change point with unknown distributions, *The Annals of Statistics*, 18, 1829-1839.

[23] Talagrand, M. (1994), Sharper bounds for Gaussian and empirical processes, *The Annals of Probability*, 22, 20-76.

[24] Van de Geer, S. (2000), *Empirical Processes in M-Estimation*, Cambridge University Press, Cambridge, UK.