Efficient Inferential Methods in Regression Models with Change Points or High Dimensional Covariates

by

Ritabrata Das

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2015

Doctoral Committee:

Professor Moulinath Banerjee, Co-Chair
Professor Bin Nan, Co-Chair
Professor Sioban D. Harlow
Professor Bhramar Mukherjee

Dedicated to the memory of Lord mamu

(my uncle, Dhrubo Jyoti Ganguly)

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank my advisers, Professor Moulinath Banerjee and Professor Bin Nan for being extremely patient, understanding and encouraging over the last five years. Having been taught by both of them in and outside a classroom, I found them both to be superb teachers. But I would remember both of them as the best mentors I could have wished for. Mouli, with all his enthusiasm for academics and beyond, was an adviser and friend molded into one. Bin, on the other hand, is without doubt one of the most patient and humble persons I have known; one's adviser is often referred to as their "academic father". I truly cannot think of a more fitting example of this than Bin. I consider it my good fortune to have them as my advisors. If I can emulate even a few of their qualities, I think I would have done good for myself.

I would like to thank Professor Sioban Harlow for agreeing to be on my committee and am extremely grateful for her inputs about the dissertation and also other projects. I have had the opportunity to be taught by some of the best teachers and collaborate with some of the best researchers during my time at the University of Michigan. I would like to thank every faculty member of the Biostatistics department for making this experience truly enriching. I feel the need to thank all the staff members of the Biostatistics department, who made all my problems their problems.

This would not be complete without thanking Professor Bhramar Mukherjee. She literally has been a friend, philosopher and guide. Working as her research assistant was both rewarding and enriching; being taught Biostat 699 by her was both taxing and fun; her inputs

as a member of my committee has been extremely helpful and for that I would like to thank her separately. But the reason I will be forever grateful to 'Bhramardi' is for giving many of us, including me, a home away from home. I can only say, my mother sleeps peacefully at night because she knows 'Bhramardi' is in Ann-Arbor.

Friends are the biggest support-system one can have and I believe no one can attest to this more than a graduate student. I would like to thank all my friends, right from high school through graduate school who have stood by me through thick and thin. It would not have been possible without them. Naming a few would be an injustice to the rest.

I would like to thank Swarnali for always being there without that gesture being reciprocated often enough. It is her presence which at many points of time made graduate life livable. I turn into a rude recluse at times but thankfully this has never stopped her from being there for me.

Finally, I would like to thank my parents for everything. I know, for a fact, this means more to them than it means to me. It is the drive, instilled in me by my parents, that makes me strive higher.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Efficient Inferential Methods in Regression Models with Change Points or High Dimensional Covariates

by

Ritabrata Das

Co-Chairs: Moulinath Banerjee and Bin Nan

This dissertation focuses on providing efficient inferential method for estimation in broken-stick model in cross-sectional as well as longitudinal studies, applying the multivariate adaptive elastic-net for statistical inference in a multiple pollutant model and proposing a variable selection procedure for multiple change-points in a broken-stick framework.

Estimation of change-point(s) in the broken-stick model has significant applications in modeling important biological phenomena. In the first project, we present a computationally economical likelihood-based approach for estimating change-point(s) efficiently in both cross-sectional and longitudinal settings. Our method, based on local smoothing in a shrinking neighborhood of each change-point, is shown via simulations to be computationally more viable than existing methods that rely on search procedures, with dramatic gains in the multiple change-point case. The proposed estimates are shown to have $\sqrt{n}$-consistency and asymptotic normality – in particular, they are asymptotically efficient in the cross-sectional setting – allowing us to provide meaningful statistical inference. As our primary and motivating application, we study the Michigan Bone Health and Metabolism Study cohort data to describe patterns of change in log estradiol levels around the final menstrual

period, for which a two change-point broken-stick model appears to be a good fit. We also illustrate our method on a plant growth data set in the cross-sectional setting.

Though there has been a considerable work done on studying the effects of coarse and fine ambient particles, how the constituent pollutants affect cardiovascular functioning is still not clearly understood. In the second project, we propose using the multivariate adaptive elastic-net to capture these effects in a multivariate autoregressive model for time series data. Because of the large number of highly correlated pollutants, a reliable method must take into account the high dimensionality as well as the multicollinearity issues. This is accomplished by using the adaptive elastic-net which deals effectively with the correlated nature of the data during variable selection. Furthermore, the selection consistency and asymptotic normality properties allow us to provide meaningful statistical inference in this set-up. The method is shown to perform well in numerical studies. As our motivating example, we study the effects of multiple pollutants on several cardiovascular end-points in a rat study based in Dearborn, Michigan, conducted by the Great Lakes Air Center for Integrative Environmental Research (GLACIER).

Finally, we look at problems where there are several covariates (large p) and some of the covariates have a simple linear effect while some have a broken-stick effect. In principle, we are looking at two separate but similar types of problems–one where we have several covariates each with possibly a broken-stick effect but with only a single change-point and the other where the broken-stick effect is exhibited in a single covariate but the number of change-points is unknown. In both settings we strive for a parsimonious yet accurate model, which necessitates an effective variable selection procedure. In a sparse setting, we illustrate the difficulty in using the popular variable selection methods and propose a post local-smoothing thresholded ridge regression as a method for identifying the non-zero linear and broken-stick type effects. We illustrate the efficiency of this approach via simulations and discuss possible routes for theoretical justification.

CHAPTER 1

Introduction

A principal task in modern statistics is to provide computationally and theoretically efficient solutions for non-standard regression set-ups. These problems find application in various disciplines of science including medicine, biology, genomics, environmental sciences and economics. Coming up with efficient estimation methods and studying their properties is a major challenge in these settings. One such problem is to devise a computationally fast as well as theoretically viable estimation method for a broken-stick model in both cross-sectional and longitudinal set-ups. This problem is studied in detail in Chapter 2. In Chapter 3, we propose an inference method based on multivariate adaptive elastic-net in a multiple pollutant set-up. Finally, in Chapter 4, an algorithm for variable selection for high dimensional sparse regression, where some of the covariate effects may be of the form of a broken-stick. This thesis covers three main topics. We provide a summary of these topics in the following section.

1.1 Fast efficient estimation method for broken-stick

The Michigan Bone Health and Metabolism Study (MBHMS) is a population-based longitudinal natural history study of ovarian aging conducted in a cohort of White women from the Tecumseh (Michigan) during their young and mid-adulthood years. The goal of *Sowers*

et al. (2008) was to describe the serum estradiol (E2) profile changes before and after Final Menstrual Period (FMP). Fig. 1 in *Sowers et al.* (2008) indicates that the mean function can be fit nicely by a piecewise linear model with multiple change-points (i.e. the broken-stick model), whose identification is of considerable significance. However, existing methods of change-point estimation in a broken-stick model are fairly slow for large sample sizes. In the particular scenario of *Sowers et al.* (2008), the effective sample-size is of order of 10^4 and hence a fast method of estimating the change-point locations precisely and providing corresponding confidence intervals is of considerable importance.

Our method, based on local smoothing in a shrinking neighborhood of each change-point, is shown via simulations to be computationally much faster than existing methods that rely on search procedures. The computational gain gets accentuated in multiple change-points models. The proposed estimates are shown to have $\sqrt{n}$-consistency and asymptotic normality – in particular, they are asymptotically efficient in the cross-sectional setting (in the sense, they have the same asymptotic distribution as the exact least squares estimate as shown in *Feder* (1975a)) – allowing us to provide meaningful statistical inference. We use our prosed estimation method to study the E2 profile, as discussed above.

1.2 High-dimensional Inference based on Multivariate Adaptive Elastic-net for Multiple Pollutant Data

With advancement in modern technology, high-dimensional regression problems are becoming more and more common in all disciplines of science. One such interesting problem arises in studying the effects of multiple pollutants on cardio-metabolic end-points. Although, the effects of fine particulate matter, as a whole, has been studied in recent years in quite some detail, how the constituent pollutants affect health responses is still up for debate. Two of the major problems in a multiple pollutant regression model is the high-dimensionality and the acute multicollinearity. We propose an inference procedure based on

multivariate adaptive elastic-net in a multivariate autoregressive model for time series data. This method is shown via simulations to effectively address both the major issues with this type of a modeling strategy. We used our method to study the effects of multiple pollutants on several cardiovascular end-points in a rat study based in Dearborn, Michigan, conducted by the Great Lakes Air Center for Integrative Environmental Research (GLACIER).

1.3 Variable selection for high-dimensional broken-stick regression

This problem is, in some sense, an extension of the problem discussed in 2. Previously, we were interested in estimating the broken-stick model in a fixed dimensional set-up. Now, with high dimensional regression techniques becoming more and more useful, the natural question is how to do variable selection in a usual sparse regression set-up with an added feature – we allow the covariates to have a broken-stick type effect on the response. This problem can be viewed from a very general setting with several covariates each with possibly multiple change-points. But to get a better understanding of the problems at hand, it makes more sense to look at two separate cases:

(i) There are several covariates, each possibly having a broken-stick effect; but we allow the broken-stick to have only one change-point.

(ii) The broken-stick effect is limited to only one covariate but it can have several change-points but the number of change-points is unknown

In both set-ups we need to arrive at a parsimonious yet accurate model, which of course necessitates an effective variable selection procedure for these models. Because of the lack of differentiability or restricted strong convexity (*Loh and Wainwright*, 2015) of the loss function, it is difficult to see how popular variable selection methods will work for this set-up. We propose post local smoothing thresholded ridge regression as a method for variable selection in such a scenario. The method is shown through numerical studies for the first

case to have a very high ability not only to separate signals from noise terms, but also to identify which signals are simply linear and which have a broken-stick type effect.

CHAPTER 2

Fast estimation of regression parameters in a broken stick model for longitudinal data

2.1 Introduction

In regression models, it is often assumed that the regression function throughout the domain of interest has the same parametric form. But it is also important to consider situations where the regression function has different functional forms in separate portions of the domain of interest. A special case of this is the continuous piecewise linear model, popularly referred to as the "broken-stick model". This model is frequently useful in environmental and biological setups where the locations of the change-points are of interest. The broken-stick with r change-points is

$$E(Y|X,Z) = \beta_0 + \beta_1 X + \sum_{k=1}^{r} \beta_{k+1} f(X,\tau_k) + Z^{\mathrm{T}}\lambda, \tag{2.1}$$

where

$$f(x,\tau) = (x-\tau)^+ = \begin{cases} x-\tau, & x > \tau; \\ 0, & x \leq \tau. \end{cases}$$

and τ_j's are ordered. Such a modelling strategy is of particular interest for the Michigan Bone Health and Metabolism Study (MBHMS).

MBHMS is a population-based longitudinal natural history study of ovarian aging conducted in a cohort of 664 White women from Tecumseh, Michigan during their young and mid-adulthood $(24 - 44)$ years. *Sowers et al.* (2008) studied the serum estradiol (E2) hormone levels in 629 women enlisted in the MBHMS over a fifteen year period starting from 1992. The goal of *Sowers et al.* (2008) was to describe the E2 profile changes before and after Final Menstrual Period (FMP). A semiparametric mixed model approach was implemented in *Sowers et al.* (2008), and smoothing splines were used for estimation. Referring to Fig. 1 in *Sowers et al.* (2008), it is clear that the mean function can be fit nicely by a piecewise linear model with multiple change-points, whose identification is of considerable significance. However, existing methods of change-point estimation in a broken-stick model are fairly slow for large sample sizes. In the particular scenario of *Sowers et al.* (2008), the effective sample-size is of order of 10^4 and hence a fast method of estimating the change-points precisely is of considerable importance.

In the early literature, it was generally assumed that either the exact location of the change-point τ is known (*Poirer*, 1973), or, at worst, it is known which two observation points τ should lie between (*Robinson*, 1964). Also, most of the early work focused on detection of whether a change-point existed at all. In this article, however, the existence of change-point(s)is assumed and the exact number of change-points, denoted by r, is assumed known. The focus here is to propose a quick estimation procedure that gets around the non-smoothness of the model without compromising asymptotic efficiency in the process.

The principal difficulty in the estimation problem arises when the locations of the change-points are unknown. For an independent and identically distributed error case, if the location of these change-points were known, we would have a standard linear regression problem. Even if a relatively small set of plausible values were known, one could perform least squares for the slope and intercept parameters for each of these plausible values to find the overall least squares estimate. However, in most scenarios, this is unlikely, and the set of plausible values over which one needs to search is typically all r tuples of ordered X_i's, leading to a

very high number of linear regressions, $\binom{n}{r}$ in principle, n being the sample size; see *Hudson* (1966).

Bellman and Roth (1969) proposed an alternative method based on dynamic linear programming but this method is even slower than the previous method. *Feder* (1975b) considered a general case of segmented regression problems and showed that the exact least squares estimate obtained by *Hudson* (1966) is asymptotically normal. In particular, for the broken-stick model, the estimate is $\sqrt{n}$– consistent.

Tishler and Zang (1981) were the first to suggest estimation of change-points using a maximum likelihood approach based on smoothing. They argued that the non-differentiability of the 'maximum' and 'minimum' operators in piecewise regression was the main problem in using maximum likelihood. However, if these operators were substituted by smoothed versions, maximum likelihood could readily be used for fast computation. *Tishler and Zang* (1981) suggested using a quadratic approximation, where the length of the interval on which f is smoothed was taken as an arbitrary small value. However, the behavior of these estimates, as the interval-length shrinks to 0, was not investigated. It is clear that unless the length of the interval is allowed to decrease with sample size, their algorithm cannot yield a consistent estimate.

Recent articles for broken-stick models include *Bhattachaya* (1987), *Huskova* (1998) and *Muggeo* (2003). While the first two deal with theory, *Muggeo* (2003) tries to develop an estimation strategy, but does not provide any asymptotic results and thus fails to address the theoretical efficiency of the approach. For a detailed description of Bayesian methods of change-point estimation refer to *Chen et al.* (2011).

In sum, the lack of a suitable method of estimation, that is optimal in terms of both precision and computational economy, has forced statisticians to fall back on the search-based algorithm of Hudson or related algorithms thereof. Our paper fills this gap in the literature.

We should note here that alternative approaches to studying 'kink-type' phenomena

have also been investigated. *Chiu et al.* (2006) suggested that in certain scenarios, instead of using the broken-stick as the true model, it might be better to use, what they referred to as, the "bent-cable model", which is a quadratic smoothing of the broken-stick model in a γ neighborhood around τ. Their change-point parameter was defined as the mid-point of the interval $(\tau - \gamma, \tau + \gamma)$ on which the smoothing was done; here both τ and γ are unknown parameters. It was shown that $\hat{\tau}_n$, the least squares estimate of τ, is $\sqrt{n}$– consistent. Also, for $\gamma_0 > 0$, $\hat{\gamma}_n$ is $\sqrt{n}$-consistent as well. In a previous article, *Chiu et al.* (2002) had shown that for $\gamma_0 = 0$, i.e., when the smooth model reduces to the broken-stick model, the asymptotics are complex and $\hat{\gamma}_n$ is at most $n^{1/3}$– consistent.

In this chapter, equation (2.1) is used as the *true* mean model. Ideally, one would want to minimize the residual sum of squares in this model by a Newton-Raphson type algorithm, but this is not viable owing to the non-differentiability of f at τ. To this end, we use a twice differentiable perturbation of f, denoted by q_n, as our working model, where q_n coincides with f outside a shrinking neighborhood of τ, say $(\tau - \gamma_n, \tau + \gamma_n)$, with γ_n, a user-specified tuning parameter, going to 0. Because q_n is differentiable, the minimization can be done by Newton's algorithm very quickly. For the iid error case, we show that our estimate of τ is indeed $\sqrt{n}$– consistent for τ, and furthermore has an asymptotic normal distribution with the same asymptotic variance as the exact least squares estimate of *Hudson* (1966). For the longitudinal model, the same method yields $\sqrt{n}$–consistent and asymptotically normal estimates for the change-points even for misspecified variance structures.

In sections 2.2 and 2.3, we introduce the model for both the cross-sectional and longitudinal set-ups respectively and outline the main steps of the estimation. The main theoretical results are presented along with the main ideas of the proofs. Section 2.4 contains simulation results indicating the efficiency of the proposed method while the method is applied to two real data —plant growth data (cross-sectional study) in section 2.5.1 and estradiol profile analysis (longitudinal study) in section 2.5.2. Proofs are provided in Appendix A.

2.2 Cross-sectional Study

We assume that the covariate X is contained in $[M_1, M_2]$. The regression parameter in the cross-sectional set-up (2.1), is denoted by $\theta^{\mathrm{T}} = (\beta^{\mathrm{T}}, \tau^{\mathrm{T}}, \lambda^{\mathrm{T}})$. We assume θ belongs to a compact set $\Theta = \mathbb{B} \times [M_1 + \delta, M_2 - \delta]^r \times \Lambda$, $\tau_k < \tau_{k+1}, k = 1, 2..., r-1$, and $|M_1|, |M_2| < \infty$ and δ is a known small positive constant, indicating the change-points need to be well-separated from the boundaries of the $X-$space. Without loss of generality, take $M_1 = 0$ and $M_2 = M$. We write $\beta^{\mathrm{T}} = (\beta_0, \beta_1, ..., \beta_{r+1}) \in \mathbb{B}$, where β_0 is the intercept and $\sum_{j=1}^{k} \beta_j$ is the slope of the k^{th} segment, $k = 1, 2..., r+1$. $\mathbb{B}$ is a compact set in $\mathbb{R}^{k+2}$; the restriction $\zeta \leq \beta_k \leq B$, $2 \leq k \leq r+1$ is imposed for the sake of identifiability. We also write $\tau^{\mathrm{T}} = (\tau_1, \tau_2, ..., \tau_r)$, with τ_k denoting the k'th smallest change-point, and again identifiability requires the conditions $\tau_k < \tau_{k+1}$, $k = 1, 2..., r-1$, $\tau_1 \geq \delta$ and $\tau_r \leq M - \delta$. Λ is assumed to be a compact set in $\mathbb{R}^l$. The errors ε_i are assumed to be independent and identically distributed with mean 0 and variance σ^2. The true parameter vector $\theta^0 = (\beta_0^0, \beta_1^0, ..., \beta_{r+1}^0, \tau_1^0, \tau_2^0, ..., \tau_r^0, \lambda^0)^{\mathrm{T}}$ is assumed to be an interior point of the compact set Θ. Our data are independent and identically distributed observations $\{Y_i, X_i, Z_i\}_{i=1}^n$ from (2.1) and henceforth, $\mathbb{P}_n$ denotes the empirical measure of the data. Note that, Y_i and X_i are scalars while Z_i is an $l-$dimensional vector of covariates with no change-points.

2.2.1 Estimation

Define,

$$M(\theta, x, y, z) = (y - E_\theta(Y|X = x, Z = z))^2 = \left[y - \left\{ \beta_0 + \beta_1 x + \sum_{k=1}^{r} \beta_{j+1} f(x, \tau_k) + z^{\mathrm{T}} \lambda \right\} \right]^2.$$

The exact least squares procedure aims to obtain the minimizer of

$$\mathbb{P}_n(M(\theta, X, Y, Z)) = \frac{1}{n} \sum_{i=1}^{n} M(\theta, X_i, Y_i, Z_i).$$

As $f(x,\tau)$ is not differentiable at τ, one cannot obtain the minimizer of $\mathbb{P}_n(M(\theta,X,Y,Z)) \equiv \mathbb{P}_n(M(\theta))$ (for notational convenience) by Newton's algorithm. So, we resort to minimize $\mathbb{P}_n(M_n(\theta))$, where

$$M_n(\theta) \equiv M_n(\theta,x,y,z) = [y - \{\beta_0 + \beta_1 x + \sum_{k=1}^{r} \beta_{j+1} q_n(x,\tau_k) + z^{\mathrm{T}}\lambda\}]^2$$

is *our working model*, a smoothed approximation of $M(\theta)$. Basically each of the f's in $M(\theta)$ is replaced by its corresponding smoothed version q_n to obtain $M_n(\theta)$.

As far as the functional form of q_n is concerned, the motivation lies in the work of *Tishler and Zang* (1981) and *Chiu et al.* (2006). *Tishler and Zang* (1981) suggested using a quadratic approximation, where the length of the interval on which f is smoothed was taken as an arbitrary small value, while *Chiu et al.* (2006) considered the length of the corresponding interval as a parameter. We consider the same functional form for q_n as in these papers, but in our model, the length of the interval on which we smooth f is a user-specified tuning parameter shrinking to 0 with n at an appropriate rate as $n \to \infty$. More specifically,

$$q_n(x,\tau) = \begin{cases} 0, & \text{if } x < \tau - \gamma_n; \\ \frac{(x-\tau+\gamma_n)^2}{4\gamma_n}, & \text{if } \tau - \gamma_n \le x \le \tau + \gamma_n; \\ (x-\tau), & \text{if } x > \tau + \gamma_n; \end{cases} \tag{2.2}$$

where γ_n is a deterministic sequence, that approaches zero as $n \to \infty$.

Define $\hat{\theta}_n = \arg\min_{\theta\in\Theta} \mathbb{P}_n(M_n(\theta))$. In Appendix A.1, we show that $\hat{\theta}_n$ is a zero of the surrogate empirical estimating function $\mathbb{U}_n(\theta) = \partial\mathbb{P}_n(M_n(\theta))/\partial\theta$, with probability increasing to 1, validating the use of the solution of $\mathbb{U}_n(\theta)$ as our numerical estimate. It is not clear whether this zero is unique but this is not an issue, since for our asymptotics, how we pick the minimizer is immaterial, meaning that the results hold true for any choice of zero of $\mathbb{U}_n(\theta)$. Our numerical results, however, suggest that generally there is a unique minimizer.

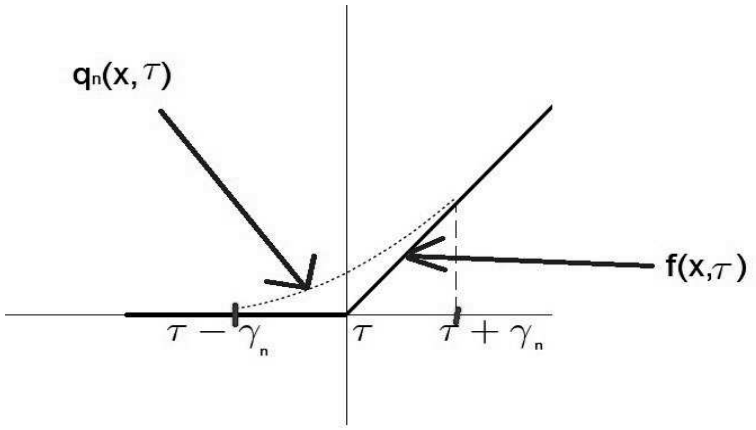

Figure 2.1: q_n is the smoothed version of f.

2.2.2 Asymptotic Results

For model (2.1) we consider the following regularity conditions.

Condition 1.1: There are r distinct change-points $\tau_1, \ldots, \tau_r$ in model (2.1) for a fixed integer $r \geq 1$; r is known.

Condition 1.2: Covariate $X \in [0, M]$, $M < \infty$, follows a continuous distribution F_X such that $\text{pr}(\tau_k < X \leq \tau_{k+1}) > 0$, $k = 0, \ldots, r$ with $\tau_0 = 0$ and $\tau_{r+1} = M$. The joint distribution of Z is denoted by F_Z.

Condition 1.3: Changes of slope parameters satisfy $0 < \zeta \leq |\beta_k| \leq B < \infty$, $k = 2, 3, \ldots, r+1$, for some constants ζ and B.

Theorem 2.1. *Under Conditions 1.1-1.3, $\hat{\theta}_n$ is a consistent estimator for θ^0 for any deterministic sequence $\gamma_n \to 0$ as $n \to \infty$.*

The proof of Theorem 2.1 is based on the argmax (argmin) continuous mapping theorem (*van der Vaart and Wellner*, 1996). First, we show that θ^0 is the unique minimizer of $P(M(\theta))$ in Θ. then, it suffices to show that $\|\mathbb{P}_n(M_n) - P(M)\| := \sup_{\theta \in \Theta} |\mathbb{P}_n(M_n(\theta)) - P(M(\theta))| = o_p(1)$; here $Pf = \int f dP$, P being the probability measure that generates the data. For the case with a single change-point and covariate Z absent, the proof is presented

in Appendix A.1. The proof for the case with multiple change-points or with other covariates is an exercise involving extensive algebraic derivations following the same line. Note that, in Appendix A.1, we also prove that θ^0 is the unique solution of $U(\theta) = \partial P(M(\theta))/\partial\theta = 0$, which implies that the estimate obtained by Newton-Raphson of the smoothed score equation converges to the true θ^0, and not to some local minima.

Theorem 2.2. *For $\gamma_n = n^{-\alpha}$ with $\alpha > 1/2$, under Conditions 1.1-1.3, we have that $n^{1/2}(\hat{\theta}_n - \theta^0)$ converges in distribution to a normal random variable with mean 0 and covariance matrix $2\sigma^2\dot{U}_*^{-1}(\theta^0)$. The kl-th element of matrix $\dot{U}_*$ is $(\dot{U}_*(\theta^0))_{kl} = 2P(H^T(\theta^0)H(\theta^0))$, where*

$$H(\theta) = \Big(\ 1, \quad X, \quad f(X,\tau_1), \quad \ldots \quad , f(X,\tau_r), \quad -\beta_2\mathbf{1}(X > \tau_1), \quad \ldots \quad , -\beta_{r+1}\mathbf{1}(X > \tau_r), \quad Z \ \Big)_{1\times(2+2r+l)}.$$

The proof of Theorem 2.2 also consists of two major steps. To this end, let us define θ_n as the minimizer of $P(M_n(\theta))$ in Θ closest to θ^0. The first step is to show that θ_n converges to θ^0 with a faster than $\sqrt{n}$ rate, which in fact is γ_n, and the second to show the asymptotic normality of $n^{1/2}(\hat{\theta}_n - \theta_n)$. Both steps rely on Taylor series expansions. For notational simplicity, we provide the proof for the case with single change-point and absence of Z in Appendix A.2. The case with multiple change-points and other covariates is again a straightforward extension.

The following Corollary shows that our proposed local smoothing method does not lose any efficiency. Its proof is provided in Appendix A.3.

Corollary 2.3. *The asymptotic distribution of our estimate, as stated in Theorem 2.2, is exactly the same as that in Feder (1975a) for the exact least squares estimate in the broken stick model.*

Remark 2.4. Please note that for the sake of notational convenience and to keep the results and proofs terse, only one variable with change-points have been included in model (2.1). However, the method will work equally well for a model consisting of multiple variables, with

multiple change-points in each variable, and the results will be analogous.

2.3 Longitudinal study

The model for the longitudinal study with a broken-stick mean function is

$$E(Y_{ij}|X_{ij}, Z_i) = \mu_{ij} = \beta_0 + \beta_1 X_{ij} + \sum_{k=1}^{r} \beta_{k+1} f(X_{ij}, \tau_k) + Z_i^T \lambda, \tag{2.3}$$

where Y_{ij} is the response of the i^{th} subject at the j^{th} time-point (t_{ij}) and X_{ij} denotes the corresponding covariate with r change-points, while Z_i are l time-invariant covariates, $j = 1, \ldots, m_i$, $i = 1, \ldots, n$.

For the regression parameters, $\beta^T = (\beta_0, \beta_1, .., \beta_{r+1})$, we have the same assumptions as in the cross-sectional model. We assume the effect sizes $\lambda \in \Lambda$ (a compact set in $\mathbb{R}^l$) and τ is the vector of change-points, as before. Here, $\theta^T = (\beta^T, \tau^T, \lambda^T)$ is our parameter of interest and θ^0 is the true value of θ.

As far as the variance function is concerned, we postulate the following form:

$$\begin{aligned} \text{Cov}(Y_{ij}, Y_{ik}) &= g(\eta, t_{ij}, t_{ik}), \\ \text{Cov}(Y_{ij}, Y_{lk}) &= 0, i \neq l; j, k = 1, \ldots, m_i; i = 1, \ldots, n, \end{aligned} \tag{2.4}$$

where η is the vector of covariance parameters. We assume that the observations across individuals are independent and the correlation between different observations of the same individual can depend on the time-points but not on the mean parameters, θ.

Y_i is used to denote the vector of m_i observations for the $i-$th individual, $i = 1, \ldots, n$. $Y = (Y_1, \ldots, Y_n)$ is the vector of all responses. We use similar definitions for X_i and X. Let $\Sigma^{(i)}$ denote the dispersion matrix of Y_i and Σ the dispersion matrix of Y. The true dispersion matrix is denoted by Σ_0, which can be written as $\Sigma(\eta^0)$, η^0 being the true value of η.

To establish the asymptotic results rigorously, the problem needs to be cast in a proper

mathematical framework. We assume that, the number of repeated measures is denoted by the random variable D which takes values in the integer-space $\{1, 2, \ldots, L\}$ with probabilities $p_1, p_2, \ldots, p_L$ respectively. Note that this L is assumed fixed and known. Also we have a triangular array of X-values,

$$\begin{array}{ccccc} X_1^{(1)} & & & & \\ X_1^{(2)} & X_2^{(2)} & & & \\ X_1^{(3)} & X_2^{(3)} & X_3^{(3)} & & \\ \vdots & & & \ddots & \\ X_1^{(L)} & X_2^{(L)} & X_3^{(L)} & \ldots & X_L^{(L)}. \end{array}$$

When $D = d$, the d-th row of this array is selected as the set of time-dependent covariates. The same is true for the measurement errors $\{\varepsilon_{ij}\}$ and they are assumed to be independent of $\{X_{ij}\}$'s. Thus,

$$Y^{(D)} = \beta_0 + \beta_1 X^{(D)} + \sum_{k=1}^{r} \beta_{k+1} f(X^{(D)}, \tau_k) + Z^T \lambda + \varepsilon^{(D)} = \mu^{(D)} + \varepsilon^{(D)}.$$

The observation for each individual consists of $(D, Y^{(D)}, X^{(D)}, Z)$ and our data comprise of n *iid* copies of this array. As with most inference methods in longitudinal studies, we allow for ignorable dropouts (*Rubin*, 1976).

2.3.1 Estimation

The estimation process is divided into three steps:

Step 1 : Assume working independence, i.e. take $\Sigma^{(i)} = I, i = 1, \ldots, n$. As for the cross-sectional study, replace each of the f's by their respective smoothed version q_n's. Now, find the corresponding estimate $\hat{\theta}_n^{(I)}$ as the solution to the estimating equation

$$\frac{\partial}{\partial \theta} \mathbb{P}_n [(Y^{(D)} - \mu_n^{(D)})^T (Y^{(D)} - \mu_n^{(D)})] = 0,$$

where $\mu_n^{(D)}$ is the smoothed version of $\mu^{(D)}$.

Step 2 **:** Then use $\hat{\theta}_n^{(I)}$ to estimate the nuisance parameter η. The specifics will depend on the nature of the covariance function g in (2.4).

Step 3 **:** Use $\hat{\eta}_n$ obtained in step 2 to estimate θ. So the final $\hat{\theta}_n$ is the solution to the estimating equation

$$\frac{\partial}{\partial\theta}\mathbb{P}_n[(Y^{(D)}-\mu_n^{(D)})^T\hat{\Sigma}_n^{-1}(Y^{(D)}-\mu_n^{(D)})]=0,$$

where $\hat{\Sigma}_n^{-1}$ is the block-diagonal dispersion matrix based on $\hat{\eta}_n$.

2.3.2 Asymptotic Results

As in the cross-sectional model, for the longitudinal model we consider similar regularity conditions.

Condition 2.1: There are r distinct change-points $\tau_1,\ldots,\tau_r$ in model (2.3) for a fixed integer $r\geq 1$; r is known.

Condition 2.2: Conditional on $D=d$, covariate $X\in[0,M]^d$, $M<\infty$, follows a continuous distribution F_X such that $\text{pr}(\tau_k<X_j\leq\tau_{k+1})>0$, for some $j=1,\ldots,d$, for all $k=0,\ldots,r$ with $\tau_0=0$ and $\tau_{r+1}=M$. Also we assume the covariates Z follow a joint distribution F_Z.

Condition 2.3: Changes of slope parameters satisfy $0<\zeta\leq|\beta_k|\leq B<\infty$, $k=2,3,\ldots,r+1$, for some constants ζ and B.

Condition 2.4: There exists a positive definite matrix W, such that estimated covariance matrix $\hat{\Sigma}_n$ satisfies $\sqrt{n}(\hat{\Sigma}_n-W)=O_p(1)$.

Theorem 2.5. *Under Conditions 2.1-2.4,*

(a) The estimator $\hat{\theta}_n$ is consistent for θ^0 given any deterministic sequence $\gamma_n\to 0$ as $n\to\infty$.

(b) For $\gamma_n = n^{-\alpha}$ with $\alpha > 1/2$, $n^{1/2}(\hat{\theta}_n - \theta^0)$ converges in distribution to a normal random variable with mean 0 and covariance matrix

$$K^{(W^{-1})} = 2\sum_{d=1}^{L} P^{(d)}[(H^T(\theta^0)W^{-1}H(\theta^0))^{-1}(W^{-1}H(\theta^0))^T\Sigma_0(W^{-1}H(\theta^0)) \\ (H^T(\theta^0)W^{-1}H(\theta^0))^{-1}]p_d$$

where

$$H(\theta) = \Big(\begin{array}{ccccccccc} 1 & X & f(X,\tau_1) & \dots & f(X,\tau_r) & -\beta_2\mathbf{1}(X>\tau_1) & \dots & -\beta_{r+1}\mathbf{1}(X>\tau_r) & Z \end{array} \Big)_{d\times(l+2r+2)};$$

here $P^{(d)}f = \int f dP^{(d)}$, $P^{(d)}$ being the probability measure that generates the data given $D = d$.

Remark 2.6. If the matrix W in condition 2.4 is indeed the true covariance matrix Σ_0, i.e., $\hat{\Sigma}_n$ is a $\sqrt{n}$-consistent estimate of Σ_0, then for $\gamma_n = n^{-\alpha}$ with $\alpha > 1/2$, $n^{1/2}(\hat{\theta}_n - \theta^0)$ converges in distribution to a normal random variable with mean 0 and covariance matrix

$$K^{(\Sigma_0^{-1})} = 2\sum_{d=1}^{L} P^{(d)}[(H^T(\theta^0)\Sigma^{0^{-1}}H(\theta^0))^{-1}]p_d.$$

The proof of Theorem 2.5 is similar to the proofs of Theorems 2.1 and 2.2 in Section 2.2. The main proof is divided into three major parts. First, we show that $\sqrt{n}(\hat{\theta}_n^{(I)} - \theta^0)$ converges to $N(0, K^{(I)})$ in distribution. Next, we prove $\sqrt{n}(\hat{\theta}_n^{(W^{-1})} - \theta^0)$ converges to $N(0, K^{(W^{-1})})$ in distribution. Here, $\hat{\theta}_n^{(W^{-1})}$ is defined as the minimizer of $\mathbb{P}_n(Y - \mu_n)^T W^{-1}(Y - \mu_n)$, which is shown to be a zero of $\mathbb{U}_n^{(W^{-1})}(\theta) = \frac{\partial}{\partial\theta}\mathbb{P}_n(Y - \mu_n)^T W^{-1}(Y - \mu_n)$, with probability increasing to 1. Finally, we show that $\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^{(W^{-1})}) = o_p(1)$, which proves Theorem 2.5.

For the sake of notational convenience, the proof is presented in Appendix A.4, for $r = 1$ and for fixed visit-times, i.e. $D \equiv m$ or equivalently $m_i = m$ for all $i = 1, \dots, n$. Also for brevity, we exclude the covariates Z in the proof.

Though the proof provided in Appendix A.4 is for a fixed number of visit-times, it holds true for variable number of visit-times, as stated in Theorem 2.5. Notice that, conditional on $D = d$ (this event has probability p_d), it is shown that $n^{1/2}(\hat{\theta}_n - \theta^0)$ converges in distribution to a normal random variable with mean 0 and covariance matrix

$$\begin{aligned} K^{(W^{-1})} &= 2P^{(d)}[(H^T(\theta^0)W^{-1}H(\theta^0))^{-1}(W^{-1}H(\theta^0))^T\Sigma_0(W^{-1}H(\theta^0)) \\ &\quad (H^T(\theta^0)W^{-1}H(\theta^0))^{-1}]. \end{aligned}$$

Now, the result of Theorem 2.5 easily follows.

Corollary 2.7. *Denoting the mean function at $X = x, Z = z$ by $\mu(x, z, \theta)$, we have $\sqrt{n}(\mu(x, z, \hat{\theta}_n) - \mu(x, z, \theta^0))$ converges in distribution to a normal random variable with mean zero and variance $a^T K^{(W^{-1})} a$, where*

$$a^T = (1, x, f(x, \tau_1^0), \ldots, f(x, \tau_r^0), -\beta_2^0 \mathbf{1}(x > \tau_1^0), \ldots, -\beta_{r+1}^0 \mathbf{1}(x > \tau_r^0), z)$$

.

This result is useful in providing pointwise confidence bands for the broken-stick mean function as illustrated in Fig 2.4. The proof is provided in the Appendix A.8

Remark 2.8. Note that the estimated confidence band for the mean at $\hat{\tau}_n$, as provided by Corollary 2.7 is discontinuous. The asymptotic distribution of $\sqrt{n}(\mu(\hat{\tau}_n, z, \hat{\theta}_n) - \mu(\tau^0, z, \theta^0))$ is not a direct application of this result, but needs separate calculations —a direct application of the delta method.

2.4 Simulations

2.4.1 Cross-sectional set-up

Simulations were conducted to compare the proposed method with the existing one. Models with one and two change-points were both considered. Sample sizes were varied, $n = 50, 200, 1000, 5000$. For each of the two models, for a fixed n, 3 different sets of values of θ were considered within the domain of interest. For each value of θ, the proposed and existing (*Hudson*, 1966) methods were both repeated $N = 1000$ times. The run-times, a measure of computational efficiency, for each of the methods were then averaged over these 1000 repetitions and over the 3 different values of θ. This was done to average out discrepancies being caused by individual θ's. Error standard deviation σ was taken to be equal to 0.1 for all cases and $M = 1$. For all simulations, α was taken to be 1. The simulations were carried out on an Intel(R) Core(TM) i7 system with 1.6 GHz and 8 GB RAM in a 64-bit OS.

2.4.1.1 Results

Table 2.1: Simulation results comparing the run-times of the existing (*Hudson*, 1966) and proposed methods for one and two change-point(s) model, with ratio of the time taken by the existing method with respect to that of the proposed one.

Sample	Mean Time (Seconds)					
Size n	One change-point			Two change-points		
	Existing	Proposed	Ratio	Existing	Proposed	Ratio
50	0.18	0.006	30	2.36	0.02	118
100	0.30	0.008	38	13.86	0.03	462
500	0.97	0.02	49	64.87	0.06	1015
1000	1.89	0.03	63	947.03	0.08	11838
5000	4.98	0.06	83	22843	0.20	114215

From Table 2.1, it is obvious that the proposed method is much faster than the exact least squares method, especially for two change-point problems. Also Tables 2.2 & 2.3 indicate that the change-point estimates of the proposed method are almost as accurate as the exact least squares estimate. The biases are close to zero for all sample sizes, especially for the large

Table 2.2: Bias and variances for the change-point estimate $\hat{\tau}_n$ compared for one change-point problem in 3 setups: A: $\theta^{\mathrm{T}} = (0.2, 1, 1, 0.6)$, B: $\theta^{\mathrm{T}} = (0.3, 1.5, 1, 0.8)$ & C: $\theta^{\mathrm{T}} = (0.3, 1.5, -1, 0.2)$ (S.D.: Average of estimated standard deviations over 1000 replications; Emp. S.D. : Sample standard deviation based on 1000 replications).

Sample Size n	Bias (S.D., Emp. S.D.) $\times 10^{-3}$ Existing	Proposed	Theoretical S.D. $\times 10^{-3}$
	Set-up A		
50	-8.2 (58.1, 60.3)	-12.4 (58.3, 61.0)	57.8
100	-4.3 (41.0, 41.2)	-5.2 (41.0, 41.3)	40.9
500	-2.1 (18.3, 18.4)	-2.9 (18.3, 18.5)	18.3
1000	-0.6 (12.9, 12.9)	-0.9 (12.9, 13.0)	12.9
5000	-0.0 (5.8, 5.8)	-0.1 (5.8, 5.8)	5.8
	Set-up B		
50	-9.2 (71.6, 73.2)	-14.1 (71.6, 73.6)	70.7
100	-4.7 (50.1, 50.3)	-5.9 (50.1, 50.4)	50.0
500	-2.8 (22.4, 22.4)	-3.3 (22.4, 22.5)	22.3
1000	-0.9 (15.8, 15.8)	-1.2 (15.8, 15.8)	15.8
5000	-0.1 (7.1, 7.1)	-0.1 (7.1, 7.1)	7.1
	Set-up C		
50	8.1 (71.6, 73.3)	12.8 (71.6, 73.5)	70.7
100	4.7 (50.1, 50.4)	6.2 (50.1, 50.5)	50.0
500	1.8 (22.4, 22.4)	2.0 (22.5, 22.6)	22.3
1000	0.2 (15.8, 15.8)	0.4 (15.8, 15.8)	15.8
5000	0.1 (7.1, 7.1)	0.1 (7.1, 7.1)	7.1

samples. The standard deviation estimates are very close to the sample standard deviations indicating our standard deviation estimates work well, especially for large samples. The estimates are also very close to the theoretical standard deviations, indicating the asymptotic efficiency of our estimates. Although the bias and variances for the β's have not been tabulated for the sake of brevity, we observed that our β estimates also have comparable Mean Squared Errors to their respective exact least squares estimates.

Table 2.3: Bias and variances for the change-point estimate $\hat{\tau}_n$ compared for two change-points problem in 3 setups: D: $\theta^{\mathrm{T}} = (0.3, 1, 1, 1, 0.2, 0.8)$, E: $\theta^{\mathrm{T}} = (0.2, 1, 2, 1, 0.4, 0.6)$ & F: $\theta^{\mathrm{T}} = (0.3, 1, -1, 1, 0.2, 0.8)$ (S.D.: Average of estimated standard deviations over 1000 replications; Emp. S.D. : Sample standard deviation based on 1000 replications).

Sample Size (n)	$\hat{\tau}_{1n}$ Bias (S.D., Emp. S.D.) $\times 10^{-3}$ Existing	Proposed	Theo. S.D. $\times 10^{-3}$	$\hat{\tau}_{2n}$ Bias $\times 10^{-3}$ (S.D., Emp. S.D.) $\times 10^{-3}$ Existing	Proposed	Theo. S.D. $\times 10^{-3}$
			Set-up D			
50	10.2 (63.2, 63.9)	20.3 (63.3, 64.1)	62.0	-11.1 (55.0, 55.5)	-19.2 (55.1, 55.6)	54.0
100	5.1 (44.2, 44.6)	6.3 (44.3, 44.8)	43.8	-4.8 (38.7, 39.2)	-5.9 (38.8, 39.3)	38.2
500	2.8 (19.7, 19.8)	3.7 (19.8, 19.9)	19.6	-3.0 (17.3, 17.5)	-4.0 (17.3, 17.5)	17.1
1000	0.9 (13.9, 13.9)	1.1 (13.9, 14.0)	13.9	-1.0 (12.2, 12.3)	-1.1 (12.2, 12.3)	12.1
5000	0.1 (6.2, 6.2)	0.2 (6.2, 6.2)	6.2	-0.1 (5.4, 5.5)	-0.1 (5.4, 5.5)	5.4
			Set-up E			
50	-32.9 (14.1, 17.9)	-41.1 (14.8, 19.0)	11.0	40.0 (22.4, 24.1)	51.2 (22.8, 24.9)	21.0
100	-7.2 (9.6, 10.2)	-9.0 (10.0, 10.5)	7.7	8.1 (15.5, 16.4)	9.7 (15.8, 16.7)	14.8
500	-5.1 (3.7, 3.7)	-6.2 (3.7, 3.8)	3.5	6.0 (6.8, 7.0)	6.8 (6.9, 7.1)	6.6
1000	-1.3 (2.5, 2.5)	-1.4 (2.5, 2.5)	2.4	1.4 (4.8, 4.8)	1.5 (4.8, 5.0)	4.7
5000	-0.1 (1.1, 1.1)	-0.1 (1.1, 1.1)	1.1	0.1 (2.1, 2.1)	0.2 (2.1, 2.2)	2.1
			Set-up F			
50	10.2 (63.2, 64.0)	19.8 (63.3, 64.7)	62.0	-10.4 (55.0, 0.155)	-19.0 (55.3, 55.8)	54.0
100	4.9 (44.0, 44.6)	6.0 (44.3, 44.8)	43.8	-5.3 (38.7, 39.1)	-6.1 (38.8, 39.3)	38.2
500	3.0 (19.7, 19.8)	4.0 (19.8, 19.9)	19.6	-3.8 (17.2, 17.5)	-4.3 (17.3, 17.4)	17.1
1000	0.9 (13.9, 14.0)	1.2 (13.9, 14.0)	13.9	-0.8 (12.1, 12.2)	-1.0 (12.2, 12.3)	12.1
5000	0.1 (6.2, 6.2)	0.1 (6.2, 6.2)	6.2	-0.1 (5.4, 5.5)	-0.1 (5.4, 5.5)	5.4

2.4.1.2 Choice of α for finite samples

Although asymptotic results were established for all $\alpha > 1/2$, what a proper choice of α should be for finite samples is a very pertinent question. We performed extensive simulations for different sample-sizes, to explore the robustness of different choices of α values.

We tried a sample situation with one change-point, $\beta_0^0 = 0.3$, $\beta_1^0 = 1.5$, $\beta_2^0 = 1$ and $\sigma = 0.1$ with covariate X-space $= [0, 1]$. The τ-values were varied between 0 and 1 and the Mean Square Errors were plotted against $\log_{10} \alpha$ values for various sample-sizes. We found that the M.S.E. vs $\log_{10} \alpha$ graphs are almost invariant with changing sample-sizes. To change the signal-to-noise ratio, the β-values were kept constant but σ was changed to 0.5 (Fig.2.2) and 1. The patterns are exactly similar for all parameter values and signal-to-

noise ratios. However, for a small n and a very large value of α, the algorithm occasionally breaks down because $\dot{\mathbb{U}}_n(\theta)$ becomes (almost) singular for computational purposes. This is clearly indicated by the very large average MSE for $\alpha = 50$ or 100 when sample-size is small ($n = 50$). So, very large α's (greater than 10) are not recommended for small samples (less than 50).We would also like to point out the robustness of the M.S.E.'s to the choice of tuning parameter in the range of α's for which the algorithm is numerically stable. This is reflected by the flat stretch of the M.S.E. curves for each n, before numerical instability sets in. In other words, so long as the algorithm works, any choice of $\alpha > 1/2$ is essentially as effective as any other. So, searching for an optimal α is unlikely to yield any significant gains. Our recommendation is to use $\alpha = 1$, which works very well in terms of M.S.E. for all sample-sizes, as low as 30. The same α value (1) is used for all data anlyses in the subsequent sections. The simulations indicate that computational efficiency is insensitive to the choice of α. A more detailed version of Fig.2.2 is provided in Appendix A.6.

2.4.2 Longitudinal set-up

Simulations were conducted for the longitudinal case as well to compare the efficiency of our proposed method to the search-based algorithm. We considered an AR(1) correlation structure with $\rho = 0.6$ to model the dependence among observations within subject. For each subject, we considered 10 observations in scenarios G and H (Table 2.4). For set-up J, we considered varying number of observations for each individual, which is uniformly distributed over integer-space $\{1, 2, \ldots, 20\}$. Error standard deviation σ was taken to be equal to 0.1 for all cases and $M = 1$. For all simulations, α was taken to be 1. The computational efficiency of our proposed method is huge compared to the search-based algorithm, as in the cross-sectional case (Table 2.1). So, in Table 2.4, we have just compared the bias and standard errors to illustrate the validity and estimation efficiency of our method.

Results from Table 2.4, clearly indicate that our method yields almost the same standard error estimates as the search-based algorithm. Although for both methods with small sample-

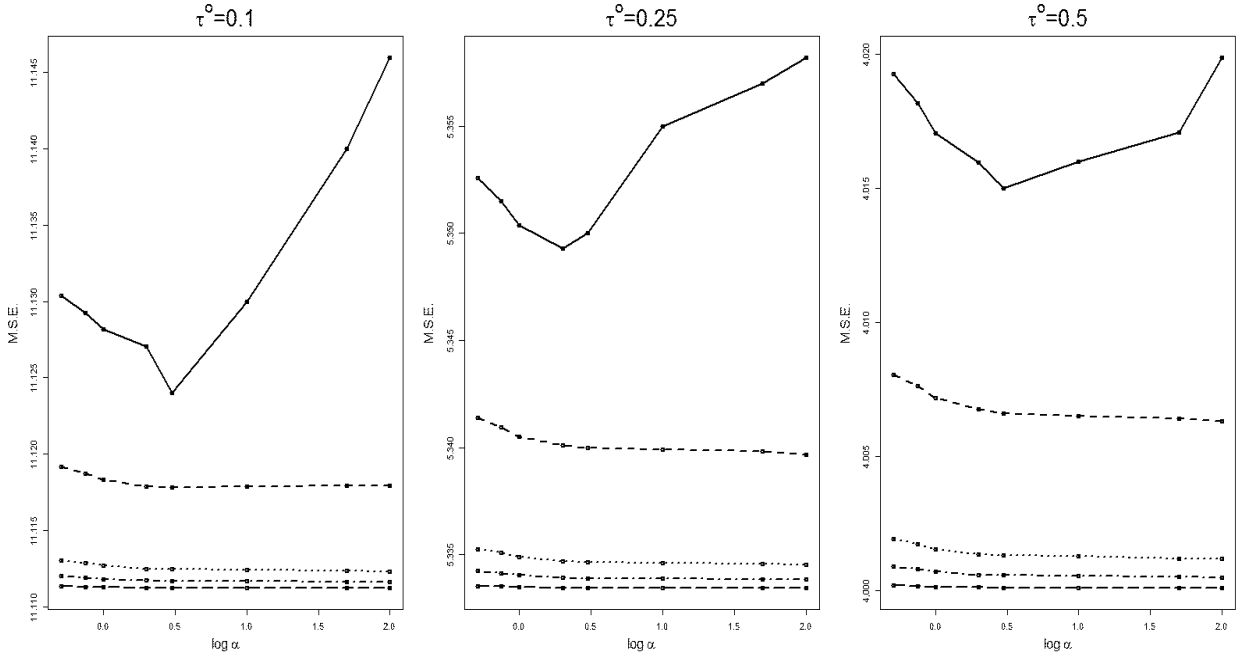

Figure 2.2: Mean Square Errors vs $\log_{10}\alpha$ for varying sample-sizes with different τ-values, where $\beta_0^0 = 0.3$, $\beta_1^0 = 1.5$, $\beta_2^0 = 1$ and $\sigma = 0.5$. From the top below, the solid line corresponds to $n = 50$, dashed line corresponds to $n = 100$, the dotted line corresponds to $n = 500$, the dot-dash line corresponds to $n = 1000$ and the longdash line corresponds to $n = 5000$.

sizes, the bias is comparatively high and the standard deviation estimates are higher than the theoretical values, the differences become smaller for larger sample sizes. The M.S.E.'s for the slope and intercept parameters also behave similar to those of the change-points.

2.5 Applications

2.5.1 Plant growth data analysis

Vernalization, a requirement for plants to experience a period of cool conditions to accelerate flowering, is an important determinant of flowering date in winter wheat. In *Brooking and Jameison* (2002), controlled environment studies were carried out to quantify the response of vernalization rate to temperature for two near-isogenic lines of the wheat cultivar

Table 2.4: Bias and variances for the change-point estimate $\hat{\tau}_n$ compared for two change-points problem in 3 longitudinal setups: G: $\theta^{\mathrm{T}} = (0.3, 1, 1, 1, 0.2, 0.8)$, H: $\theta^{\mathrm{T}} = (0.2, 1, 2, 1, 0.4, 0.6)$ & J: $\theta^{\mathrm{T}} = (0.2, 1, 2, 1, 0.4, 0.6)$. 10 observations per individual in set-ups G and H. In set-up J, number of observations per individual $D \sim$ Discrete Uniform $\{1, 2, \ldots, 20\}$(S.D.: Average of estimated standard deviations over 1000 replications; Emp. S.D. : Sample standard deviation based on 1000 replications).

Sample Size (n)	$\hat{\tau}_{1n}$ Bias (S.D., Emp. S.D.) $\times 10^{-3}$ Existing	Proposed	Theo. S.D. $\times 10^{-3}$	$\hat{\tau}_{2n}$ Bias $\times 10^{-3}$ (S.D., Emp. S.D.) $\times 10^{-3}$ Existing	Proposed	Theo. S.D. $\times 10^{-3}$
			Set-up G			
10	5.4 (74.2, 74.6)	6.9 (74.3, 74.8)	62.8	-5.3 (56.7, 57.3)	-6.6 (56.8, 58.3)	43.2
50	3.2 (27.7, 27.9)	4.0 (27.8, 27.9)	21.4	-3.3 (21.3, 21.5)	-4.3 (21.3, 21.5)	19.8
100	1.1 (19.8, 20.0)	1.4 (19.9, 20.1)	16.0	-1.2 (15.3, 15.4)	-1.5 (15.4, 15.5)	14.1
500	0.2 (7.7, 7.7)	0.3 (7.9, 8.2)	6.9	-0.2 (6.4, 6.5)	-0.2 (6.4, 6.5)	5.9
			Set-up H			
10	-7.5 (13.6, 13.2)	-9.7 (13.8, 13.5)	8.7	8.1 (25.5, 26.4)	9.8 (25.8, 27.7)	17.1
50	-5.3 (6.7, 6.7)	-6.8 (6.7, 6.8)	4.6	6.3 (9.9, 10.2)	7.1 (9.8, 10.1)	8.2
100	-1.4 (4.5, 4.4)	-1.6 (4.5, 4.5)	2.9	1.4 (6.8, 6.4)	1.7 (6.8, 6.1)	5.5
500	-0.1 (2.8, 2.8)	-0.1 (2.8, 2.8)	1.9	0.1 (4.1, 4.1)	0.2 (4.1, 4.2)	2.8
			Set-up J			
10	5.8 (77.1, 77.6)	7.1 (77.2, 77.8)	64.4	-6.1 (57.7, 59.4)	-6.8 (57.2, 59.1)	44.1
50	3.5 (28.5, 28.7)	4.1 (28.5, 28.7)	22.9	-3.9 (22.0, 22.2)	-4.4 (21.8, 22.0)	20.6
100	1.2 (20.3, 20.6)	1.4 (20.4, 20.6)	16.6	-1.3 (15.7, 15.8)	-1.5 (15.8, 15.9)	14.9
500	0.2 (7.9, 8.0)	0.3 (8.0, 8.2)	7.2	-0.2 (6.8, 6.9)	-0.2 (6.8, 6.9)	6.4

Batten: Spring Batten, vernalization insensitive; and Winter Batten, vernalization sensitive. Plants were sampled for dissection at intervals during the treatment and post-treatment period, until the flag leaf could be distinguished. The authors investigated the co-ordination of primordium initiation and leaf appearance, quantified by the Haun stage. The authors observed that Spring Batten plants grown under fully inductive conditions, 25/20°C, 16 hrs photoperiod, produced eight leaves on average, and the rate of primordium initiation per emerged leaf increased markedly with the transition from leaf initiation to spikelet initiation. This represents an important phase transition in the growth of the plant. From Fig. 2.3, it is quite clear that the model which best fits the scenario is a broken-stick model with two change-points. The authors had estimated the change-points by naked eye and then fitted three line-segments for the three regions. We provide a fast as well as statistically rigorous

analysis using the approach developed in this paper.

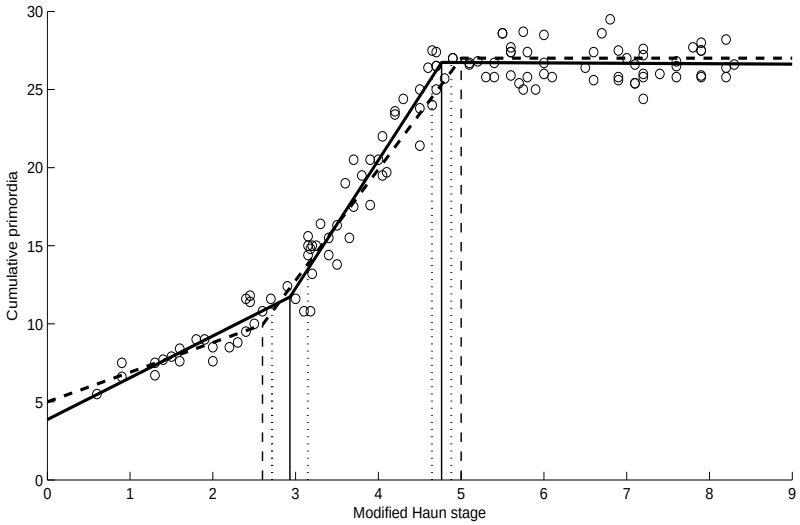

Figure 2.3: Co-ordination of primordium initiation and leaf emergence from Spring Batten treatments resulting in a final leaf number of 8 *Brooking and Jameison* (2002). The solid bold line represents the one estimated by our approach while the broken line represents the one estimated by *Brooking and Jameison* (2002). The dotted vertical lines give the confidence intervals for the estimated change-points given by the solid lines while the vertical broken-lines indicate the eye-estimated change-points.

The change-point estimates of *Brooking and Jameison* (2002) by naked eye were 2.6 and 5 on the Haun stage scale, whereas ours are $2.931(2.715, 3.147)$ and $4.764(4.647, 4.881)$, with 95% confidence intervals provided in parentheses. From Fig. 2.3 we see that the estimates in *Brooking and Jameison* (2002) do not lie within our confidence intervals, emphasizing the importance of a principled analysis such as the one we have proposed. The main conclusion in *Brooking and Jameison* (2002) was that the rate of primordium initiation per emerged leaf, the slope parameter, jumped from 1.9 primordia per leaf to 7.11 primordia per leaf and then became constant. Our estimates of the slopes of the three segments are $2.67(2.46, 2.88)$, $8.19(7.84, 8.54)$ and $-0.02(-0.16, 0.12)$ primordia per leaf. Our estimates, qualitatively, corroborate their conclusion that there are two sharp phase transitions in the

growth pattern whereby the initial growth rate gets more than tripled and then becomes more or less constant.

2.5.2 Estradiol hormone profile analysis

We applied our proposed method to analyze the longitudinal estradiol data as discussed in Section 2.1. For our purpose, we considered only women whose Final Menstrual Period (FMP) had already been observed. This was done so as to avoid scenarios with censored FMP's (*Lu et al.*, 2010). Among all these women, eight were left out either because their observed FMP was too early or too late or had less than three data-points. The remaining sample of $n = 156$ women with identified FMP was our sample of interest who in total gave 1396 observations, with each woman contributing 3 to 10 observations over time, covering 11 years before to 10 years after FMP. This gave an average of about 8.95 observations per woman. There were 75(48%) smokers at baseline and the baseline BMI mean(SD) was 27.4(6.56). Please note that the data we use here have longer follow-up and hence more subjects with identified FMP, compared to the data on which the analysis in Fig. 1 in *Sowers et al.* (2008) is based. A log transformation was applied to the Estradiol hormone level to make the normality assumption more plausible.

Denote by Y_{ij} the jth log-transformed E2 value measured at day t_{ij} centered around FMP T_i, for the ith woman and by $SMOKE_i$ and BMI_i baseline smoking habit (0 meaning smoker at baseline, 1 otherwise) and the baseline body mass index, centered at the grand mean, respectively. We consider the following model:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_2 f(X_{ij}, \tau_1) + \beta_3 f(X_{ij}, \tau_2) + \lambda_1 SMOKE_i + \lambda_2 BMI_i + b_i + U_i(t_{ij}) + \varepsilon_{ij} \quad (2.5)$$

where $X_{ij} = t_{ij} - T_i$, the b_i are random intercepts following a $N(0, \phi)$ distribution, the $U_i(t)$ are mean zero Gaussian processes modeling serial correlation and ε_{ij} are independent measurement errors following a $N(0, \sigma^2)$ distribution. We assume $U_i(t)$ is a nonhomogenous

OrnsteinUhlenbeck process, which satisfies $\text{Var}(U_i(t)) = \xi(t)$ where $log\xi(t) = \xi_0 + \xi_1 t + \xi_2 t^2$ and $\text{corr}(U_i(t), U_i(s)) = \rho^{|t-s|}$. We also assume that for each i, ε_i, b_i and $U_i(t)$ are independent of one another. Further, we assume, $-11.9 \leq \tau_1 < \tau_2 \leq 9.9$ (in general, we assume for all our theoretical results that the covariate X is contained in some compact interval $[M_1, M_2]$; here from the nature of the study and previous work we knew the scope of the study was between 12 years before and 10 years after the FMP) and that $10^{-6} \leq |\beta_2| \leq 10^6$ and $10^{-6} \leq |\beta_3| \leq 10^6$ for the sake of identifiability. Also, the variance function part does not include any mean function parameters and so even in the presence of unknown change-points, the model remains identifiable.

As illustrated in section 2.3, we estimate the regression parameters in a three step procedure. In the first step, we assume working independence to estimate $\hat{\theta}_n^{(I)}$. Then $\eta = (\phi, \sigma^2, \xi_0, \xi_1, \xi_2, \rho)$ is estimated by maximizing the conditional log-likelihood,

$$l(\eta) = -1/2 \sum_{i=1}^{n} \left[(Y_i - \mu_{n,i}^{(I)})^T \Sigma(\eta) (Y_i - \mu_{n,i}^{(I)}) + log\left|\Sigma^{(i)}(\eta)\right| \right]. \tag{2.6}$$

Therefore, $\hat{\Sigma}_n = \Sigma(\hat{\eta}_n)$ which is subsequently used in Step 3 to obtain $\hat{\theta}_n$. Condition 2.4 is verified to hold for this model; in fact W here turns out to be $\Sigma_0 = \Sigma(\eta^0)$. The proof for this is provided in Appendix A.7.

Our results indicate the presence of change-points at -2.174 $(-2.554, -1.794)$ and 1.733 $(1.513, 1.953)$ years (Table 2.5).

Table 2.5: Regression parameter estimates along with their respective standard errors

Parameter	Estimate	Standard Error
β_0	4.116	0.139
β_1	-0.006	0.002
β_2	-0.259	0.009
β_3	0.199	0.008
τ_1	-2.192	0.197
τ_2	1.738	0.11
λ_1	0.047	0.072
λ_2	0.005	0.005

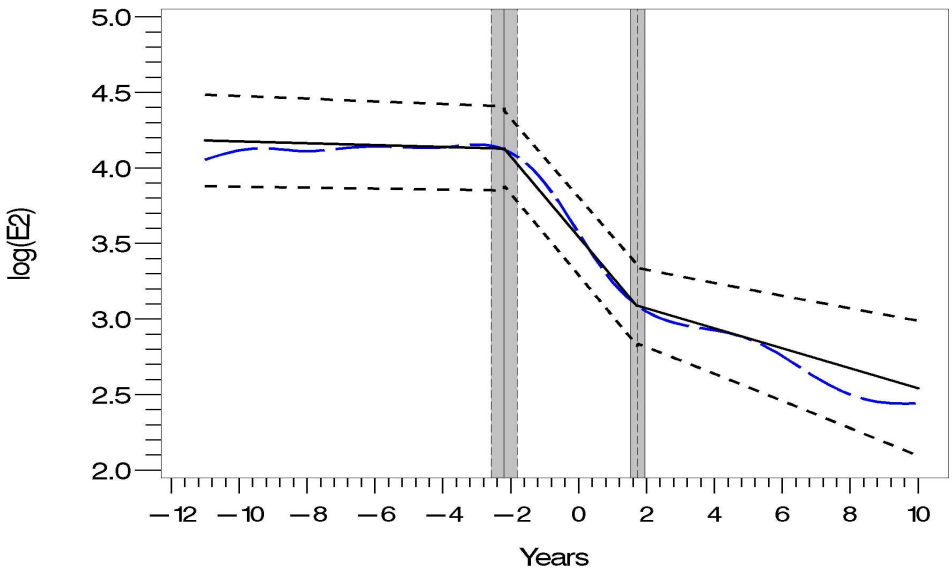

Figure 2.4: E2 profile analysis at baseline mean BMI for a non-smoker: the solid line represents the mean estimator using two change-point broken-stick model, the short-broken lines the corresponding pointwise 95% confidence bands; the long-broken lines represent the smooth estimator of the mean function from semiparametric mixed effects model using the same method as in *Sowers et al.* (2008); the shaded regions represent the 95% confidence intervals for the two change-points.

In *Sowers et al.* (2008), the change-points had been roughly thought to be around 2 years before and after FMP. Although this was a good estimate, we can see that actually the 95% confidence interval for the second change point does not contain 2 years after FMP, indicating the change of estradiol levels actually happen slightly sooner than anticipated in *Sowers et al.* (2008). Also BMI and smoking habits do not seem to alter this pattern significantly. But, our contribution, above all, is providing statistically meaningful inference about the change-points. Also the form of the confidence bands indicate that a two-change point model is indeed a good fit for the E2-hormone profile.

2.6 Discussion

We have proposed a method of estimating change-points in a broken stick model which is computationally much more efficient than existing methods, and demonstrated that it is

asymptotically as efficient. The method of estimation is also numerically stable. An added advantage of this method is that, as shown in section 2.3, it can be readily extended to generalized linear models with repeated measures, examples of which are abundant. The estimates in those frameworks have shown similar desirable asymptotic properties.

It seems reasonable to assume that this idea should work equally efficiently in estimating change-points in a time-series framework, at least under short-range dependence. For instance, estimation of change-points is of considerable significance in climatic series data (*Lund et al.*, 1995; *Lund and Reeves*, 2002) and such data sets tend to be really large. Hence our idea would likely prove even more economic in this setting. This is underscored by the fact that even for a sample size of 50, our method is more than a hundred times faster than the exact least squares method with multiple change points, and at large samples, thousands of times faster. Also, in a linear spline model with knot-locations unknown (number of knots known), the proposed method provides a faster alternative for locating these knots.

We cannot stress enough that this is a very generic idea which can be used for computational economy in several settings without giving up on asymptotic efficiency. For example, the same idea should be applicable for estimating change-points in a multivariable setup where the change-points are observed in more than one variable. While, for a search based algorithm, the computational time will increase many folds with the number of variables having change-points, it will scale much more favorably for our approach.

However, we would like to point out if the investigators feel that the linearity of broken-stick model is not best suited for their data, our method of estimation or for that matter any method of estimation based on the broken-stick model may not be reliable.

Also if the coefficients of two consecutive regimes are very close, then trying to fit separate segments for the two regimes is strongly discouraged. We performed extensive simulations and both our approach and the search-based approach yield poor estimates. Thus before fitting a broken-stick model, we would strongly suggest the investigators check that the assumptions for the model are valid.

In this article, we were interested in modeling the mean hormone profile of all subjects in the cohort discussed in Section (2.5.2). A possible way to model individual-specific hormone profiles is via multi-path change-points models. Major work done in regards to multi-path change-points include *Joseph and Wolfson* (1993) and *Asgharian and Wolfson* (2001). Most of this literature has treated change-point as the observation at which a transition has occurred, rather than a point in the $X-$space. Broken-stick models with random change-points and random intercept-slopes is a possible interesting avenue for future work in this field. The simplest possible model with one change-point is:

$$E(Y|X) = \beta_0 + \beta_1 X + \beta_2 (X - \tau)^+,$$

where $\theta^T = (\beta_0, \beta_1, \beta_2, \tau)$ follows, say, a multivariate $N(\theta^0, \Upsilon)$ distribution. Estimating methods will rely on minimizing criterion functions involving several integrals and is beyond the scope of this work.

Although this chapter focuses only on estimating change-points in a situation where their exact number is known beforehand, this approach can also be deployed for detection of change-points, where likelihood ratio type test-statistics can be computed much faster in comparison to search-based algorithms.

CHAPTER 3

High-dimensional Inference based on Multivariate Adaptive Elastic-net for Multiple Pollutant Data

3.1 Introduction

While studying the adverse health effects of air pollution, it is common practice to assess the effects of composite mass as a single pollutant. But single-pollutant modeling in air-pollution epidemiology does not suffice for gaining significant insight into understanding the exact association of air pollution with adverse health effects, i.e. exactly what biological mechanisms are linked with pollutants, and thus provide scientific support for certain regulatory public health guidelines. Estimating the adverse health effects in presence of multiple pollutants can aid significantly (*Dominici et al.*, 2010; *Johns et al.*, 2012; *Brown et al.*, 2007). Air pollution is not a single mass, rather a composite of ambient particles,gases and vapors whose compositions vary spatio-temporally and depend on a variety of issues (for instance, meteorological conditions). Thus, clearly, treating air-pollution as a single mass will lead to missing out on much of the information inside the data.

Fine particulate matter ($PM_{2.5}$; aerodynamic diameter < 2.5 micron) has been one of the most frequently studied pollutants in air-pollution epidemiology. Recent studies have shown high ambient levels of $PM_{2.5}$ are associated with cardiovascular morbidity and mortality (*Min et al.*, 2009; *Brook et al.*, 2010). Individuals with the metabolic syndrome (MetS) are

believed to be more susceptible to the adverse health effects of this type of pollutant (*Min et al.*, 2009; *NCEP*, 2001; *Brook and Rajagopalan*, 2012).

Using a rat model of experimental MetS, *Wagner et al.* (2014) hypothesized that the cardiovascular responses caused by $PM_{2.5}$ would be higher among individuals with diet-induced MetS. However, compared with traditional mass-based PM standards,identifying the most harmful constituent elements will assist policy makers in developing better targeted air pollution regulations. But, teasing out the exact health effects of constituent elements of the complex mixture of ambient $PM_{2.5}$ remains challenging.

In this study, four male rats were fed high fructose diet (HFrD) to induce MetS and four were fed normal diet (ND) and then exposed in real time to concentrated ambient paricles (CAPs) for nine days. Data related to several cardiovascular end-points (Heart Rate, Systolic Blood Pressure, Diastolic Blood Pressure, Pulse Pressure, Mean Arterial Pressure, Temperature and QA Interval) were recorded at five-minute intervals from 7:30 am to 3:30 pm each day. The concentration of 28 constituent elements in the concentrated ambient $PM_{2.5}$ were measured between the same hours of the day, but at thirty minute intervals. Extensive CAP characterization, including use of a Semicontinuous Elements in Aerosol Sampler (SEAS), was performed, and positive matrix factorization was applied to investigate source factors. SEAS (*Kidwell and J.M.*, 2001, 2004) uses high-resolution inductively coupled plasmamass spectrometry (ICP-MS) to perform every thirty minutes multi-elemental analysis of $PM_{2.5}$ samples. The increased temporal resolution of the dataand thus the number of data pointscoupled with repeated measures on individual animals, increases the statistical power to observe an effect. U.S. Environmental Protection Agency (EPA) Positive Matrix Factorization (PMF) 3.0 (*Agency)*, 2010) was used to investigate source factors; PMF is a variant of factor analysis that constrains factor loadings and factor scores to nonnegative values and has been described in detail in *Paatero and Tapper* (1994). For more details on these procedures, please refer to *Morishita et al.* (2011) and *Rohr et al.* (2011).

The cardiovascular measures were averaged over thirty minute intervals to make the

elements and responses data conformable. From the nature of the gathered data, it is quite evident that there is an AR(1) type correlation between consecutive cardiovascular measures. Two simple modeling strategies facilitate the understanding of the effects of these constituent elemental pollutants:

(i) One approach looks at the simple association between a single element and a single response in each diet group separately. The major problem with this type of naive modeling is firstly, that it gives us only the marginal association and secondly, more than a thousand such models exist, in totality, leading to multiple hypotheses testing issues.

(ii) The other simple way is to jointly model all the 28 pollutants in a single multiple linear model (one for each diet group). However, the elemental pollutants are all quite highly correlated (median absolute correlation=0.43, maximum absolute correlation=0.81) and thus this model has inherent multicollinearity issues.

Thus the principal challenge in estimating the health effects of multiple pollutants is to devise a modeling strategy which can handle the multicollinearity issues as well as the high-dimensional nature of the data. Classification and regression tree (CART) (*Breiman et al.*, 1984; *Hastie et al.*, 2001), Deletion/ Substitution/Addition (DSA) (*Haight et al.*, 2010; *Sinisi and van der Laan*, 2004), Supervised principal component analysis (SPCA) (*Roberts and Martin*, 2006; *Bair et al.*, 2006), Partial least-squares regression (PLSR) (*Wold et al.*, 2001), Bayesian model averaging (BMA) (*Hoeting et al.*, 1999; *Raftery*, 1996; *Raftery et al.*, 1997) and Least absolute shrinkage and selection operator (LASSO)(*Mauderly et al.*, 2010,?) have been suggested as possible methodologies to determine the health effects of multiple pollutants. *Sun et al.* (2013) provides a detailed comparison of these methods.

Variable-selection methods like LASSO and elastic-net are suitable for dimension reduction in these problems, however *Sun et al.* (2013) notes that inference based on these methods are not trivial. However, it is interesting to note that the adaptive versions of these variable-selection methods actually do have asymptotic normality of the estimated effect-sizes of the selected variables (*Zou and Zhang*, 2009). Also, it is important to point out that

adaptive elastic-net (*Zou and Zhang*, 2009) can handle correlated predictors better than adaptive LASSO (*Zou*, 2006). Of course, because we have several correlated cardiovascular end-points, multivariate adaptive elastic-net (*Caner and Zhang*, 2014) is going to be more efficient for estimating the health effects of the pollutants.

On the basis of these ideas, we propose a general inferential method based on multivariate adaptive elastic-net for correlated high-dimensional autoregressive set-ups in Section 3.2. The effectiveness of this method is verified via simulations in Section 3.3. Finally, in Section 3.4, we analyze the multiple pollutant data for the rat study described previously.

3.2 Methods

Consider the general case where we have k types of responses and n individuals. Also there are p pollutants. We will be considering the following model:

$$y_j^{(t)} = \alpha_j y_j^{(t-1)} + X^{(t)}\beta_j + \varepsilon_j^{(t)}; \;\; j = 1, \ldots, k; \;\; t = 1, \ldots, T,$$

where $y_j^{(t)}$ is the j^{th} response variable and $X^{(t)}$ is the matrix of pollutant concentrations at t^{th} time-point.

For notational convenience, let us define the following:

$$y^{(t)} = \begin{pmatrix} y_1^{(t)} \\ \vdots \\ y_k^{(t)} \end{pmatrix}, \varepsilon^{(t)} = \begin{pmatrix} \varepsilon_1^{(t)} \\ \vdots \\ \varepsilon_k^{(t)} \end{pmatrix},$$

$\beta^T = (\beta_1^T, \ldots, \beta_k^T)$ and $\alpha^T = (\alpha_1, \ldots, \alpha_k)$.

Also, let $Z_1^{(t)}$ =Diag$(y^{(t-1)}, \ldots, y^{(t-1)})$ and $Z_2^{(t)}$ =Diag$(X^{(t)}, \ldots, X^{(t)})$. Finally, defining

$Z^{(t)} = (Z_1^{(t)} : Z_2^{(t)})$ and $\theta^T = (\alpha^T, \beta^T)$, the above model simplifies to

$$y^{(t)} = Z^{(t)}\theta + \varepsilon^{(t)}.$$

Without loss of generality we assume that $y^{(t)}$ is centered so that we have no intercept in the model. Also assume $\text{Cov}(\varepsilon_i^{(t)}, \varepsilon_j^{(t)}) = \sigma_{ij}$ for $i, j = 1, \ldots, k$. Clearly, $\sigma_{ij} = \sigma_{ji}$. Now, define Σ_{ij} as an $n \times n$ matrix with all entries zero except the first entry which is equal to σ_{ij}, $i, j = 1, \ldots, k$. Clearly, the dispersion matrix of $\varepsilon^{(t)}$ is the $nk \times nk$ block matrix Σ, with Σ_{ij} the blocks. Also, $\theta^{0T} = (\alpha^{0T}, \beta^{0T})$ is the true parameter vector.

Now, our variable selection method is divided into three steps:

In the first step, we assume working independence of the responses i.e. $\Sigma = I_{nk}$. Then we use adaptive elastic net (*Zou and Zhang*, 2009) to estimate the parameter vector θ i.e.

$$\hat{\theta}_n^{(I)} = (1 + \frac{\lambda_2}{n}) \arg\min_{\theta} \big\{ \sum_{t=1}^{T} ||y^{(t)} - Z^{(t)}\theta||_2^2 + \lambda_2||\theta||_2^2 + \lambda_1 \Sigma_{j=1}^{(k+1)p} \hat{w}_j|\theta_j| \big\},$$

where $\{\hat{w}_j\}_{j=1}^{(k+1)p}$ are adaptive weights as described in *Zou and Zhang* (2009). Note that, $\{\hat{w}_j\}_{j=1}^{p}$ are all set as zero (i.e. no penalization on α) to maintain the AR(1) correlation structure.

In the next step, we use $\hat{\theta}_n^{(I)}$ obtained from previous step to calculate the residuals and thereby estimate the σ_{ij}'s and hence the Σ matrix.

Finally, we estimate θ by minimizing the weighted adaptive elastic net criterion function as described by *Caner and Zhang* (2014) i.e.

$$\hat{\theta}_n = (1 + \frac{\lambda_2}{n}) \arg\min_{\theta} \big\{ \sum_{t=1}^{T} (y^{(t)} - Z^{(t)}\theta)^T \hat{\Sigma}_n^{-1} (y^{(t)} - Z^{(t)}\theta) + \lambda_2||\theta||_2^2 + \lambda_1 \Sigma_{j=1}^{(k+1)p} \hat{w}_j|\theta_j| \big\}.$$

Again in this situation, to preserve the AR(1) correlation structure, $\{\hat{w}_j\}_{j=1}^{p}$ are all set as zero. For notational consistency, we let $\hat{\theta}_n^T = (\hat{\alpha}_n^T, \hat{\beta}_n^T)$.

We make the same assumptions as in *Zou and Zhang* (2009) and *Caner and Zhang* (2014):

$(A1)$ We use $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ to denote the minimum and maximum eigen values of a positive definite matrix M, respectively. Then, we assume

$$b \leq \lambda_{\min}\left(\frac{1}{n}(Z^{(t)})^T(Z^{(t)})\right) \leq \lambda_{\max}\left(\frac{1}{n}(Z^{(t)})^T(Z^{(t)})\right) \leq B,$$

where b and B are two positive constants.

$(A2)$ $\lim_{n\to\infty} \frac{\max_{i=1,\ldots,n}\sum_{j=1}^{p}((Z^{(t)}))_{ij}^2}{n} = 0$;

$(A3)$ $E\left[|\varepsilon|^{2+\delta}\right] < \infty$ for some $\delta > 0$.

$(A4)$ $\lim_{n\to\infty} \frac{\log(p)}{\log(n)} = \nu$ for some $0 \leq \nu < 1$.

$(A5)$ $\lim_{n\to\infty} \frac{\lambda_2}{n} = 0$ and $\lim_{n\to\infty} \frac{\lambda_1}{\sqrt{n}} = 0$.

$(A6)$ $||\hat{\Sigma}_n^{-1} - W||_2^2 \to^P 0$, where W is finite and positive.

Under assumptions $(A1)-(A6)$, following the theoretical results of *Zou and Zhang* (2009) and *Caner and Zhang* (2014), we get,

(i) Consistency in Selection:

$$P(A_n = A) \to 1$$

where $A = \{j : \beta_j^0 \neq 0\}$ *is the set of true signals and* $A_n = \{j : \hat{\beta}_{n,j} \neq 0\}$ *is the set of estimated signals.*

(ii) Asymptotic normality:

$$\delta^T \frac{I + \lambda_2\Omega_A^{-1}}{1 + \lambda_2/n}\Omega_A^{1/2}(\hat{\beta}_{n,A} - \beta_A^0) \to N(0,1)$$

in distribution, where $\Omega_A = Z_{2,A}^T\Sigma^{-1}Z_{2,A}$ *is a matrix of dimension equal to the cardinality of* A*, say* p_A *and* δ *is a vector of norm* 1.

On the set $\{A_n = A\}$,

$$\delta^T \frac{I + \lambda_2\Omega_A^{-1}}{1 + \lambda_2/n}\Omega_A^{1/2}(\hat{\beta}_{n,A} - \beta_A^0) = \delta^T \frac{I + \lambda_2\Omega_{A_n}^{-1}}{1 + \lambda_2/n}\Omega_{A_n}^{1/2}(\hat{\beta}_{n,A_n} - \beta_{A_n}^0),$$

where $\Omega_{A_n} = Z_{2,A_n}^T \hat{\Sigma}_n^{-1} Z_{2,A_n}$ is a matrix of dimension equal to the cardinality of A_n, say p_{A_n} and δ is a vector of norm 1.

This means that, on the set $\{A_n = A\}$,

$$\delta^T \frac{I + \lambda_2 \Omega_{A_n}^{-1}}{1 + \lambda_2/n} \Omega_{A_n}^{1/2} (\hat{\beta}_{n,A_n} - \beta_{A_n}^0) \xrightarrow{d} N(0,1). \tag{3.1}$$

Now from *(i)*, we have $P(A_n = A) \to 1$. This implies that, (3.1) holds, where $\Omega_{A_n} = Z_{2,A_n}^T \hat{\Sigma}_n^{-1} Z_{2,A_n}$ is a matrix of dimension equal to the cardinality of A_n, say p_{A_n} and δ is a vector of norm 1.

We are going to use this asymptotic result for our inferential procedures.

3.3 Simulations

Extensive simulations were conducted to evaluate the performance of this inferential procedure. We considered $k = 10$, $n = 10$ and $T = 100$. The value of p was varied over $\{30, 50, 80\}$. α_j^0 was taken to be 0.5 for all $j = 1, \ldots, 10$. All σ_{ij}'s $(i \neq j)$ were fixed at 0.6 while σ_{ii}'s were taken as 1. We considered $\beta_j^0 = (2.5, 2.5, 2.5, -2.5, -2.5, 0, 0, \ldots, 0)$ for all $j = 1, \ldots, 10$ (note the dimension of β_j^0 varies with changing p).

To model the correlation between the covariates, rows of X are considered as iid realizations from $\mathrm{N}(0, \Omega)$. The following correlation structures were looked at:

A. $\Omega_{ii} = 1$ for all $i = 1, \ldots, p$, $\Omega_{ij} = 0.8$ for $i, j = 1, \ldots, 10, i \neq j$ and $\Omega_{ij} = 0.1$ for $i, j = 11, \ldots, p, i \neq j$.

B. $\Omega_{ij} = (0.9)^{|i-j|}$, $i, j = 1, \ldots, p$.

C. $\Omega_{ii} = 1$ and $\Omega_{ij} = 0.8$, $i, j = 1, \ldots, p, i \neq j$.

Now, we used our inferential procedure to calculate confidence intervals for each of the β_j's in each of the nine situations (three choices of correlation structure and three choices of p). In each situation, the experiment was repeated 1000 times. A particular co-ordinate

of β was defined as a signal if its 95% confidence interval did not contain 0. CI_j is the 95% confidence interval of β_j, $j = 1, \ldots, 10p$. The performance of the procedure was evaluated by two separate measures:

I. Coverage Probability (CP) at j^{th} co-ordinate: $\hat{\mathbb{P}}(\beta_j^0 \in CI_j)$ and

II. True Positives (TP) at j^{th} co-ordinate: $\hat{\mathbb{P}}(0 \notin CI_j | \beta_j^0 \neq 0)$ and False Positives (FP) at j^{th} co-ordinate: $\hat{\mathbb{P}}(0 \notin CI_j | \beta_j^0 = 0)$,

where $\hat{\mathbb{P}}$ is the empirical probability. In each situation, for each β_j, these two measures were calculated. We also provide the simulation results for the same data-sets using simple adaptive elastic-net for one response at a time. The results have been summarized in Table 3.1.

From the high percentages of the True Positives measure for the proposed method, it is evident that the method is very well equipped to identify the correct signals. Also, for such highly correlated set-ups, the method provides a fair control over detection of false positives (maximum of False Positives measure below 20%). Finally, the coverage probabilities are not very far away from the nominal coverage probability (95%), indicating the efficiency of the inference procedure. Comparing the results in Table 3.1, we can clearly see that there is a very nominal gain in False Positives and Coverage Probability measures but a sizeable loss in terms of True Positives measures, which clearly indicates the effectiveness of the proposed inference procedure.

3.4 Rat-data analysis

We used our proposed inferential strategy to analyze the rat-data as discussed in Section 3.1. We have seven cardiovascular end-points and 28 elemental pollutant concentrations, measured at 30-minute intervals. We have two diet-groups (HFrD and ND), each of which we are going to model separately. So, for our example, $k = 7$, $n = 4$, $p = 28$ and $T = 71$. The results of the analysis are presented in Figures 3.1-3.7.

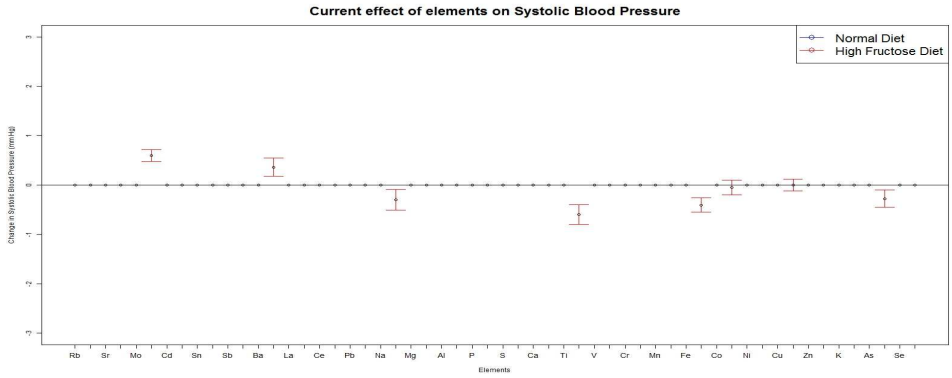

Figure 3.1: Confidence intervals of effects of standardized pollutant concentrations on Systolic Blood Pressure

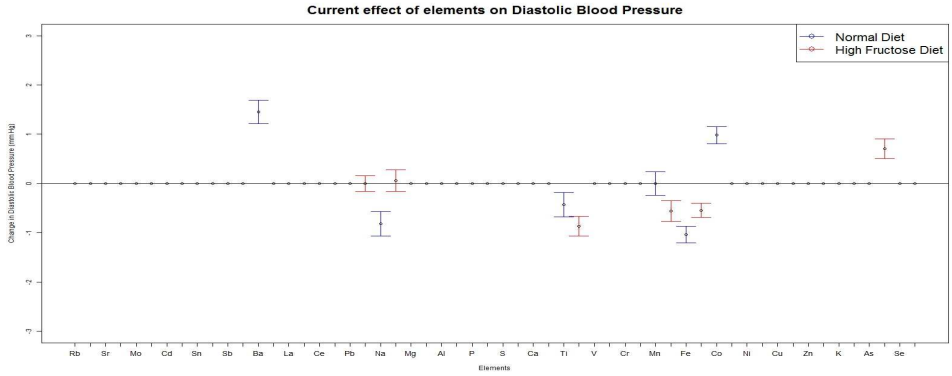

Figure 3.2: Confidence intervals of effects of standardized pollutant concentrations on Diastolic Blood Pressure

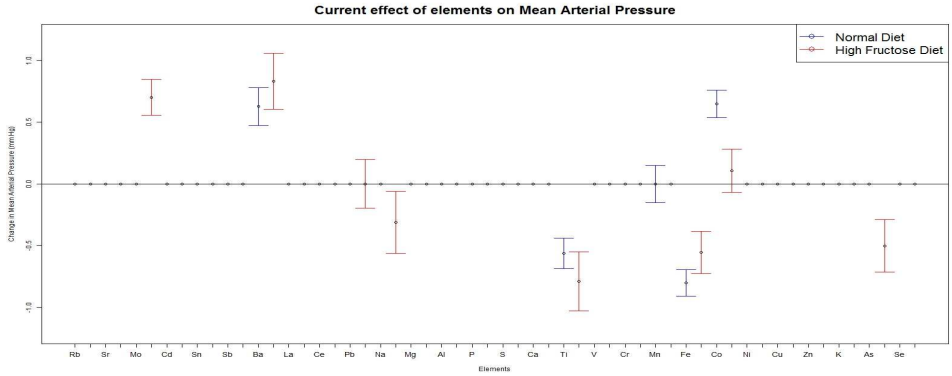

Figure 3.3: Confidence intervals of effects of standardized pollutant concentrations on Mean Arterial Pressure

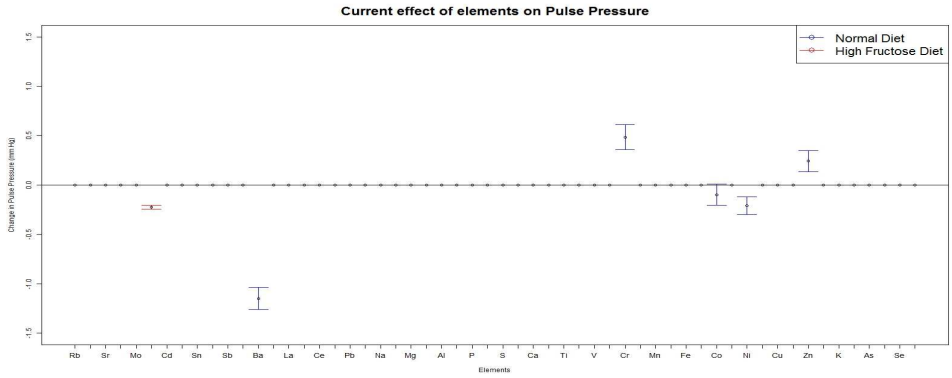

Figure 3.4: Confidence intervals of effects of standardized pollutant concentrations on Pulse Pressure

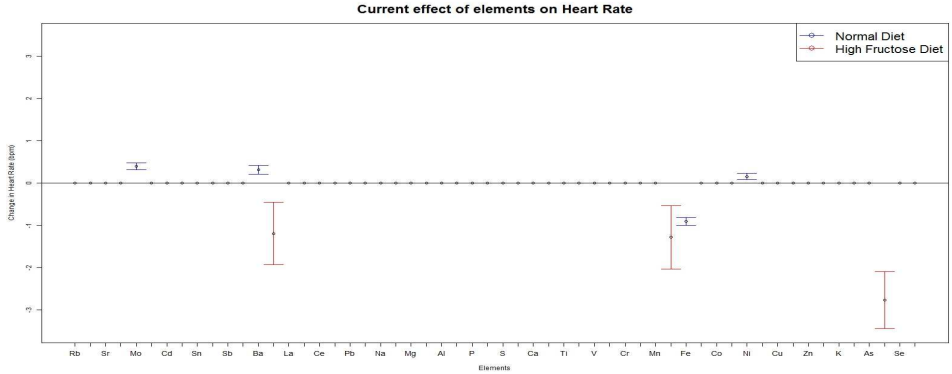

Figure 3.5: Confidence intervals of effects of standardized pollutant concentrations on Heart Rate

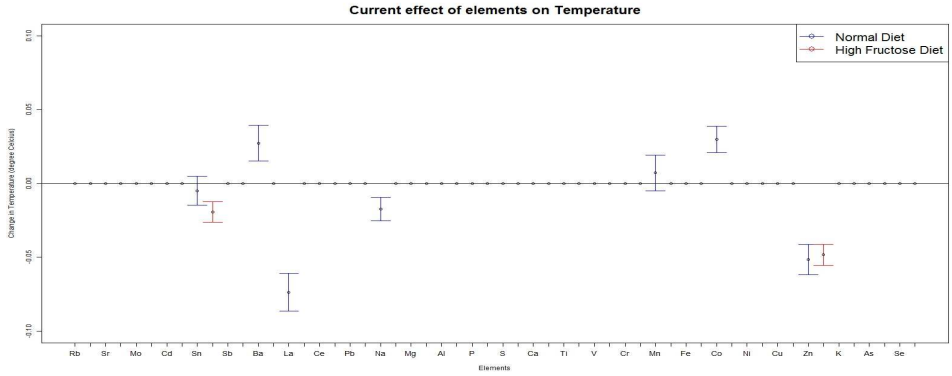

Figure 3.6: Confidence intervals of effects of standardized pollutant concentrations on Temperature

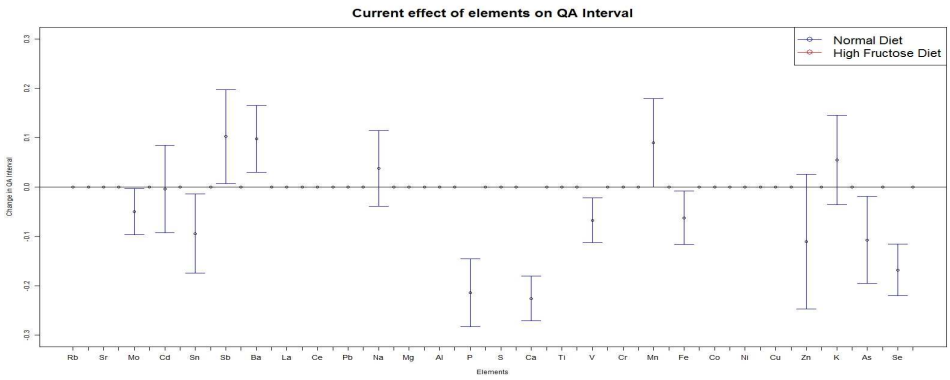

Figure 3.7: Confidence intervals of effects of standardized pollutant concentrations on QA Interval

Note that, the figures contain the confidence intervals of only the variables selected by the procedure; the variables which are not selected are set to zero, and hence by definition, they are not significant.

A major concern with such data-sets where we have so many correlated pollutants is whether the statistical significance observed is truly significant or merely a statistical artifact. Thus, possibly we could have a scenario where two positively correlated pollutants, none of which are actually signals, are both deemed to be significant in the analysis, albeit having opposite signs. Simply speaking, in this situation we have two false positives. To address this concern, we would like to point out that our simulations indicate even for such strongly correlated data-sets this procedure does provide a fair amount of control on the rate of false positives. Also, as a post-analysis sensitivity check we tried the following– if two highly correlated pollutants were both deemed to be significant, we randomly dropped one of the two pollutants. Then we reran the entire procedure with the remaining 27 pollutants to see how sensitive the behavior of the other pollutant was to this. There was, understandably, a slight change in the estimated effect sizes and confidence intervals for the pollutants but there were no major changes to the significance status or signs of the pollutant effects (only twice there were changes in significance status, but both times the pollutants were marginally

significant). This type of sensitivity analysis verifies that our results are not statistical artifact but are 'true' results.

Given that a single emission source produces particles that contain multiple elements, it is not surprising that we observed correlations with two or more elements. The site of the current study in an industrial area of Dearborn is within a mile of a steel mill, automotive assembly facilities, and an oil refinery. As such our findings of associations of iron, barium, titanium, cobalt and molybdenum are consistent with the high concentration of metal-associated industries in the area.

Consistent with our biological hypothesis, our results demonstrate significant differences in the effect of the pollutants in the two diet groups. Also, specific pollutants are associated with an increase in a certain cardiovascular measure (e.g. Barium with an increase of systolic blood pressure in the HFrD diet group) while some others are associated with a decrease in the same (e.g. Iron is linked with a decrease of systolic blood pressure in the HFrD diet group). This outcome should not be interpreted to indicate one element is harmful while the other is protective. It is worth remembering that either an increase or a decrease of a cardiovascular end-point like systolic blood pressure can be harmful. Furthermore these elements may be an indicator of a certain type of PM combustion source (e.g., metal coating, steel processing, or oil combustion, etc), and may not themselves be mediating the cardiovascular effect. The statistical interpretation of these results, while quite straightforward, need to be incorporated cautiously into the calculation for biological plausibility to inform the assessment of PM2.5 exposure for cardiovascular health risks.

3.5 Discussion

The aim of this work was to propose a high-dimensional inference method for multiple pollutants model and then use that procedure to analyze this important data-set.

Note that, because this inference procedure is based on adaptive elastic-net, this method

will always work for effective $p <$ effective n. However, for the effective $p >$ effective n scenario, this method may not work. There are certain situations when even in this set-up this procedure might work, but that is difficult to say beforehand (for details please refer to *Zou and Hastie* (2005)). In our analysis, we have ignored the interaction effects of pollutants for two main reasons. Firstly, including interaction effects would have led us to a situation where $n < p$ and also because interpretation of variable-selection based results in the presence of interactions is not straightforward. For instance, one might run into a situation where the main effect of a pollutant is not selected but the interactions are, in which case the interpretation is not valid. This can be avoided by doing a group-penalty like in group LASSO (*Yuan and Lin*, 2006; *Meier et al.*, 2008). However, statistical inference based on this type of method is not known.

So, we believe before including such terms in the model and making the inference procedure harder, it is worth interpreting the results of this simpler model carefully and understanding the biological mechanisms of pollutant toxicity and thus be able to guide public health regulatory standards.

Table 3.1: Simulation results comparing the proposed method with inference method based on simple adaptive elastic-net: min:= minimum over all the co-ordinates, max:= maximum over all the co-ordinates and mean:= mean over all the co-ordinates

p	Corr. Struct.	Proposed method									Simple adaptive elastic-net based method								
		TP (%)			FP (%)			CP (%)			TP (%)			FP (%)			CP (%)		
		min	max	mean	min	max	mean	min	max	mean	min	max	mean	min	max	mean	min	max	mean
30	A	90.4	100	96.9	7.3	15.4	11.5	84.6	96.0	89.2	79.2	93.2	85.7	7.0	15.0	11.3	85.0	97.1	89.9
	B	91.8	100	98.2	5.9	9.2	7.1	90.8	96.4	93.6	82.7	94.9	87.2	5.7	8.9	6.9	91.1	97.7	94.2
	C	89.9	99.9	96.0	8.8	17.3	13.7	82.7	95.8	87.2	77.1	90.7	82.4	8.6	16.9	13.6	83.1	97.0	88.1
50	A	90.1	99.9	96.2	8.2	15.9	12.8	84.1	95.3	87.4	78.7	92.1	84.9	8.0	15.5	12.7	84.5	96.2	88.4
	B	91.2	100	97.1	6.7	10.0	8.4	90.0	95.8	91.7	81.3	93.8	86.8	6.5	9.7	8.2	90.3	96.8	92.6
	C	89.0	99.7	94.9	11.6	17.5	15.1	82.5	95.1	85.2	76.1	89.9	83.0	11.5	17.3	14.9	82.7	96.0	86.4
80	A	89.4	99.8	95.8	9.3	17.1	13.5	82.9	95.0	86.6	77.9	90.7	84.0	9.0	16.8	13.3	83.2	95.9	87.9
	B	90.2	99.9	96.4	7.8	10.8	9.1	89.2	95.2	90.4	80.1	92.9	86.2	7.6	10.6	8.8	89.4	96.2	91.7
	C	88.8	99.7	94.0	12.9	18.6	16.0	81.4	94.8	84.1	74.8	87.8	80.9	12.7	18.3	15.9	81.7	95.7	85.6

CHAPTER 4

Variable selection for high-dimensional broken-stick regression

4.1 Introduction

As we have discussed in Chapter 2, the regression function need not have the same parametric form in the entire domain of interest. The broken-stick is a specific model of interest, where the regression function is comprised of line-segments, joined at certain unknown change-points. The broken-stick with r change-points is defined in Eq. (2.1). For the serum estradiol hormone profile analysis that we discussed in detail in Chapter 2, we fit a broken stick with two change-points. The main reasons for using this model were the results from *Sowers et al.* (2008b) and also the fact that scientists felt that there were changes in the hormone secretion twice, once before and one after the Final Menstrual Period (FMP). However, a major limitation of the Michigan Bone Health and Metabolism Study (MBHMS) was that several of the subjects did not have records for more advanced years, i.e. more than 8 years after FMP. With a more detailed study, where we have more observations in the latter years, epidemiologists think that there might be a third or even a fourth change-point. This scenario raises the question: do we know how many change-points we should include in our model?

To get a better insight, first let us consider the standard linear model,

$$y_i = x_i^{\mathrm{T}}\beta + \varepsilon_i;\ i = 1, \ldots, n, \tag{4.1}$$

where for the $i-$th subject, y_i is the observed response, x_i is the $p-$dimensional observed covariate (or design points associated with y_i) and ε_i's are independent and identically distributed random errors with mean 0 and unknown variance σ^2; β is a $p-$dimensional vector of unknown parameters.

For this model, the theory is very well established for the traditional set-up where p is fixed and $p < n$. However, with advancements in modern technology in several biological, medical and economics-related fields, we have data-sets with growing p, comparable or even larger than n. This makes standard inferencial procedures invalid.

A rich literature exists on variable selection procedures, i.e. identifying the non-zero entries of β in (4.1), for the case of $p < n$, as well as for $p > n$, especially in the last decade. For further details, see *Fan and Peng* (2004); *Hunter and Li* (2005); *Meinshausen and Bhlmann* (2006); *Zhao and Yu* (2006); *Zou* (2006); *Wang et al.* (2007); *Fan and Lv* (2008); *Zhang and Huang* (2008); *Meinshausen and Yu* (2009); *Wang* (2009); *Fan and Lv* (2010).

So returning to our motivating example of the serum estradiol hormone-profile analysis, one possible solution is to start with a complicated overfit model, say 10 change-points and then do some sort of variable selection to see which of these change-points are actually true. To illustrate this point further, if one takes a look at the model (2.1), it is quite clear that β_{j+1} and τ_j, $j = 1, \ldots, r$ are very closely related; in the sense that if for some j, $\beta_{j+1} = 0$, then there is no change in regime i.e. for that particular j, τ_j is a nonsensical parameter. This means that if we can do some sort of variable selection on the β's, which are changes of slope parameters, we should be able to identify which of the change-points are actually meaningful and be able to select the best model, i.e., the one with the appropriate number

of change-points. We introduce a problem which has the same flavor, but is more suitable to introduce our variable selection procedure. We believe this procedure can be extended to solve the problem discussed above.

4.1.1 Problem Formulation

The problem we are going to look into is when we have p covariates, each of which could possibly have a broken-stick, with one change-point, effect on Y. We assume of all the p covariates, several have no effects, some have simple linear effects and the remaining have broken-stick, with one change-point, effects. Mathematically speaking, this means,

$$y_i = \beta_0 + \sum_{j=1}^{p} m_{\theta_j}(x_{ij}) + \varepsilon_i,\ i = 1, \ldots, n, \tag{4.2}$$

where,

$$m_{(\beta_1,\beta_2,\tau)}(x) = \beta_1 x + \beta_2 f(x, \tau). \tag{4.3}$$

Here, $f(x, \tau) = (x - \tau)^+$, as defined in (2.1) and $\theta_j = (\beta_{1j}, \beta_{2j}, \tau_j)$. Also, we define $\theta = (\beta_0, \theta_1, \theta_2, \ldots, \theta_p)^{\mathrm{T}}$ as the $(3p + 1)-$dimensional vector of all the unknown parameters.

We assume that $X \in [0, 1]$. From the above representation, it appears as if we are assuming that each of the p covariates has one change-point. For instance, say, x_j has a change-point at τ_j where the linear effect size of x_j changes from β_{1j} to $\beta_{1j} + \beta_{2j}$. But, for the model we described above, we assumed that not all the regressors have a broken-stick type effect. So, for the sake of identifiability, we assume that if, for example, x_k does not have a broken-stick type effect, $\beta_{2k} = 0$ and $\tau_k = 1/2$. This leads to the following characterization

of θ:

$$\begin{aligned} \text{no effect} &\Rightarrow \beta_1 = 0, \beta_2 = 0, \tau = 1/2, \\ \text{effect with no change-point} &\Rightarrow \beta_1 \neq 0, \beta_2 = 0, \tau = 1/2 \text{ and} \\ \text{broken-stick effect with one change-point} &\Rightarrow \beta_2 \neq 0, \tau \in (0,1). \end{aligned} \tag{4.4}$$

We assume that most of the X_j's, $j = 1, \ldots, p$, do not have an effect on the Y which, of course, means that most of the β's are zero. This agrees perfectly with the usual assumption of sparsity that we get for high-dimensional regression problems. No literature exists on variable selection for such regression models. Thus, providing a proper variable selection procedure is of particular importance in this set-up.

The commonly used variable selection procedures such as LASSO, elastic-net, SCAD, etc will not readily work for this model. We illustrate why these methods fail to work in this set-up in the following section (4.2). In Section 4.3, we introduce post local smoothing thresholded ridge regression as an effective variable selection procedure for this model. In Section 4.4, we provide simulation results which indicate the effectiveness of the proposed method. Finally, in Section 4.6, we try to indicate our ideas about how we plan to develop theoretical results for our approach and possible extensions of our methods.

4.2 Difficulty with existing variable selection procedures

From equations (4.2) and (4.3), we can see that our basic objective is to minimize the least squares criterion:

$$\sum_{i=1}^{n} [y_i - \beta_0 - \sum_{j=1}^{p} \{\beta_{1j} x_{ij} + \beta_{2j} f(x_{ij}, \tau_j)\}]^2 \tag{4.5}$$

under sparsity assumptions. The total number of unknown parameters in this model is $3p+1$. For now, we assume that $n > 3p + 1$.

Our aim is to solve this problem from a penalized regression point of view, i.e., to minimize the regularized least squares criterion. So we define,

$$\hat{\theta}_n = \arg\min_{\theta} \sum_{i=1}^{n} [y_i - \beta_0 - \sum_{j=1}^{p} \{\beta_{1j} x_{ij} + \beta_{2j} f(x_{ij}, \tau_j)\}]^2 + p_\lambda(\beta, (\tau - \frac{1}{2})), \tag{4.6}$$

where β is the $(2p+1)$-dimensional vector of β_0, all β_{1j}'s and β_{2j}'s, while τ is the p-dimensional vector of all τ_j's and p_λ is an appropriate penalty depending on the unknown parameter λ.

It is quite clear that because of the sparsity assumption, we want most of the β-parameters to be zero, hence the penalty on the β's. It is worth noting that we are overfitting, in some sense, i.e. using a one change-point broken-stick model when in truth, say, that particular x_j has a simple linear effect. In this kind of a scenario, because of the intrinsic co-dependence of the β_2 and the τ parameters, we need to be cautious. We might run into a situation where τ_j is estimated very close to one of the boundaries and the β_{2j} parameter is assigned a nonsensical estimate or the other way round, β_{2j} is estimated very close to zero and the τ_j parameter is assigned a meaningless estimate. We have avoided this type of identifiability issue by our definitions in (4.4). Thus, we want to make sure that the τ_j-estimates are always well separated from the boundary and in case β_{2j} is estimated at zero (or very close to zero), then the τ_j estimate automatically becomes $1/2$ (or very close to $1/2$). Thus our aim is to shrink the $\beta-$estimates towards zero, for sparsity, and shrink the $\tau-$estimates towards $1/2$. This is the motivation behind putting a penalty on $|\tau_j - 1/2|$.

Now, the least squares criterion in (4.5) is clearly not differentiable because the function $f(x, \tau)$ is not differentiable at τ. In general, we find that a usual necessity for most of the commonly used variable selection procedures, reliant on coordinate descent type algorithms, is to have a differentiable loss function (*Loh and Wainwright*, 2015). We aim to resolve this by replacing the $f-$ function in the least squares criterion by the twice-differentiable

smooth local approximation q_n as we had done in Chapter 2. This makes our loss function differentiable; but it still needs to satisfy at least the restricted strong convexity assumption as in *Loh and Wainwright* (2015); *Negahban et al.* (2012); *Agarwal et al.* (2012); *Wang et al.* (2014) to use the popular variable selection approaches such as LASSO, elastic-net or SCAD. It is not clear if this assumption is going to be satisfied by our loss function and thus we cannot justify the usage of these procedures.

4.3 Post local smoothing thresholded ridge regression

From arguments given in the previous section, it is clear that it is not possible to use the traditional variable selection procedures for our set-up. However, from the previous arguments we also understand that we do need some kind of penalty for shrinking the β estimates towards zero and τ estimates towards $1/2$. We choose a ridge regression type penalty as this is a smooth penalty which allows us to use simple Newton-Raphson type algorithms for finding the minimum of the penalized regression criterion function, by solving for the zero of the score function, corresponding to this penalized regression objective function. However, clearly, this does not perform variable selection; to that end we use hard thresholding. The algorithm is described below:

Step I: We replace the $f-$function in (4.5) by the twice differentiable local approximation q_n, where q_n is as in Chapter 2. The functional form of q_n is:

$$q_n(x,\tau) = \begin{cases} 0, & \text{if } x < \tau - \gamma_n; \\ \frac{(x-\tau+\gamma_n)^2}{4\gamma_n}, & \text{if } \tau - \gamma_n \leq x \leq \tau + \gamma_n; \\ (x-\tau), & \text{if } x > \tau + \gamma_n; \end{cases} \tag{4.7}$$

where γ_n is a deterministic sequence, that approaches zero as $n \to \infty$ (Figure 2.1).

Step II: We choose a ridge-regression type penalty, i.e.,

$$\tilde{\theta}_n = \arg\min_{\theta} \sum_{i=1}^{n} [y_i - \beta_0 - \sum_{j=1}^{p} \{\beta_{1j} x_{ij} + \beta_{2j} q_n(x_{ij}, \tau_j)\}]^2 + \lambda [\beta_0^2 + \sum_{j=1}^{p} \{\beta_{1j}^2 + \beta_{2j}^2 + (\tau_j - \frac{1}{2})^2\}]. \quad (4.8)$$

We look at the score function corresponding to the criterion function in Eq. (4.8). We solve for this score function by Newton-Raphson algorithm, as we did in Chapter 2 and define the resulting estimate as $\tilde{\theta}_n$.

Step III: Now we do a hard thresholding on the β's, i.e. $\hat{\beta} = \tilde{\beta}\mathbf{1}(|\tilde{\beta}| < a)$. Note that if $|\tilde{\beta}_{1j}| \geq a$ and $|\tilde{\beta}_{2j}| < a$, then x_j has only linear effect, while if both $|\tilde{\beta}_{1j}| < a$ and $|\tilde{\beta}_{2j}| < a$, then x_j has no effect. If $|\tilde{\beta}_{2j}| < a$, then we make $\hat{\tau}_j = 0.5$. This is our final estimate, $\hat{\theta}_n$.

Step IV: We choose the tuning parameters λ and a by cross-validation. For our simulations, in the following section, we use 5-fold cross-validation and minimize the average prediction mean squared error. Here, we are basically adopting ideas from the thresholded ridge regression procedure for linear models, with fixed design, as described in *Shao and Deng* (2012).

4.4 Simulation study

We considered model (4.2) with normally distributed ε_i. To vary the signal-to-noise ratio, the standard deviation of ε_i was varied over the following three set-ups: $\sigma = 0.1$ (A), 0.3 (B), 0.6 (C). Three sets of variables and sample sizes were also considered, $(p, n) = (30, 100)$, $(100, 320)$ and $(300, 920)$. A set of $X_1, \ldots, X_n$ were independently generated with X_i coming from a truncated (in $(0, 1)$) multivariate normal distribution with $(\mu = 0.5 * \mathbf{1}, \Sigma)$, where the diagonal elements of Σ are all equal to $1/6$ and off-diagonal elements of Σ are all equal to 0.1.

Set-up 1: $(p, n) = (30, 100)$. Of these 30 $x-$variables, the first two have a broken-stick, with single change-point effect ($\beta_{1j} = 1$, $\beta_{2j} = 0.5$ for $j = 1, 2$ and $\tau_1 = 0.4$ and $\tau_2 = 0.6$),

the next three have only linear effects ($\beta_{1j} = 1$, $\beta_{2j} = 0$ and $\tau_j = 0.5$, for $j = 3, 4, 5$) and the last 25 do not have any effects (i.e. $\beta_{1j} = \beta_{2j} = 0$ and $\tau_j = 0.5$, for $j = 6, 7, \ldots, 30$).

Set-up 2: $(p, n) = (100, 320)$. Of these 100 $x-$variables, the first five have a broken-stick, with single change-point effect ($\beta_{1j} = 1$, $\beta_{2j} = 0.5$ and $\tau_j = 0.1(j+2)$ for $j = 1, \ldots, 5$), the next five have only linear effects ($\beta_{1j} = 1$, $\beta_{2j} = 0$ and $\tau_j = 0.5$, for $j = 6, \ldots, 10$) and the last ninety do not have any effects (i.e. $\beta_{1j} = \beta_{2j} = 0$ and $\tau_j = 0.5$, for $j = 11, 12, \ldots, 100$).

Set-up 3: $(p, n) = (300, 920)$. Of these 300 $x-$variables, the first 15 have a broken-stick, with single change-point effect ($\beta_{1j} = 1$, $\beta_{2j} = 0.5$ and $\tau_j = 0.1(1+8j/15)$ for $j = 1, \ldots, 15$), the next 15 have only linear effects ($\beta_{1j} = 1$, $\beta_{2j} = 0$ and $\tau_j = 0.5$, for $j = 16, 17, \ldots, 30$) and the last 270 do not have any effects (i.e. $\beta_{1j} = \beta_{2j} = 0$ and $\tau_j = 0.5$, for $j = 31, 32, \ldots, 300$).

For all the set-ups, β_0 is taken as 5. The proposed algorithm described in Section 4.3 is then used to perform variable selection for the simulated data-sets and the experiment is repeated 1000 times for each set-up. We noted for each set-up, the percentage of times the method is able to identify the variable accurately, i.e. whether it is a signal with one change-point, or a signal with no change-point or just a noise term. The high percentages of correct variable identification, as noted in Table 4.1, indicate the effectiveness of the proposed method.

In Table 4.2, we have recorded the bias and standard deviations of the estimates over the 1000 replications for Set-up 1. We can see there seems to be a small systematic bias present for the estimates. However, because the accuracy of variable identification is so high, as illustrated in Table 4.1, we can see that there is almost no bias in the extra parameters (i.e. for the signals with no change-point, the β_2 parameter is almost bias-free and similarly for the noise terms both the β parameters are almost bias free). This makes sense because we are able to identify the variables accurately; hence in both these situations most of these noise-type parameters are thresholded to zero. The same pattern is repeated in Tables 4.3 and 4.4 where the biases are recorded for Set-ups 2 and 3. It appears from the simulation results the bias does decrease with increasing sample-size(n). However, even if the estimates

Table 4.1: Percentage of times the method was able to identify the variable correctly

	Signal with single change-point	Signal with no change-point	Noise term
	Set-up A		
Set-up 1	99.42	99.60	99.15
Set-up 2	99.49	99.61	99.23
Set-up 3	99.64	99.60	99.28
	Set-up B		
Set-up 1	95.83	96.65	96.83
Set-up 2	95.98	96.82	96.97
Set-up 3	96.04	96.91	97.02
	Set-up C		
Set-up 1	92.26	92.86	92.47
Set-up 2	92.43	93.05	92.70
Set-up 3	92.64	93.19	92.89

Table 4.2: Bias and standard deviations, in parentheses, of the estimates over 1000 replications for Set-up 1

		A	B	C
Single change point signals	β_{11}	0.04 (0.02)	0.09 (0.03)	0.17 (0.06)
	β_{21}	-0.06 (0.03)	-0.10 (0.05)	-0.16 (0.07)
	τ_1	0.02 (0.01)	0.02 (0.01)	0.03 (0.01)
	β_{12}	0.04 (0.02)	0.10 (0.03)	0.18 (0.06)
	β_{22}	-0.06 (0.03)	-0.10 (0.04)	-0.17 (0.07)
	τ_2	-0.02 (0.01)	-0.03 (0.01)	-0.03 (0.01)
No change point signals	β_{13}	0.05 (0.02)	0.08 (0.03)	0.17 (0.06)
	β_{14}	0.04 (0.02)	0.08 (0.03)	0.16 (0.06)
	β_{15}	0.04 (0.02)	0.09 (0.03)	0.18 (0.06)
	β_2	0.0001	0.0004	0.0008
Noise terms	β_1	0.0002	0.0006	0.0011
	β_2	< 0.0001	< 0.0001	0.0001

Table 4.3: Bias and standard deviations, in parentheses, of the estimates over 1000 replications for Set-up 2

		A	B	C
Single change point signals	β_{11}	0.03 (0.01)	0.08 (0.02)	0.15 (0.06)
	β_{21}	-0.06 (0.02)	-0.09 (0.05)	-0.15 (0.07)
	τ_1	0.01 (< 0.01)	0.01 (0.01)	0.02 (0.01)
	β_{12}	0.04 (0.02)	0.10 (0.03)	0.17 (0.07)
	β_{22}	-0.05 (0.02)	-0.09 (0.04)	-0.17 (0.06)
	τ_2	0.01 (< 0.01)	0.02 (< 0.01)	0.01 (< 0.01)
	β_{13}	0.04 (0.01)	0.09 (0.03)	0.16 (0.05)
	β_{23}	-0.06 (0.03)	-0.09 (0.05)	-0.16 (0.06)
	τ_3	< 0.01 (< 0.01)	< 0.01 (< 0.01)	0.01 (< 0.01)
	β_{14}	0.03 (0.01)	0.09 (0.03)	0.17 (0.05)
	β_{24}	-0.06 (0.03)	-0.10 (0.05)	-0.16 (0.06)
	τ_4	-0.01 (< 0.01)	-0.02 (< 0.01)	-0.02 (0.01)
	β_{15}	0.03 (0.01)	0.08 (0.03)	0.16 (0.05)
	β_{25}	-0.05 (0.02)	-0.09 (0.04)	-0.17 (0.07)
	τ_5	-0.01 (< 0.01)	-0.01 (< 0.01)	-0.02 (< 0.01)
No change point signals	β_{16}	0.05 (0.02)	0.08 (0.03)	0.16 (0.06)
	β_{17}	0.05 (0.02)	0.09 (0.03)	0.16 (0.05)
	β_{18}	0.04 (0.01)	0.09 (0.03)	0.18 (0.06)
	β_{19}	0.05 (0.01)	0.08 (0.02)	0.17 (0.05)
	$\beta_{1,10}$	0.04 (0.02)	0.09 (0.03)	0.18 (0.06)
	β_2	< 0.0001	0.0002	0.0005
Noise terms	β_1	0.0001	0.0004	0.0010
	β_2	< 0.0001	< 0.0001	< 0.0001

are asymptotically unbiased for a proper choice of the tuning parameters (λ_n and a_n), it does appear that the rate at which the bias converges to zero is quite slow. This will need to be investigated in more detail.

4.5 Extension to multiple change-points model

4.5.1 The idea

We would also like to extend this method to broken-stick effects with multiple change-points, as described in Section 4.1. The principal difficulty is that we need to penalize the distance between consecutive change-points, which will lead to penalties of the form

Table 4.4: Average bias of the estimates over 1000 replications for Set-up 3

		A	B	C
Single change point signals	β_1	0.03	0.09	0.17
	β_2	-0.05	-0.10	-0.16
	τ	0.0008	0.0002	-0.0003
No change point signals	β_1	0.04	0.08	0.15
	β_2	< 0.0001	0.0001	0.0003
Noise terms	β_1	< 0.0001	0.0002	0.0007
	β_2	< 0.0001	< 0.0001	< 0.0001

$\lambda \sum_j |\tau_{j+1} - \tau_j|$. Quite clearly, this penalty is not separable in the coordinates of θ, because of which the problem becomes trickier. However, it is our understanding that if we have a good idea about the maximum possible number of change-points, then our proposed method can be extended to the multiple change-points set-up as well. For instance, let us look at the E2 hormone profile analysis that we performed in Chapter 2. If we do not believe our initial assumption that this is a broken-stick model with two change-points, then we can use a modified version of our algorithm in this chapter to determine the correct number of change-points. Say, we assume that there are, at most, possibly k change-points. Then, we modify Step II of our algorithm, as described in Section 4.3 to the following:

$$\tilde{\theta}_n = \arg\min_{\theta} \quad [\sum_{i=1}^{n} \{y_i - \beta_0 - \beta_1 x_i - \sum_{j=1}^{k} \beta_{j+1} q_n(x_i, \tau_j)\}^2 + \lambda_1 \{\sum_{j=0}^{k} \beta_j^2 + (\tau_1 - \frac{1}{k})^2 + (\tau_k - \frac{k-1}{k})^2\} + \lambda_2 \sum_{j=1}^{k-1} (\tau_{j+1} - \tau_j - \frac{1}{k})^2], \quad (4.9)$$

the idea being we do not want the change-point estimates too close to the boundaries or consecutive change-points too close to one another. After this step, the remaining steps remain unaltered. We seek to find the most parsimonious yet accurate model. We investigated the merit of this approach via some simple simulations.

4.5.2 Simulations

We considered the simple model where the mean function is a broken-stick with three change-points and normally distributed errors with mean zero and standard deviation 0.1. The covariate X is generated from a uniform distribution in (0,1). Sample-sizes (n) are varied over 50, 100, 500 and 1000. The initial guess about the maximum possible number of change-points as defined in Eq (4.9) is denoted by k; k is varied over 10, 6 and 4. For all these set-ups, the change-points' locations were fixed at 0.4, 0.6 and 0.8 and change of slope parameters equal to 0.5 (i.e. $\beta_2 = \beta_3 = \beta_4 = 0.5$). For each set-up, the experiment was repeated 1000 times and the results are presented in Table 4.5.2.

It is quite evident from the results in Table 4.5.2, for all (n, k) pairs, the approach detects the correct model with a high probability. Also, as one would expect, this proportion increases with increasing sample-sizes as well as when we have a better initial guess about the correct number of change-points in the broken-stick.

4.6 Future work

We expect to have selection consistency for the proposed algorithm in the same way as in *Shao and Deng* (2012), for proper choice of $\lambda_n \to 0$ and $a_n \to 0$. Here λ_n is the tuning parameter in the ridge regression step (Step II) while a_n is the cut-off tuning parameter in Step III. For linear regression with $n > p$, *Shao and Deng* (2012) established selection consistency for gaussian errors, i.e., for any constant $t > 0$,

$$P(\digamma_{\hat{\theta}_n, a_n} = \digamma_{\theta^0, a_n}) = 1 - O(n^{-t}),$$

where $\digamma_{\xi, c_n}$ denotes the set of indices of components of ξ whose absolute values are larger than c_n. For the $n > p$ set-up, these results, we believe, can be extended to the random design model. We aim to establish a similar result for our broken-stick model.

Firstly, for a fixed p set-up, we wish to look at the score functions corresponding to Eq. (4.8) and put everything into a general M-estimation framework, as we did in Chapter 2 to understand the asymptotic properties of $\hat{\theta}_n$. Of course, this would mean we need to understand at what rates λ_n and a_n converge to zero. Once, we have an understanding of the asymptotic behavior of our estimate for a fixed p, we hope to be able to provide good theoretical insight into a growing p set-up for this problem.

For a fixed p, when the number of change-points for the broken-stick model is known, an efficient method of estimation has been presented in Chapter 2. Variable selection procedures for large p in a linear regression set-up are very well known in literature. However, there is no known method of variable selection when the covariates can possibly have a broken-stick type effect. This is the gap in literature that we are aiming to bridge. In this chapter, we have provided a numerically efficient method for broken-stick effect with single change-point. Our immediate aim is to show theoretical efficiency of the proposed method, in terms of selection/prediction consistency, as described above.

Finally, we would like to investigate how the proposed extension of our method will work for detecting the number of change-points in a multiple change-points model. Because this approach is closely related to the problem we have looked at in more detail in this chapter, we feel this should work. However, we note that the simulation study for the multiple change-points model is quite limited and more comprehensive investigation, both on the simulation and theoretical fronts are called for. This will be pursued in the future.

Table 4.5: Distribution (in percentages) of the estimated number of change-points by the proposed approach

Estimated number of change-points	n=50			n=100			n=500			n=1000		
	$k = 10$	$k = 6$	$k = 4$	$k = 10$	$k = 6$	$k = 4$	$k = 10$	$k = 6$	$k = 4$	$k = 10$	$k = 6$	$k = 4$
≤ 1	0.7	0.7	0.3	0.3	0.2	0.2	0.1	0.0	0.0	0.0	0.0	0.0
2	14.4	14.0	8.6	12.6	11.1	8.5	10.9	10.1	7.1	10.2	7.8	5.9
3 (truth)	64.9	68.8	79.0	70.2	74.8	80.8	73.9	78.1	84.0	76.2	82.0	86.9
4	16.3	14.7	12.1	15.4	12.9	10.5	14.1	11.3	8.9	13.0	10.0	7.2
≥ 5	3.7	1.8	0.0	1.5	1.0	0.0	1.0	0.5	0.0	0.6	0.2	0.0

APPENDIX

APPENDIX A

Proofs for Chapter 2

A.1 Proof of Theorem 2.1

It is clearly seen that $U(\theta^0) = 0$, here $U(\theta) = \partial P(M(\theta))/\partial\theta$ is the population score function. The proof of consistency is based on the following two facts:

FACT 1: θ^0 is the unique minimizer of $P(M(\theta))$ in Θ.

FACT 1$*$: θ^0 is also the unique solution of $U(\theta) = 0$. Note that, this is not required for proving Theorem 2.1; however, this does imply that the estimate obtained by Newton-Raphson of the smoothed score equation converges to the true θ^0, and not to some local minima.

FACT 2: $\|\mathbb{P}_n(M_n) - P(M)\| := \sup_{\theta\in\Theta} |\mathbb{P}_n(M_n(\theta)) - P(M(\theta))| = o_p(1)$.

Now, from Fact 1, we get θ^0 is the unique minimizer of $P(M(\theta))$ in Θ. This along with Fact 2, proves consistency of $\hat{\theta}_n$ by the argmax (argmin) continuous mapping theorem (*van der Vaart and Wellner*, 1996). Note that, since θ^0 is an interior point of Θ and $\hat{\theta}_n$, the minimizer of $\mathbb{P}_n(M_n(\theta))$ is consistent for θ^0, hence $\hat{\theta}_n$ is also an interior point of Θ with probability increasing to 1. Of course, this also implies with probability increasing to 1, $\hat{\theta}_n$ is a zero of $\mathbb{U}_n(\theta)$.

To show Fact 1, observe that,

$$P(M(\theta)) = P(M(\theta^0)) + P\left[(\beta_0 - \beta_0^0) + (\beta_1 - \beta_1^0)X + \{\beta_2 f(X,\tau) - \beta_2^0 f(X,\tau^0)\}\right]^2.$$

This, of course, implies that θ^0 minimizes $P(M(\theta))$ in Θ. Now, assuming θ^1 is also a minimizer of $P(M(\theta))$, we get

$$(\beta_0^1 - \beta_0^0) + (\beta_1^1 - \beta_1^0)x + \{\beta_2^1 f(x,\tau^1) - \beta_2^0 f(x,\tau^0)\} = 0$$

holds for all $x \in [0, M]$. Putting $x = 0$, we get $\beta_0^1 = \beta_0^0$. So, now we have,

$$(\beta_1^1 - \beta_1^0)x + \{\beta_2^1 f(x,\tau^1) - \beta_2^0 f(x,\tau^0)\} = 0$$

holds for all $x \in [0, M]$. Next, if we take $0 < x < \min(\tau^0, \tau^1)$, then we get $(\beta_1^1 - \beta_1^0)x = 0 \Rightarrow \beta_1^1 = \beta_1^0$. Thus, we are now left with

$$\beta_2^1 f(x,\tau^1) - \beta_2^0 f(x,\tau^0) = 0$$

holds for all $x \in [0, M]$. Now, taking $x = \min(\tau^0, \tau^1)$, we have $\tau^0 = \tau^1$ because $\beta_2 \neq 0$. Finally, we get,

$$(\beta_2^1 - \beta_2^0) f(x,\tau^0) = 0$$

holds for all $x \in [0, M]$, implying $\beta_2^1 = \beta_2^0$. So, we have $\theta^1 = \theta^0$, implying that θ^0 is indeed the unique minimizer of $P(M(\theta))$ in Θ. We now have established Fact 1.

To establish Fact 2, observe that $|\mathbb{P}_n(M_n(\theta)) - P(M(\theta))| \leq |\mathbb{P}_n(M_n(\theta)) - P(M_n(\theta))| + |P(M_n(\theta) - M(\theta))|$. Direct calculation yields $P(M_n(\theta) - M(\theta)) = 2P[\{f(X,\tau) - q_n(X,\tau)\}\{Y - \beta_0 - \beta_1 X - \beta_2 \frac{f(X,\tau)+q_n(X,\tau)}{2}\})] = O(\gamma_n) = o(1)$. Now we show that, $(\mathbb{P}_n - P)(M_n(\theta)) = o_p(1)$. Observe that $M_n(\theta) = (Y - \beta_0 - \beta_1 X - \beta_2 q_n(X,\tau))^2$. Now, Θ being compact, it is clear

that β_0, $\beta_1 X$ and $\beta_2 q_n(X,\tau)$ are all bounded monotones as functions of θ. Theorem 2.7.5 in *van der Vaart and Wellner* (1996) shows that bounded monotone functions have a bracketing number of order $(1/\epsilon)$, wrt $L_1(P)$ norm and hence a similar bound on the covering number with respect to the same norm. So they have bounded uniform entropy integral (BUEI) property and hence belongs to a Glivenko-Cantelli class. Theorem 3 in *van der Vaart and Wellner* (1999) shows that simple operations such as addition or multiplication preserves the Glivenko-Cantelli property and hence $M_n(\theta)$ belongs to a Glivenko-Cantelli class. This implies that $(\mathbb{P}_n - P)(M_n(\theta)) = o_p(1)$ (*van der Vaart and Wellner*, 1996), which establishes Fact 2, and hence Theorem 2.1.

To show Fact 1$*$, we consider the simpler model $E(Y|X) = \beta_1 X + \beta_2 f(X,\tau)$ by absorbing the intercept into Y, and then show that $\theta^0 = (\beta_1^0, \beta_2^0, \tau^0)$ is the unique zero of $U(\theta)$, here,

$$\begin{aligned} U(\theta) &= \begin{pmatrix} -2P[X(Y - \beta_1 X - \beta_2 f(X,\tau))] \\ -2P[f(X,\tau)(Y - \beta_1 X - \beta_2 f(X,\tau))] \\ -2\beta_2 P[-\mathbf{1}(X > \tau)(Y - \beta_1 X - \beta_2 f(X,\tau))] \end{pmatrix} \\ &= \begin{pmatrix} -2P[X(\beta_1^0 X + \beta_2^0 f(X,\tau^0) - \beta_1 X - \beta_2 f(X,\tau))] \\ -2P[f(X,\tau)(\beta_1^0 X + \beta_2^0 f(X,\tau^0) - \beta_1 X - \beta_2 f(X,\tau))] \\ -2\beta_2 P[-\mathbf{1}(X > \tau)(\beta_1^0 X + \beta_2^0 f(X,\tau^0) - \beta_1 X - \beta_2 f(X,\tau))] \end{pmatrix}. \end{aligned}$$

Now to solve the system of equations $U(\theta) = 0$, we eliminate β_1 to obtain,

$$\begin{aligned} &\frac{P(Xf(X,\tau^0))P(Xf(X,\tau)) - P(f(X,\tau)f(X,\tau^0))P(X^2)}{P^2(Xf(X,\tau)) - P(f^2(X,\tau))P(X^2)} \\ &\quad = \frac{P(Xf(X,\tau^0))P(X1(X>\tau)) - P(1(X>\tau)f(X,\tau^0))P(X^2)}{P(Xf(X,\tau))P(X1(X>\tau)) - P(1(X>\tau)f(X,\tau))P(X^2)} \\ &\quad = \frac{\beta_2}{\beta_2^0}. \end{aligned} \tag{A.1}$$

From the first equality in (A.1) we obtain

$$\begin{aligned}
&P(X^2)P(1(X>\tau)f(X,\tau^0))[P(f^2(X,\tau))-P(f(X,\tau)f(X,\tau^0))]\\
&\quad+P(f(X,\tau)f(X,\tau^0))P(Xf(X,\tau))P(X1(X>\tau))\\
&\quad-P^2(Xf(X,\tau))P(1(X>\tau)f(X,\tau^0))\\
&\quad+P(1(X>\tau)f(X,\tau))P(Xf(X,\tau^0))P(Xf(X,\tau))\\
&\quad-P(f^2(X,\tau))P(X1(X>\tau))P(Xf(X,\tau^0))=0. \qquad \text{(A.2)}
\end{aligned}$$

We now show that this cannot be true for any $\tau \neq \tau^0$. Suppose $\tau > \tau^0$. Clearly,

$$P(f^2(X,\tau))-P(f(X,\tau)f(X,\tau^0))=P[f(X,\tau)(f(X,\tau)-f(X,\tau^0))]<0,$$

which implies

$$P(X^2)P(1(X>\tau)f(X,\tau^0))[P(f^2(X,\tau))-P(f(X,\tau)f(X,\tau^0))]<0.$$

Let $h_k(\tau)=P(X^k 1(X>\tau))$. We have

$$\begin{aligned}
&P(f(X,\tau)f(X,\tau^0))P(Xf(X,\tau))P(X1(X>\tau))-P^2(Xf(X,\tau))P(1(X>\tau)f(X,\tau^0))\\
&\quad+P(1(X>\tau)f(X,\tau))P(Xf(X,\tau^0))P(Xf(X,\tau))\\
&\quad-P(f^2(X,\tau))P(X1(X>\tau))P(Xf(X,\tau^0))=[h_1^2(\tau)-h_2(\tau)h_0(\tau)]\tau\tau^0(g(\tau^0)-g(\tau)),
\end{aligned}$$

where $g(\tau)=\{h_2(\tau)-\tau h_1(\tau)\}/\tau$. Now, $h_k'(\tau)=\tau h_{k-1}'(\tau)$. So, $g'(\tau)=\{h_2'(\tau)-\tau h_1'(\tau)\}/\tau^2-\{h_2(\tau)\}/\tau^2=-\{h_2(\tau)\}/\tau^2<0$. Hence, we have $g(\tau^0)-g(\tau)>0$. Also, the Cauchy-Schwarz inequality yields $h_1^2(\tau)-h_2(\tau)h_0(\tau)<0$, thus $[h_1^2(\tau)-h_2(\tau)h_0(\tau)]\tau\tau^0(g(\tau^0)-g(\tau))<0$. Hence for all $\tau>\tau^0$ the left hand side of (A.2) is negative. Similarly we can show that it is positive for all $\tau<\tau^0$. Thus (A.2) yields $\tau=\tau^0$.

Clearly from the second equality in (A.1) we have $\beta_2 = \beta_2^0$, then $\beta_1 = \beta_1^0$ and thus $\theta = \theta^0$. This proves Fact 1∗

A.2 Proof of Theorem 2.2

Since $n^{1/2}(\hat{\theta}_n - \theta^0) = n^{1/2}(\hat{\theta}_n - \theta_n) + n^{1/2}(\theta_n - \theta^0)$, the asymptotic distribution of $\hat{\theta}_n$ is a direct result of the following two facts when $\gamma_n = n^{-\alpha}$ with $\alpha > 1/2$.

FACT 3: $\|\theta_n - \theta^0\| = O(\gamma_n)$.

FACT 4: $n^{1/2}(\hat{\theta}_n - \theta_n)$ converges in distribution to $N(\mathbf{0}, 2\sigma^2 \dot{U}_*^{-1}(\theta^0))$.

We first show Fact 3. Observe that a simple Taylor series expansion of $U_n(\theta_n)$ around θ^0 yields

$$U_n(\theta_n) - U_n(\theta^0) = \begin{pmatrix} \dot{U}_{1n}(\tilde{\theta}_n^{(1)}) \\ \dot{U}_{2n}(\tilde{\theta}_n^{(2)}) \\ \dot{U}_{3n}(\tilde{\theta}_n^{(3)}) \\ \dot{U}_{4n}(\tilde{\theta}_n^{(4)}) \end{pmatrix} (\theta_n - \theta^0) = A_n(\theta_n - \theta^0),$$

where each of $\tilde{\theta}_n^{(i)}$, $i = 1, 2, 3, 4$, lies on the straight line joining θ_n and θ^0 and $\dot{U}_{in} = \frac{\partial U_n}{\partial \beta_{i-1}}$, $i = 1, 2, 3$, and $\dot{U}_{4n} = \frac{\partial U_n}{\partial \tau}$. Now, we know from the proof of Fact 1, $\sup_{\theta \in \Theta} |P(M_n(\theta)) - P(M_n(\theta))| = o(1)$, hence $\theta_n - \theta^0 = o(1)$. This in turn implies that for sufficiently large n, θ_n is an interior point of Θ and hence, a zero of $U_n(\theta_n)$.

Now, $U(\theta^0) = 0$. Thus for sufficiently large n, the above equality becomes

$$U_n(\theta^0) - U(\theta^0) = -A_n(\theta_n - \theta^0).$$

It is clearly seen from

$$
\begin{aligned}
& U_n(\theta) - U(\theta) \\
= & \begin{pmatrix}
-2\beta_2 P(f(X,\tau) - q_n(X,\tau)) \\
-2\beta_2 P[X(f(X,\tau) - q_n(X,\tau))] \\
2P[(f(X,\tau) - q_n(X,\tau))(Y - \beta_0 - \beta_1 X - \beta_2(q_n(X,\tau) + f(X,\tau)))] \\
2\beta_2 P[(-\mathbf{1}(X \geq \tau) - \frac{\partial}{\partial \tau} q_n(X,\tau))(Y - \beta_0 - \beta_1 X - \beta_2(q_n(X,\tau) + f(X,\tau)))]
\end{pmatrix},
\end{aligned}
$$

that $\|U_n - U\| = O(\gamma_n)$. Thus Fact 3 is established if A_n is invertible for large n. Now,

$$
\dot{U}_n(\theta) = \begin{pmatrix}
2 & 2P(X) & 2P(q_n(X,\tau)) & 2\beta_2 P(\frac{\partial}{\partial \tau} q_n(X,\tau)) \\
2P(X) & 2P(X^2) & 2P(X q_n(X,\tau)) & 2\beta_2 P(X \frac{\partial}{\partial \tau} q_n(X,\tau)) \\
2P(q_n(X,\tau)) & 2P(X q_n(X,\tau)) & 2P(q_n^2(X,\tau)) & a_n(\theta) \\
2\beta_2 P(\frac{\partial}{\partial \tau} q_n(X,\tau)) & 2\beta_2 P(X \frac{\partial}{\partial \tau} q_n(X,\tau)) & a_n(\theta) & b_n(\theta)
\end{pmatrix},
$$

where,

$$
\begin{aligned}
a_n(\theta) = & \ 2\beta_2 P\left\{ q_n(X,\tau) \frac{\partial}{\partial \tau} q_n(X,\tau) \right\} - 2P\left\{ (Y - \beta_0 - \beta_1 X - \beta_2 q_n(X,\tau)) \frac{\partial}{\partial \tau} q_n(X,\tau) \right\}, \\
b_n(\theta) = & \ 2\beta_2^2 P\left\{ \left(\frac{\partial}{\partial \tau} q_n(X,\tau) \right)^2 \right\} - 2\beta_2 P\left\{ (Y - \beta_0 - \beta_1 X - \beta_2 q_n(X,\tau)) \frac{\partial^2}{\partial \tau^2} q_n(X,\tau) \right\}.
\end{aligned}
$$

A direct calculation shows that all the elements of $\dot{U}_n(\theta)$ apart from $b_n(\theta)$ converge uniformly to continuous finite limits. Since $\|U_n - U\| = O(\gamma_n) = o(1)$ and θ^0 is the unique root of $U(\theta)$, we have $\theta_n \to \theta^0$, thus $\theta_n^{(i)} \to \theta^0$, $i = 1, 2, 3, 4$ as $n \to \infty$. Hence, by continuity, all the elements of A_n apart from $b_n(\tilde{\theta}_n^{(4)})$ converge to a finite limit.

To check the convergence of $b_n(\tilde{\theta}_n^{(4)})$, we write $b_n(\theta) = b_n^{(1)}(\theta) - b_n^{(2)}(\theta)$ where

$$
\begin{aligned}
b_n^{(1)}(\theta) &= 2\beta_2^2 P\left\{ \left(\frac{\partial}{\partial \tau} q_n(X,\tau) \right)^2 \right\}, \\
b_n^{(2)}(\theta) &= 2\beta_2 P\left\{ (Y - \beta_0 - \beta_1 X - \beta_2 q_n(X,\tau)) \frac{\partial^2}{\partial \tau^2} q_n(X,\tau) \right\}.
\end{aligned}
$$

It can be clearly seen that $b_n^{(1)}(\theta) \to 2\beta_2^2 P(\mathbf{1}(X > \tau))$, a continuous function, uniformly. Hence $b_n^{(1)}(\tilde{\theta}_n^{(4)}) \to 2\beta_2^{(0)^2} P(\mathbf{1}(X > \tau^0))$.

For notational simplicity we denote $\tilde{\theta}_n^{(4)} = (\tilde{\beta}_{0n}, \tilde{\beta}_{1n}, \tilde{\beta}_{2n}, \tilde{\tau}_n)$. Then,

$$\begin{aligned}
&P\left\{(Y - \tilde{\beta}_{0n} - \tilde{\beta}_{1n}X - \tilde{\beta}_{2n}q_n(X,\tilde{\tau}_n))\frac{\partial^2}{\partial\tau^2}q_n(X,\tilde{\tau}_n)\right\} \\
&\quad = P\left\{\varepsilon\frac{\partial^2}{\partial\tau^2}q_n(X,\tilde{\tau}_n)\right\} + P\left\{\frac{\partial^2}{\partial\tau^2}q_n(X,\tilde{\tau}_n)\left\{(\beta_0^{(0)} - \tilde{\beta}_{0n}) + (\beta_1^{(0)} - \tilde{\beta}_{1n})X\right.\right. \\
&\qquad \left.\left. +(\beta_2^{(0)} - \tilde{\beta}_{2n})f(X,\tau^0) + \tilde{\beta}_{2n}(f(X,\tau^0) - q_n(X,\tilde{\tau}_n))\right\}\right\} \\
&\quad = \mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV} + \mathrm{V}.
\end{aligned}$$

We then have, $\mathrm{I} = P[\varepsilon\partial^2 q_n(X,\tilde{\tau}_n)/\partial\tau^2] = 0$, and

$$\begin{aligned}
\mathrm{II} &= P\left|\frac{1}{2\gamma_n}(\beta_0^{(0)} - \tilde{\beta}_{0n})\mathbf{1}(\tilde{\tau}_n - \gamma_n \le X \le \tilde{\tau}_n + \gamma_n)\right| \\
&= |\beta_0^{(0)} - \tilde{\beta}_{0n}|P\left|\frac{1}{2\gamma_n}\mathbf{1}(\tilde{\tau}_n - \gamma_n \le X \le \tilde{\tau}_n + \gamma_n)\right| \\
&= o(1)O(1) \\
&= o(1).
\end{aligned}$$

Similarly the result of convergence to 0 can be shown for III, IV and V. Also, $\tilde{\beta}_{2n} \to \beta_2{}^{(0)}$. Thus we have $b_n^{(2)}(\tilde{\theta}_n^{(4)}) \to 0$.

So we have established $\dot{U}_n(\theta)$ converges uniformly, implying A_n converges as well. Let $\dot{U}_*(\theta^0)$ denote the limit of A_n, where

$$\dot{U}_*(\theta^0) = 2\begin{pmatrix}
1 & P(X) & P(f(X,\tau^0)) & -\beta_2^0 P(\mathbf{1}(X > \tau^0)) \\
P(X) & P(X^2) & P(Xf(X,\tau^0)) & -\beta_2^0 P(X\mathbf{1}(X > \tau^0)) \\
P(f(X,\tau^0)) & P(Xf(X,\tau^0)) & P(f^2(X,\tau^0)) & -\beta_2^0 P(f(X,\tau^0)) \\
-\beta_2^0 P(\mathbf{1}(X > \tau^0)) & -\beta_2^0 P(X\mathbf{1}(X > \tau^0)) & -\beta_2^0 P(f(X,\tau^0)) & (\beta_2^0)^2 P(\mathbf{1}(X > \tau^0))
\end{pmatrix}.$$

Now, for any vector $a = (a_1, \ldots, a_4)^{\mathrm{T}} \neq 0$, we have

$$a^{\mathrm{T}} \dot{U}_*(\theta^0) a = 2P\{a_1 + a_2 X + a_3 f(X, \tau^0) - a_4 \beta_2^0 1(X > \tau^0)\}^2 > 0,$$

which implies that $\dot{U}_*(\theta^0)$ is positive definite and hence nonsingular. Thus A_n is nonsingular for large enough n, and we have

$$|\theta_n - \theta^0| = {A_n}^{-1} \|U_n - U\| = O(\gamma_n).$$

We next show Fact 4. Denote $\mathbb{G}_n = n^{1/2}(\mathbb{P}_n - P)$. Again, observe that by Taylor Series expansion of $U_n(\theta_n)$ around $\hat{\theta}_n$ yields

$$U_n(\hat{\theta}_n) - U_n(\theta_n) = \begin{pmatrix} \dot{U}_{1n}(\theta_n^{*(1)}) \\ \dot{U}_{2n}(\theta_n^{*(2)}) \\ \dot{U}_{3n}(\theta_n^{*(3)}) \\ \dot{U}_{4n}(\theta_n^{*(4)}) \end{pmatrix} (\hat{\theta}_n - \theta_n) = B_n(\hat{\theta}_n - \theta_n),$$

where $\theta_n^{*(i)}$, $i = 1, 2, 3, 4$ lie on the straight line joining θ_n and $\hat{\theta}_n$. Since, with probability increasing to 1, $\mathbb{U}_n(\hat{\theta}_n) = U_n(\theta_n) = 0$, the following equality holds with probability increasing to 1:

$$\mathbb{U}_n(\hat{\theta}_n) - U_n(\hat{\theta}_n) = -B_n(\hat{\theta}_n - \theta_n).$$

Now,

$$
\begin{aligned}
n^{1/2}[\mathbb{U}_n(\hat{\theta}_n) - U_n(\hat{\theta}_n)] &= -2\mathbb{G}_n \begin{pmatrix} Y - \beta_0 - \beta_1 X - \beta_2 q_n(X,\tau) \\ X(Y - \beta_0 - \beta_1 X - \beta_2 q_n(X,\tau)) \\ q_n(X,\tau)(Y - \beta_0 - \beta_1 X - \beta_2 q_n(X,\tau)) \\ \beta_2 \frac{\partial}{\partial \tau} q_n(X,\tau)(Y - \beta_0 - \beta_1 X - \beta_2 q_n(X,\tau)) \end{pmatrix}_{\theta=\hat{\theta}_n} \\
&= -2\mathbb{G}_n \begin{pmatrix} g_n^{(1)}(\hat{\theta}_n) \\ g_n^{(2)}(\hat{\theta}_n) \\ g_n^{(3)}(\hat{\theta}_n) \\ g_n^{(4)}(\hat{\theta}_n) \end{pmatrix} \\
&= -2\mathbb{G}_n(g_n(\hat{\theta}_n)).
\end{aligned}
$$

Clearly, each of the $g_n^{(i)}$, $i = 1, \ldots, 4$, as a function of θ is addition and/or multiplication of bounded monotones. We know from Theorem 2.7.5 in *van der Vaart and Wellner* (1996), that bounded monotones are Donsker. Also addition/multiplication of bounded functions preserves Donsker property as in Theorem 2.10.6 in *van der Vaart and Wellner* (1996). Hence, each of the $g_n^{(i)}$, $i = 1, \ldots, 4$, belongs to a Donsker class. The same can be said about the following vector of functions

$$
g = \begin{pmatrix} g^{(1)} \\ g^{(2)} \\ g^{(3)} \\ g^{(4)} \end{pmatrix} = \begin{pmatrix} Y - \beta_0 - \beta_1 X - \beta_2 f(X,\tau) \\ X(Y - \beta_0 - \beta_1 X - \beta_2 f(X,\tau)) \\ f(X,\tau)(Y - \beta_0 - \beta_1 X - \beta_2 f(X,\tau)) \\ -\beta_2 \mathbf{1}(X > \tau)(Y - \beta_0 - \beta_1 X - \beta_2 f(X,\tau)) \end{pmatrix}.
$$

Now, uniformly with respect to θ, we have,

$$\begin{aligned} P(g_n^{(1)}(\theta) - g^{(1)}(\theta))^2 &= \beta_2^2 \int (f(x,\tau) - q_n(x,\tau))^2 dH(x) \\ &\leq c\beta_2^2 \int_{\tau-\gamma_n}^{\tau+\gamma_n} \gamma_n^2 dH(x) \\ &= c\beta_2^2 \gamma_n^3 O(1) = o(1); \end{aligned}$$

$$\begin{aligned} P(g_n^{(2)}(\theta) - g^{(2)}(\theta))^2 &= \beta_2^2 \int (f(x,\tau) - q_n(x,\tau))^2 X^2 dH(x) \\ &\leq c\beta_2^2 \int_{\tau-\gamma_n}^{\tau+\gamma_n} \gamma_n^2 X^2 dH(x) = o(1); \end{aligned}$$

$$\begin{aligned} P(g_n^{(3)}(\theta) - g^{(3)}(\theta))^2 &= \beta_2^2 \int (f^2(x,\tau) - q_n^2(x,\tau))^2 dH(x) \\ &= c_1\beta_2^2 \int (f(x,\tau) - q_n(x,\tau))^2 dH(x) \\ &\leq c_2\beta_2^2 \int_{\tau-\gamma_n}^{\tau+\gamma_n} \gamma_n^2 dH(x) = o(1); \end{aligned}$$

$$\begin{aligned} P(g_n^{(4)}(\theta) - g^{(4)}(\theta))^2 &= \beta_2^2 \int \left(-f(x,\tau)\mathbf{1}(X > \tau) - q_n(x,\tau)\frac{\partial}{\partial \tau} q_n(X,\tau) \right)^2 dH(x) \\ &\leq c_2\beta_2^2 \gamma_n^3 O(1) = o(1). \end{aligned}$$

Hence $P(g_n^{(i)}(\hat{\theta}_n) - g^{(i)}(\hat{\theta}_n))^2 = o_p(1)$, $i = 1, \ldots, 4$. Then, by the asymptotic equicontinuity property, we have $\mathbb{G}_n(g_n(\hat{\theta}_n) - g(\hat{\theta}_n)) = o_p(1)$ (*van der Vaart and Wellner*, 1996). It can also be shown that $P(g^{(i)}(\hat{\theta}_n) - g^{(i)}(\theta^0))^2 = o_p(1), i = 1, \ldots, 4$. Hence, again by the asymptotic equicontinuity property, we have $\mathbb{G}_n(g(\hat{\theta}_n) - g(\theta^0)) = o_p(1)$ (*van der Vaart and Wellner*,

1996). Now,

$$\begin{aligned} n^{1/2}[\mathbb{U}_n(\hat{\theta}_n) - U_n(\hat{\theta}_n)] &= -2\mathbb{G}_n(g_n(\hat{\theta}_n)) \\ &= -2\mathbb{G}_n(g_n(\hat{\theta}_n) - g(\hat{\theta}_n)) - 2\mathbb{G}_n(g(\hat{\theta}_n) - g(\theta^0)) - 2\mathbb{G}_n(g(\theta^0)) \\ &= -2\mathbb{G}_n(g(\theta^0)) + o_p(1). \end{aligned}$$

Thus by the central limit theorem, the above expression converges in distribution to a normal random variable with mean zero and variance matrix

$$V(\theta^0) = 4 \begin{pmatrix} P(g^{(1)})^2 & P(g^{(1)}g^{(2)}) & P(g^{(1)}g^{(3)}) & P(g^{(1)}g^{(4)}) \\ P(g^{(1)}g^{(2)}) & P(g^{(2)})^2 & P(g^{(2)}g^{(3)}) & P(g^{(2)}g^{(4)}) \\ P(g^{(1)}g^{(3)}) & P(g^{(2)}g^{(3)}) & P(g^{(3)})^2 & P(g^{(3)}g^{(4)}) \\ P(g^{(1)}g^{(4)}) & P(g^{(2)}g^{(4)}) & P(g^{(3)}g^{(4)}) & P(g^{(4)})^2 \end{pmatrix} = 2\sigma^2 \dot{U}_*(\theta^0).$$

By the same argument, as for A_n, we can show that B_n converges to $\dot{U}_*(\theta^0)$ in probability. Thus $n^{1/2}(\hat{\theta}_n - \theta_n)$ converges to $N(0, 2\sigma^2 \dot{U}_*^{-1}(\theta^0))$ in distribution.

A.3 Proof of Corollary 2.3

We now show that the asymptotic distribution of $\hat{\theta}_n$ is same as that shown by *Feder* (1975a) for the exact least square estimates. Remember our broken-stick model with 1 change-point for $Z = 0$ is

$$E(Y|X) = \beta_0 + \beta_1 X + \beta_2 (X - \tau)^+. \tag{A.3}$$

Example 1 discussed by *Feder* (1975a) is a similar model as (2.1). His model is as follows:

$$E(Y|X) = \begin{cases} \psi_{11} + \psi_{12}X, & \text{if } 0 \leq x < \tau \\ \psi_{21} + \psi_{22}X, & \text{if } \tau \leq x \leq M \end{cases} \tag{A.4}$$

where $\tau = (\psi_{11} - \psi_{21})/(\psi_{22} - \psi_{12})$ to ensure continuity. Denote $\psi = (\psi_{11}, \psi_{12}, \psi_{21}, \psi_{22})^{\mathrm{T}}$ and ψ^0 is the true parameter value while as usual τ^0 is the true value of τ.

Quite clearly, the parameter τ is same for both models. *Feder* (1975a) showed that $n^{1/2}(\hat{\psi_{\mathbf{n}}} - \psi^{\mathbf{0}})$ converges in distribution to $N(0, G^{-1})$, where

$$\begin{aligned} \sigma^2 G &= \begin{pmatrix} P(\mathbf{1}(X \leq \tau^0)) & P(X\mathbf{1}(X \leq \tau^0)) & 0 & 0 \\ P(X\mathbf{1}(X \leq \tau^0)) & P(X^2\mathbf{1}(X \leq \tau^0)) & 0 & 0 \\ 0 & 0 & P(\mathbf{1}(X > \tau^0)) & P(X\mathbf{1}(X > \tau^0)) \\ 0 & 0 & P(X\mathbf{1}(X > \tau^0)) & P(X^2\mathbf{1}(X > \tau^0)) \end{pmatrix} \\ &= \begin{pmatrix} G_{11} & 0_2 \\ 0_2 & G_{22} \end{pmatrix}. \end{aligned}$$

Now, $\beta_0 = \psi_{11}$, $\beta_1 = \psi_{12}$, $\beta_2 = \psi_{22} - \psi_{12} \Rightarrow \tau = (\psi_{11} - \psi_{21})/\beta_2$. Observe that,

$$(1, \tau^0, -1, -\tau^0)(\hat{\psi_n} - \psi^0) = \hat{\beta}_{2n}(\hat{\tau}_n - \tau^0).$$

So as in *Feder* (1975a), we have

$$(\hat{\tau}_n - \tau^0) = \frac{1}{\hat{\beta}_{2n}}(1, \tau^0, -1, -\tau^0)(\hat{\psi_n} - \psi^0).$$

Thus,

$$n^{1/2}(\hat{\tau}_n - \tau^0) = \frac{1}{\beta_2^0}(1, \tau^0, -1, -\tau^0)n^{1/2}(\hat{\psi_n} - \psi^0) + o_p(1),$$

which yields

$$n^{1/2}(\hat{\theta}_n - \theta^0) = \begin{pmatrix} I_2 & 0_2 \\ A(\theta^0) & -A(\theta^0) \end{pmatrix} n^{1/2}(\hat{\psi}_n - \psi^0) + o_p(1),$$

where

$$A \equiv A(\theta^0) = \begin{pmatrix} 0 & -1 \\ \frac{1}{\beta_2^0} & \frac{\tau^0}{\beta_2^0} \end{pmatrix}.$$

Letting $D = \begin{pmatrix} I_2 & 0_2 \\ I_2 & -A^{-1} \end{pmatrix}$, we have $n^{1/2}(\hat{\theta}_n - \theta^0) = D^{-1}n^{1/2}(\hat{\psi}_n - \psi^0)$. So according to Feder's work, the asymptotic variance of $n^{1/2}(\hat{\theta}_n - \theta^0)$ is $(D^{\mathrm{T}}GD)^{-1}$. Now $A^{-1}(\theta^0) = \begin{pmatrix} \tau^0 & \beta_2^0 \\ -1 & 0 \end{pmatrix}$, we have $D^{\mathrm{T}}GD = \begin{pmatrix} G_{11} + G_{22} & -G_{22}A^{-1} \\ (-G_{22}A^{-1})^{\mathrm{T}} & (A^{-1})^{\mathrm{T}}G_{22}A^{-1} \end{pmatrix} = (1/2\sigma^2)\dot{U}_*(\theta^0)$. Thus by Feder's calculation, the asymptotic variance of $n^{1/2}(\hat{\theta}_n - \theta^0)$ is $2\sigma^2\dot{U}_*^{-1}(\theta^0))$, which proves Corollary 2.3.

A.4 Proof of Theorem 2.5

The three main steps of this proof has been already outlined in Section 2.3.2. The proof of Theorem 2.5 relies on the following lemma:

Lemma A.1. *Under Conditions 2.1-2.3 in Section 2.3.2, and for $D \equiv m$, for any positive definite matrix $V_{m\times m}$ and $\gamma_n = n^{-\alpha}$ with $\alpha > 1/2$, $n^{1/2}(\hat{\theta}_n^{(V)} - \theta^0)$ converges to $N(0, K^{(V)})$ in distribution, where*

$$K^{(V)} = 2P[(H^T(\theta^0)VH(\theta^0))^{-1}H^T(\theta^0)V^T\Sigma_0VH(\theta^0)(H^T(\theta^0)VH(\theta^0))^{-1}].$$

Here, $\hat{\theta}_n^{(V)} := \arg\min_{\theta\in\Theta} \mathbb{P}_n[(Y-\mu_n)^TV(Y-\mu_n)]$, which is shown to be a zero of $\mathbb{U}_n^{(V)}(\theta) = \frac{\partial}{\partial\theta}\mathbb{P}_n(Y-\mu_n)^TV(Y-\mu_n)$, with probability converging to 1.

The proof of lemma A.1 is similar to the proof of Theorem 2.2. The details of the proof have been provided in the following section.

Now to prove Theorem 2.5, using $V = I$, we obtain from Lemma A.1 that $n^{1/2}(\hat{\theta}_n^{(I)} - \theta^0)$ converges to $N(0, K^{(I)})$ in distribution, for $\gamma_n = n^{-\alpha}$, with $\alpha > 1/2$.

Now because of Condition 2.4 $\sqrt{n}(\hat{\Sigma}_n^{-1} - W^{-1}) = O_p(1)$ implying,

$$\begin{aligned}
& \sqrt{n}(\mathbb{U}_n(\hat{\theta}_n^{(W^{-1})}) - \mathbb{U}_n^{(W^{-1})}(\hat{\theta}_n^{(W^{-1})})) \\
= & \frac{\partial}{\partial\theta}\mathbb{P}_n\left[(Y-\mu_n)^T\{\sqrt{n}(\hat{\Sigma}_n^{-1} - W^{-1})\}(Y-\mu_n)\right]\Big|_{\theta=\hat{\theta}_n^{(W^{-1})}} \\
= & O_p(1)\frac{\partial}{\partial\theta}\mathbb{P}_n\left[(Y-\mu_n)^T(Y-\mu_n)\right]\Big|_{\theta=\hat{\theta}_n^{(W^{-1})}} \\
= & O_p(1)\mathbb{U}_n^{(I)}(\hat{\theta}_n^{(W^{-1})}).
\end{aligned}$$

Again $||\mathbb{U}_n^{(I)} - U^{(I)}|| = o_p(1)$ (proof in A.5). This implies that, $\mathbb{U}_n^{(I)}(\hat{\theta}_n^{(W^{-1})}) = U^{(I)}(\hat{\theta}_n^{(W^{-1})}) + o_p(1)$. Also $\hat{\theta}_n^{(W^{-1})}$ converges in probability to θ^0 and $U^{(I)}$ is continuous, which together imply that $U^{(I)}(\hat{\theta}_n^{(W^{-1})}) = U^{(I)}(\theta^0) + o_p(1) = o_p(1)$, since $U^{(I)}(\theta^0) = 0$. Thus, we obtain $\mathbb{U}_n^{(I)}(\hat{\theta}_n^{(W^{-1})}) = o_p(1)$, which implies that

$$\sqrt{n}(\mathbb{U}_n(\hat{\theta}_n^{(W^{-1})}) - \mathbb{U}_n^{(W^{-1})}(\hat{\theta}_n^{(W^{-1})})) = o_p(1).$$

Also, with probability increasing to 1, $\mathbb{U}_n(\hat{\theta}_n) = \mathbb{U}_n^{(W^{-1})}(\hat{\theta}_n^{(W^{-1})}) = 0$, which implies that the following holds with probability increasing to 1:

$$\sqrt{n}(\mathbb{U}_n(\hat{\theta}_n) - \mathbb{U}_n(\hat{\theta}_n^{(W^{-1})})) = -\sqrt{n}(\mathbb{U}_n(\hat{\theta}_n^{(W^{-1})}) - \mathbb{U}_n^{(W^{-1})}(\hat{\theta}_n^{(W^{-1})})) = o_p(1).$$

Taylor series expansion of $\mathbb{U}_n(\hat{\theta}_n)$ around $\hat{\theta}_n^{(W^{-1})}$ provides

$$\sqrt{n}(\mathbb{U}_n(\hat{\theta}_n) - \mathbb{U}_n(\hat{\theta}_n^{(W^{-1})})) \;=\; \dot{\mathbb{U}}_n(\tilde{\theta_n}^*)\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^{(W^{-1})}),$$

for some $\tilde{\theta_n}^*$ lying on the straight line joining $\hat{\theta}_n$ and $\hat{\theta}_n^{(W^{-1})}$. As shown for A_n in the proof of Theorem 2, we can show that $\dot{\mathbb{U}}_n(\tilde{\theta_n}^*)$ converges in probability to $\dot{U}_*(\theta^0)$, which in turn implies that $\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^{(W^{-1})}) = o_p(1)$.

Also, notice that, clearly with $V = W^{-1}$, from Lemma A.1, $n^{1/2}(\hat{\theta}_n^{(W^{-1})} - \theta^0)$ converges to $N(0, K^{(W^{-1})})$ in distribution. Along with the fact that $\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^{(W^{-1})}) = o_p(1)$, we have proved Theorem 2.5.

A.5 Proof of Lemma A.1

For some positive definite $m \times m$ matrix V, define

$$U^{(V)}(\theta) = \frac{\partial}{\partial\theta} P(M^{(V)}(\theta)) = \frac{\partial}{\partial\theta} P[(Y-\mu)^T V (Y-\mu)],$$

$$U_n^{(V)}(\theta) = \frac{\partial}{\partial\theta} P(M_n^{(V)}(\theta)) = \frac{\partial}{\partial\theta} P[(Y-\mu_n)^T V (Y-\mu_n)],$$

$$\mathbb{U}_n^{(V)}(\theta) = \frac{\partial}{\partial\theta} \mathbb{P}_n(M_n^{(V)}(\theta)) = \frac{\partial}{\partial\theta} \mathbb{P}_n[(Y-\mu_n)^T V (Y-\mu_n)].$$

Define $\theta_n^{(V)}$ and $\hat{\theta}_n^{(V)}$ as minimizers of $P(M_n^{(V)}(\theta))$ and $\mathbb{P}_n(M_n^{(V)}(\theta))$ respectively in Θ, which can be shown to be zeros of $U_n^{(V)}(\theta)$ and $\mathbb{U}_n^{(V)}(\theta)$ respectively, with probability increasing to 1. Now,

$$\mathbb{U}_n^{(V)}(\theta) - U_n^{(V)}(\theta) = (\mathbb{P}_n - P)\left[\frac{\partial}{\partial\theta}(Y-\mu_n)^T V (Y-\mu_n)\right].$$

So as in the proof of Theorem 2.1, the expression in parentheses above belongs to Glivenko-Cantelli class and hence $||\mathbb{U}_n^{(V)} - U_n^{(V)}|| = o_p(1)$.

Observe,

$$\begin{aligned} U_n^{(V)}(\theta) &= \frac{\partial}{\partial\theta}[P(Y-\mu_n)^T V(Y-\mu_n)] \\ &= P[2\{\frac{\partial}{\partial\theta}(Y-\mu_n)^T\}V(Y-\mu_n)] \\ &= -2P[\{\frac{\partial{\mu_n}^T}{\partial\theta}\}V(Y-\mu_n)] \\ &= -2P[D_n(\theta)^T V(Y-\mu_n)], \end{aligned}$$

where

$$D_n(\theta) = \begin{pmatrix} 1^T \\ X^T \\ q_n(X,\tau)^T \\ \beta_2 \frac{\partial}{\partial\tau} q_n(X,\tau)^T \end{pmatrix}_{4\times m}.$$

Hence,

$$U_n^{(V)}(\theta) = -2P \begin{pmatrix} (Y-\mu_n)^T V 1 \\ (Y-\mu_n)^T V X \\ (Y-\mu_n)^T V q_n(X,\tau) \\ \beta_2 (Y-\mu_n)^T V \frac{\partial}{\partial\tau} q_n(X,\tau) \end{pmatrix}^T_{1\times 4}.$$

Similarly, $U^{(V)}(\theta) = -2P[D(\theta)^T V(Y-\mu)]$, where,

$$D(\theta)^T = \begin{pmatrix} 1^T \\ X^T \\ f(X,\tau)^T \\ -\beta_2 \mathbf{1}(X>\tau)^T \end{pmatrix}_{4\times m},$$

and so,

$$U^{(V)}(\theta) \;=\; -2P \begin{pmatrix} (Y-\mu)^T V 1 \\ (Y-\mu)^T V X \\ (Y-\mu)^T V f(X,\tau) \\ -\beta_2 (Y-\mu)^T V \mathbf{1}(X>\tau) \end{pmatrix}^T_{1\times 4}.$$

Clearly, as in proof of Theorem 2.1, $||U_n^{(V)} - U^{(V)}|| = O(\gamma_n)$.
Hence, $||\mathbb{U}_n^{(V)} - U^{(V)}|| \leq ||\mathbb{U}_n^{(V)} - U_n^{(V)}|| + ||U_n^{(V)} - U^{(V)}|| = o_p(1)$, which yields $||\hat{\theta}_n^{(V)} - \theta^0|| = o_p(1)$.

Now Taylor series expansion of $U_n^{(V)}(\theta_n^{(V)})$ around θ^0 yields,

$$U_n^{(V)}(\theta_n^{(V)}) - U_n^{(V)}(\theta^0) = \begin{pmatrix} \dot{U}_{1n}^{(V)}(\tilde{\theta}_n^{(1)}) \\ \dot{U}_{2n}^{(V)}(\tilde{\theta}_n^{(2)}) \\ \dot{U}_{3n}^{(V)}(\tilde{\theta}_n^{(3)}) \\ \dot{U}_{4n}^{(V)}(\tilde{\theta}_n^{(4)}) \end{pmatrix} (\theta_n^{(V)} - \theta^0) = A_n(\theta_n^{(V)} - \theta^0),$$

where each of $\tilde{\theta}_n^{(i)}$, $i = 1, 2, 3, 4$, lies on the straight line joining $\theta_n^{(V)}$ and θ^0 and $\dot{U}_{in}^{(V)} = \frac{\partial U_n^{(V)}}{\partial \beta_{i-1}}$, $i = 1, 2, 3$, and $\dot{U}_{4n} = \frac{\partial U_n^{(V)}}{\partial \tau}$. Now, $U^{(V)}(\theta^0) = U_n^{(V)}(\theta_n^{(V)}) = 0$ for sufficiently large n. Hence, the following is true for sufficiently large n:

$$U_n^{(V)}(\theta^0) - U^{(V)}(\theta^0) = -A_n(\theta_n^{(V)} - \theta^0).$$

Now we show A_n is invertible for large enough n. Observe,

$$\dot{U}_n^{(V)}(\theta) = 2 \begin{pmatrix} 1^T V 1 & P(1^T V X) & P(1^T V q_n(X,\tau)) & \beta_2 P(1^T V \frac{\partial}{\partial \tau} q_n(X,\tau)) \\ P(1^T V X) & P(X^T V X) & P(X^T V q_n(X,\tau)) & \beta_2 P(X^T V \frac{\partial}{\partial \tau} q_n(X,\tau)) \\ P(1^T V q_n(X,\tau)) & P(X^T V q_n(X,\tau)) & P(q_n^T(X,\tau) V q_n(X,\tau)) & a_n^V(\theta) \\ \beta_2 P(1^T V \frac{\partial}{\partial \tau} q_n(X,\tau)) & \beta_2 P(X^T V \frac{\partial}{\partial \tau} q_n(X,\tau)) & a_n^V(\theta) & b_n^V(\theta) \end{pmatrix}$$

where,

$$\begin{aligned} a_n^V(\theta) =& \beta_2 P\left\{q_n^T(X,\tau) V \tfrac{\partial}{\partial \tau} q_n(X,\tau)\right\} - P\left\{(Y - \beta_0 - \beta_1 X - \beta_2 q_n(X,\tau))^T V \tfrac{\partial}{\partial \tau} q_n(X,\tau)\right\}, \\ b_n^V(\theta) =& \beta_2^2 P\left\{\left(\tfrac{\partial}{\partial \tau} q_n(X,\tau)\right)^T V \left(\tfrac{\partial}{\partial \tau} q_n(X,\tau)\right)\right\} \\ & - \beta_2 P\left\{(Y - \beta_0 - \beta_1 X - \beta_2 q_n(X,\tau))^T V \tfrac{\partial^2}{\partial \tau^2} q_n(X,\tau)\right\}. \end{aligned}$$

As in proof of Theorem 2.2, it can be shown that all the elements of $\dot{U}_n^{(V)}(\theta)$ converge uniformly to a finite limit and hence, $\dot{U}_n^{(V)}(\theta)$ converges uniformly to a finite limit, implying A_n converges too. Let $\dot{U}^{(V)}(\theta^0)$ denote the limit of A_n, where $\dot{U}^{(V)}(\theta) = 2P(H^T(\theta^0) V H(\theta^0))$,

$$H(\theta) = \begin{pmatrix} 1 & X & f(X,\tau) & -\beta_2 \mathbf{1}(X > \tau) \end{pmatrix}_{m \times 4}$$

As before, it can be easily seen that $\dot{U}^{(V)}(\theta^0)$ is positive definite. Thus A_n is nonsingular for large enough n, and we have for sufficiently large n,

$$|\theta_n^{(V)} - \theta^0| = {A_n}^{-1} \|U_n^{(V)} - U^{(V)}| = O(\gamma_n).$$

Denote $\mathbb{G}_n = n^{1/2}(\mathbb{P}_n - P)$. Again, observe that Taylor Series expansion of $U_n^{(V)}(\theta_n^{(V)})$

around $\hat{\theta}_n^{(V)}$ yields

$$U_n^{(V)}(\hat{\theta}_n^{(V)}) - U_n^{(V)}(\theta_n^{(V)}) = \begin{pmatrix} \dot{U}_{1n}^{(V)}(\theta_n^{*(1)}) \\ \dot{U}_{2n}^{(V)}(\theta_n^{*(2)}) \\ \dot{U}_{3n}^{(V)}(\theta_n^{*(3)}) \\ \dot{U}_{4n}^{(V)}(\theta_n^{*(4)}) \end{pmatrix} (\hat{\theta}_n^{(V)} - \theta_n^{(V)}) = B_n(\hat{\theta}_n^{(V)} - \theta_n^{(V)}),$$

where $\theta_n^{*(i)}$, $i = 1, 2, 3, 4$ lie on the straight line joining $\theta_n^{(V)}$ and $\hat{\theta}_n^{(V)}$. Now, with probability increasing to 1, $\mathbb{U}_n^{(V)}(\hat{\theta}_n^{(V)}) = U_n^{(V)}(\theta_n^{(V)}) = 0$. Thus, the following equality is true with probability increasing to 1:

$$\mathbb{U}_n^{(V)}(\hat{\theta}_n^{(V)}) - U_n^{(V)}(\hat{\theta}_n^{(V)}) = B_n(\hat{\theta}_n^{(V)} - \theta_n^{(V)}).$$

Now,

$$\begin{aligned} n^{1/2}[\mathbb{U}_n^{(V)}(\hat{\theta}_n^{(V)}) - U_n^{(V)}(\hat{\theta}_n^{(V)})] &= -2\mathbb{G}_n \begin{pmatrix} (Y-\mu_n)^T V 1 \\ (Y-\mu_n)^T V X \\ (Y-\mu_n)^T V q_n(X,\tau) \\ \beta_2 (Y-\mu_n)^T V \frac{\partial}{\partial \tau} q_n(X,\tau) \end{pmatrix}_{\theta=\hat{\theta}_n^{(V)}} \\ &= -2\mathbb{G}_n \begin{pmatrix} g_n^{(1)}(\hat{\theta}_n^{(V)}) \\ g_n^{(2)}(\hat{\theta}_n^{(V)}) \\ g_n^{(3)}(\hat{\theta}_n^{(V)}) \\ g_n^{(4)}(\hat{\theta}_n^{(V)}) \end{pmatrix} \\ &= -2\mathbb{G}_n(g_n(\hat{\theta}_n^{(V)})). \end{aligned}$$

Clearly, each of the $g_n^{(i)}$, $i = 1, \ldots, 4$, is addition and/or multiplication of bounded monotone functions and hence belongs to a Donsker class (*van der Vaart and Wellner*, 1996).

The same can be said about the following vector of functions

$$g = \begin{pmatrix} g^{(1)} \\ g^{(2)} \\ g^{(3)} \\ g^{(4)} \end{pmatrix} = \begin{pmatrix} (Y-\mu)^T V 1 \\ (Y-\mu)^T V X \\ (Y-\mu)^T V f(X,\tau) \\ -\beta_2 (Y-\mu)^T V \mathbf{1}(X > \tau) \end{pmatrix}.$$

As in the proof of Theorem 2.2, $P(g_n^{(i)}(\hat{\theta}_n^{(V)}) - g^{(i)}(\hat{\theta}_n^{(V)})^2 = o_p(1)$, $i = 1, \ldots, 4$. Then, by the asymptotic equicontinuity property, we have $\mathbb{G}_n(g_n(\hat{\theta}_n^{(V)}) - g(\hat{\theta}_n^{(V)})) = o_p(1)$ (*van der Vaart and Wellner*, 1996). It can also be shown that $P(g^{(i)}(\hat{\theta}_n^{(V)}) - g^{(i)}(\theta^0))^2 = o_p(1), i = 1, \ldots, 4$. Hence, again by the asymptotic equicontinuity property, we have $\mathbb{G}_n(g(\hat{\theta}_n^{(V)}) - g(\theta^0)) = o_p(1)$ (*van der Vaart and Wellner*, 1996). Now,

$$\begin{aligned} n^{1/2}[\mathbb{U}_n^{(V)}(\hat{\theta}_n^{(V)}) - U_n^{(V)}(\hat{\theta}_n^{(V)})] &= -2\mathbb{G}_n(g_n(\hat{\theta}_n^{(V)})) \\ &= -2\mathbb{G}_n(g_n(\hat{\theta}_n^{(V)}) - g(\hat{\theta}_n^{(V)})) - 2\mathbb{G}_n(g(\hat{\theta}_n^{(V)}) - g(\theta^0)) \\ &\quad -2\mathbb{G}_n(g(\theta^0)) \\ &= -2\mathbb{G}_n(g(\theta^0)) + o_p(1). \end{aligned}$$

Thus by the central limit theorem, the above expression converges in distribution to a normal random variable with mean zero and covariance matrix

$$4 \begin{pmatrix} P(g^{(1)})^2 & P(g^{(1)}g^{(2)}) & P(g^{(1)}g^{(3)}) & P(g^{(1)}g^{(4)}) \\ P(g^{(1)}g^{(2)}) & P(g^{(2)})^2 & P(g^{(2)}g^{(3)}) & P(g^{(2)}g^{(4)}) \\ P(g^{(1)}g^{(3)}) & P(g^{(2)}g^{(3)} & P(g^{(3)})^2 & P(g^{(3)}g^{(4)}) \\ P(g^{(1)}g^{(4)}) & P(g^{(2)}g^{(4)}) & P(g^{(3)}g^{(4)}) & P(g^{(4)})^2 \end{pmatrix} = 2P(H^T(\theta^0)V^T\Sigma_0 V H(\theta^0)).$$

As shown earlier for A_n, we can prove B_n converges to $\dot{U}^{(V)}(\theta^0)$ in probability. Thus

$n^{1/2}(\hat{\theta}_n^{(V)} - \theta_n^{(V)})$ converges to $N(0, K^{(V)})$ in distribution, where

$$K^{(V)} = 2P[(H^T(\theta^0)VH(\theta^0))^{-1}H^T(\theta^0)V^T\Sigma_0 VH(\theta^0)(H^T(\theta^0)VH(\theta^0))^{-1}].$$

Now, if $\gamma_n = n^{-\alpha}$ with $\alpha > 1/2$, then $n^{1/2}(\theta_n^{(V)} - \theta^0) = o(1)$, implying that $n^{1/2}(\hat{\theta}_n^{(V)} - \theta^0)$ converges to $N(0, K^{(V)})$ in distribution.

A.6 Detailed Figure: Choice of α

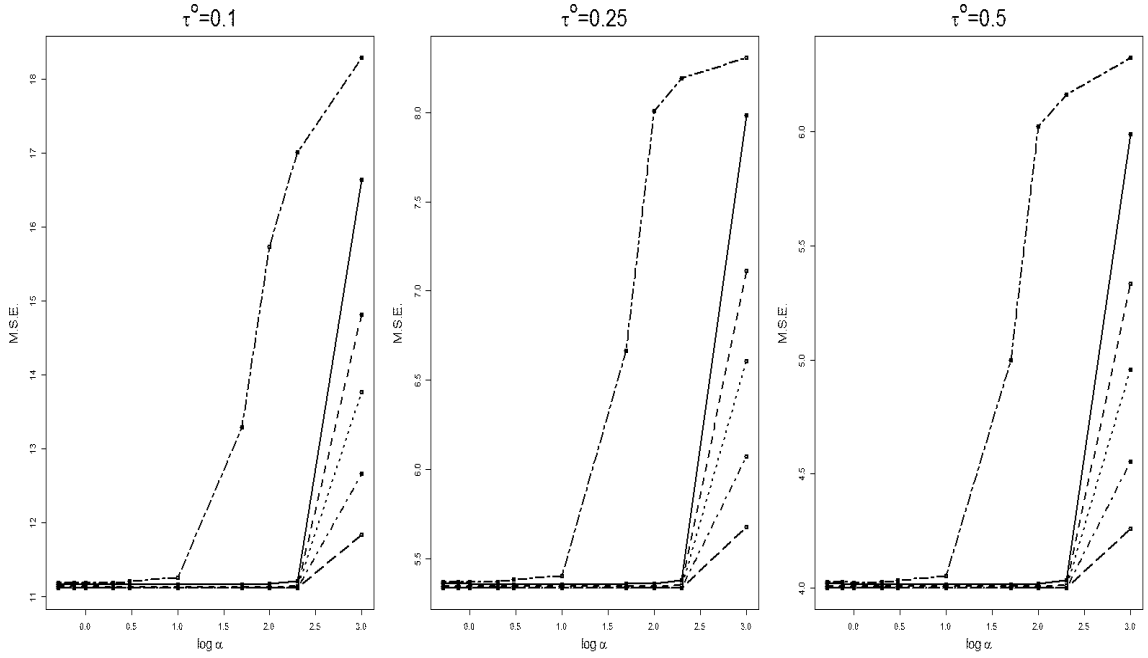

Figure A.1: Mean Square Errors vs $\log_{10}\alpha$ for varying sample-sizes with different τ-values, where $\beta_0^0 = 0.3$, $\beta_1^0 = 1.5$, $\beta_2^0 = 1$ and $\sigma = 0.5$. From the top below,the shortdash-longdash line corresponds to $n = 30$, the solid line corresponds to $n = 50$, dashed line corresponds to $n = 100$, the dotted line corresponds to $n = 500$, the dot-dash line corresponds to $n = 1000$ and the longdash line corresponds to $n = 5000$.

A.7 Verifying Condition 2.4 for model 2.5

Define,

$$\begin{aligned} U^{\eta}(\theta,\eta) &= P\left[\{(Y-\mu)(Y-\mu)^T-\Sigma^{-1}(\eta)\}\frac{\partial\Sigma^{-1}(\eta)}{\partial\eta}\right] \\ U_n^{\eta}(\theta_n^{(I)},\eta) &= P\left[\{(Y-\mu_n^{(I)})(Y-\mu_n^{(I)})^T-\Sigma^{-1}(\eta)\}\frac{\partial\Sigma^{-1}(\eta)}{\partial\eta}\right] \\ \mathbb{U}^{\eta}(\theta,\eta) &= \mathbb{P}_n\left[\{(Y-\mu_n^{(I)})(Y-\mu_n^{(I)})^T-\Sigma^{-1}(\eta)\}\frac{\partial\Sigma^{-1}(\eta)}{\partial\eta}\right] \end{aligned}$$

Now, as before, it easily follows, $||U_n^{\eta}-U^{\eta}||=O(\gamma_n)$. Also, observe that, (θ^0,η^0) is the unique zero of U^{η}. Define (θ_n,η_n) as a zero of U_n^{η}. This implies that $|(\theta_n,\eta_n)-(\theta^0,\eta^0)|=O(\gamma_n)=o(1)$.

Next we show that, $\{(Y-\mu_n)(Y-\mu_n)^T-\Sigma^{-1}(\eta)\}\frac{\partial\Sigma^{-1}(\eta)}{\partial\eta}$ belongs to Glivenko-Cantelli class, which implies that $||\mathbb{U}^{\eta}-U^{\eta}||=o(1)$, implying $|\hat{\eta}_n-\eta_n|=o_p(1)$, hence $|\hat{\eta}_n-\eta^0|=o_p(1)$.

Consider the case when $d=2$ for simplicity.

$$\Sigma(\eta)=\begin{pmatrix} \phi+\sigma^2+\xi(t_{i1}) & \phi+\rho^{|t_{i1}-t_{i2}|}\sqrt{\xi(t_{i1})\xi(t_{i2})} \\ \phi+\rho^{|t_{i1}-t_{i2}|}\sqrt{\xi(t_{i1})\xi(t_{i2})} & \phi+\sigma^2+\xi(t_{i2}) \end{pmatrix}.$$

Clearly this is a non-singular matrix and $\eta\to|\Sigma(\eta)|$ is a continuous map in a compact neighborhood around η^0. Hence, minimum value of $|\Sigma(\eta)|$ is attained, which is strictly greater than zero. Now,

$$\Sigma^{-1}(\eta)=\frac{1}{|\Sigma(\eta)|}\begin{pmatrix} \phi+\sigma^2+\xi(t_{i2}) & -\phi-\rho^{|t_{i1}-t_{i2}|}\sqrt{\xi(t_{i1})\xi(t_{i2})} \\ -\phi-\rho^{|t_{i1}-t_{i2}|}\sqrt{\xi(t_{i1})\xi(t_{i2})} & \phi+\sigma^2+\xi(t_{i1}) \end{pmatrix}.$$

In a compact neighborhood around η^0, since $X\in[M_1,M_2]$, implying the t_{ij}'s are also finite, each element of $\Sigma^{-1}(\eta)$ is bounded. Thus each elements of $\Sigma^{-1}(\eta)$ is addition and/or multiplication of bounded monotone functions and hence belongs to Glivenko-Cantelli as

well as Donsker classes (*van der Vaart and Wellner*, 1996).

Next, $\frac{\partial \Sigma^{-1}(\eta)}{\partial \eta} = -\Sigma^{-1}(\eta)\frac{\partial \Sigma(\eta)}{\partial \eta}\Sigma^{-1}(\eta)$.

Now,

$$\frac{\partial}{\partial \eta}\ \mathrm{Var}(Y_{ij}) = \begin{pmatrix} 1 \\ 1 \\ \xi(t_{ij}) \\ \xi(t_{ij})t_{ij} \\ \xi(t_{ij}){t_{ij}}^2 \\ 0 \end{pmatrix},$$

while,

$$\frac{\partial}{\partial \eta}\ \mathrm{Cov}(Y_{ij}) = \begin{pmatrix} 1 \\ 0 \\ \rho^{|t_{ij}-t_{ik}|}\sqrt{\xi(t_{ij})\xi(t_{ik})} \\ \rho^{|t_{ij}-t_{ik}|}\sqrt{\xi(t_{ij})\xi(t_{ik})}(t_{ij}+t_{ik})/2 \\ \rho^{|t_{ij}-t_{ik}|}\sqrt{\xi(t_{ij})\xi(t_{ik})}({t_{ij}}^2+{t_{ik}}^2)/2 \\ |t_{ij}-t_{ik}|\rho^{|t_{ij}-t_{ik}|-1}\sqrt{\xi(t_{ij})\xi(t_{ik})} \end{pmatrix}.$$

These are all nice functions in compact neighborhood around η^0 and t_{ij}'s are finite since X belongs in $[M_1, M_2]$. Clearly, these functions are measureable and belong to a finite-dimensional vector-space. Hence, they belong to VC class (*van der Vaart and Wellner*, 1996). Also matrix multiplication is just composition of products and sums and hence $\{(Y-\mu_n)(Y-\mu_n)^T-\Sigma^{-1}(\eta)\}\frac{\partial \Sigma^{-1}(\eta)}{\partial \eta}$ indeed belongs to Glivenko-Cantelli and Donsker classes for $d=2$ (*van der Vaart and Wellner*, 1996). For, a general d, this follows by induction.

Observe that, Taylor series expansion of $U_n^{\eta}(\theta_n^{(I)}, \hat{\eta}_n)$ around $(\theta_n^{(I)}, \eta_n)$ yields

$$\begin{aligned} U_n^{\eta}(\theta_n^{(I)}, \hat{\eta}_n) &= U_n^{\eta}(\theta_n^{(I)}, \eta_n) + \frac{\partial}{\partial \eta}U_n^{\eta}(\theta_n^{(I)}, \eta_n^*)(\hat{\eta}_n - \eta_n) \\ &= \mathbb{U}^{\eta}(\theta_n^{(I)}, \hat{\eta}_n) + \frac{\partial}{\partial \eta}U_n^{\eta}(\theta_n^{(I)}, \eta_n^*)(\hat{\eta}_n - \eta_n) \end{aligned}$$

where η_n^* lies on the straight line joining $\hat{\eta}_n$ and η_n. This is true because $U_n^\eta(\theta_n^{(I)}, \eta_n) = \mathbb{U}^\eta(\theta_n^{(I)}, \hat{\eta}_n) = 0$. This implies that

$$\mathbb{G}_n \left[\{(Y - \mu_n^{(I)})(Y - \mu_n^{(I)})^T - \Sigma^{-1}(\eta)\} \frac{\partial \Sigma^{-1}(\eta)}{\partial \eta} \right] = -\frac{\partial}{\partial \eta} U_n^\eta(\theta_n^{(I)}, \eta_n^*) \sqrt{n}(\hat{\eta}_n - \eta_n).$$

Now, $\{(Y - \mu_n)(Y - \mu_n)^T - \Sigma^{-1}(\eta)\} \frac{\partial \Sigma^{-1}(\eta)}{\partial \eta}$ belongs to Donsker class. Thus following arguments similar to those in the proof of Lemma A.1, we get $\sqrt{n}(\hat{\eta}_n - \eta_n) = O_p(1)$. Thus for proper choice of γ_n, we get $\sqrt{n}(\hat{\eta}_n - \eta^0) = O_p(1)$. Now, since $\eta \to \Sigma(\eta)$ is a differentiable function, we get $\sqrt{n}(\Sigma(\hat{\eta}_n) - \Sigma(\eta^0)) = O_p(1)$, i.e., $\sqrt{n}(\hat{\Sigma}_n - \Sigma^0) = O_p(1)$. Hence, condition 2.4 is verified.

A.8 Asymptotics of the Mean Function

Observe that $\mu(x, \theta) \equiv E(Y \mid X = x) = \beta_0 + \beta_1 x + \beta_2 f(x, \tau)$ for the one change-point model. We will work out the calculations with this model for simplicity of presentation, but the idea can be readily extended for the general case.

We are looking at the asymptotics of $\sqrt{n}(\mu(x, \hat{\theta}_n) - \mu(x, \theta^0))$. Clearly the complicated part is $\sqrt{n}[\hat{\beta}_{2,n} f(x, \hat{\tau}_n) - \beta_2^0 f(x, \tau^0)]$. Observe,

$$\begin{aligned}
\sqrt{n}[\hat{\beta}_{2,n} f(x, \hat{\tau}_n) - \beta_2^0 f(x, \tau^0)] &= \sqrt{n}(\hat{\beta}_{2,n} - \beta_2^0)[(x - \hat{\tau}_n)\mathbf{1}(x > \hat{\tau}_n) - (x - \tau^0)\mathbf{1}(x > \tau^0)] \\
&\quad + \sqrt{n}(\hat{\beta}_{2,n} - \beta_2^0)(x - \tau^0)\mathbf{1}(x > \tau^0) \\
&\quad + \beta_2^0 \sqrt{n}[(x - \hat{\tau}_n)\mathbf{1}(x > \hat{\tau}_n) - (x - \tau^0)\mathbf{1}(x > \tau^0)].
\end{aligned}$$

Now,

$$\begin{aligned}
(x - \hat{\tau}_n)\mathbf{1}(x > \hat{\tau}_n) - (x - \tau^0)\mathbf{1}(x > \tau^0) &= -(\hat{\tau}_n - \tau^0)\mathbf{1}(x > \hat{\tau}_n) \\
&\quad + (x - \tau^0)[\mathbf{1}(x > \hat{\tau}_n) - \mathbf{1}(x > \tau^0)].
\end{aligned}$$

Notice that, $(\hat{\tau}_n - \tau^0) = o_p(1)$ implies that $\sqrt{n}[\mathbf{1}(x > \hat{\tau}_n) - \mathbf{1}(x > \tau^0)] = o_p(1)$, which in turn implies that

$$\begin{aligned}\sqrt{n}[(x - \hat{\tau}_n)\mathbf{1}(x > \hat{\tau}_n) - (x - \tau^0)\mathbf{1}(x > \tau^0)] &= -\sqrt{n}(\hat{\tau}_n - \tau^0)\mathbf{1}(x > \hat{\tau}_n) + o_p(1) \\ &= -\sqrt{n}(\hat{\tau}_n - \tau^0)\mathbf{1}(x > \tau^0) + o_p(1).\end{aligned}$$

This means that

$$\begin{aligned}\sqrt{n}[\hat{\beta}_{2,n} f(x, \hat{\tau}_n) - \beta_2^0 f(x, \tau^0)] &= -\sqrt{n}(\hat{\beta}_{2,n} - \beta_2^0)(\hat{\tau}_n - \tau^0)\mathbf{1}(x > \tau^0) \\ &\quad +\sqrt{n}(\hat{\beta}_{2,n} - \beta_2^0)(x - \tau^0)\mathbf{1}(x > \tau^0) \\ &\quad -\beta_2^0\sqrt{n}(\hat{\tau}_n - \tau^0)\mathbf{1}(x > \tau^0) + o_p(1) \\ &= (x - \tau^0)\mathbf{1}(x > \tau^0)\sqrt{n}(\hat{\beta}_{2,n} - \beta_2^0) \\ &\quad -\beta_2^0\mathbf{1}(x > \tau^0)\sqrt{n}(\hat{\tau}_n - \tau^0) + o_p(1).\end{aligned}$$

From the above calculation, we have,

$$\begin{aligned}\sqrt{n}(\mu_n(x, \hat{\theta}_n) - \mu(x, \theta^0)) &= \sqrt{n}(\hat{\beta}_{0,n} - \beta_0^0) + x\sqrt{n}(\hat{\beta}_{1,n} - \beta_1^0) + f(x, \tau^0)\sqrt{n}(\hat{\beta}_{2,n} - \beta_2^0) \\ &\quad -\beta_2^0\mathbf{1}(x > \tau^0)\sqrt{n}(\hat{\tau}_n - \tau^0) + o_p(1) \\ &= a^T\sqrt{n}(\hat{\theta}_n - \theta^0) + o_p(1),\end{aligned}$$

where $a^T = (1, x, f(x, \tau^0), -\beta_2^0\mathbf{1}(x > \tau^0))$, which proves Corollary 2.3.

BIBLIOGRAPHY

BIBLIOGRAPHY

Agarwal, A., S. N. Negahban, and M. J. Wainwright (2012), Fast global convergence of gradient methods for high-dimensional statistical recovery, *Annals of Statistics*, *40*(5), 2452–2482.

Agency), U. E. U. E. P. (2010), Epa positive matrix factorization (pmf) 3.0 model, *Aerosol Science and Technology*.

Asgharian, M., and D. B. Wolfson (2001), Covariates in multipath change-point problems: Modelling and consistency of the mle, *Canadian Journal of Statistics*, *29*(4), 515–528, doi:10.2307/3316005.

Bair, E., T. Hastie, D. Paul, and R. Tibshirani (2006), Prediction by supervised principal components, *J Am Stat Assoc*, *101:18*.

Bellman, R., and R. Roth (1969), Curve fitting by segmented straight lines, *Journal of the American Statistical Asscociation*, *64*, 1079–1084.

Bhattachaya, P. K. (1987), Maximum likelihood estimation of a change-point, *J. Miltivar. Analysis*, *23*, 183–208.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984), Classification and regression treess, *Monterey, CA: Wadsworth and Brooks*.

Brook, R. D., and S. Rajagopalan (2012), Air pollution and type 2 diabetes: mechanistic insights, *Diabetes*, *61*(12), 3037–3045.

Brook, R. D., et al. (2010), Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the american heart association, *Circulation*, *121*(21), 2331–2378.

Brooking, I. R., and P. D. Jameison (2002), Temperature and photoperiod response of vernalization in near-isogenic lines of wheat, *Field Crops Research*, *79*, 21–38.

Brown, J., J. Graham, L. Chen, E. Postlethwait, A. Ghio, W. Foster, and T. Gordon (2007), Panel discussion review: session four-assessing biological plausibility of epidemiological findings in air pollution research, *J Expo Sci Environ Epidemiol*, *17*, S97–S105.

Caner, M., and H. Zhang (2014), Adaptive elastic net for generalized methods of moments, *Journal of Business And Economic Statistics*, *32*.

Chen, C., J. Chan, R. Gerlach, and W. Hsieh (2011), A comparison of estimators for regression models with change points, *Statistics and Computing*, *21*(3), 395–414, doi: 10.1007/s11222-010-9177-0.

Chiu, G., R. Lockhart, and R. Routledge (2002), Bent-cable asymptotics when the bend is missing, *Statistics and Probability Letters*, *59*, 9–16.

Chiu, G., R. Lockhart, and R. Routledge (2006), Bent-cable regression theory and applications, *Journal of the American Statistical Asscociation*, *101*, 542–553.

Dominici, F., R. Peng, C. Barr, and M. Bell (2010), Protecting human health from air pollution: shifting from a single-pollutant to a multipollutant approach, *Epidemiology*, *21*, 187–194, doi:10.1097/EDE.0b013e3181cc86e8.

Fan, J., and J. Lv (2008), Sure independence screening for ultrahigh dimensional feature space, *Journal of the American Statistical Asscociation*, *70*, 849–911.

Fan, J., and J. Lv (2010), A selective overview of variable selection in high dimensional feature space, *Statist. Sinica*, *20*, 101–148.

Fan, J., and H. Peng (2004), Nonconcave penalized likelihood with a diverging number of parameters, *Ann. Statist.*, *32*, 928–961.

Feder, P. I. (1975a), On asymptotic distribution theory in segmented regression problems-identified case, *The Annals of Statistics*, *3*, 49 – 83.

Feder, P. I. (1975b), The log likelihood ratio in segmented regression, *The Annals of Statistics*, *3*, 84 – 97.

Haight, T. J., Y. Wang, M. J. van der Laan, and I. B. Tager (2010), A cross-validation deletion–substitution–addition model selection algorithm: Application to marginal structural models, *Computational Statistics and Data Analysis*, *54*(12), 3080 – 3094, doi: http://dx.doi.org/10.1016/j.csda.2010.02.002.

Hastie, T., R. Tibshirani, and F. J (2001), The elements of statistical learning, *New York: Springer-Verlag*.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999), Bayesian model averaging: a tutorial (with comments by m. clyde, david draper and e. i. george, and a rejoinder by the authors, *Statistical Science*, *14*(4), 382–417, doi:10.1214/ss/1009212519.

Hudson, D. J. (1966), Fitting segmented curves whose join points have to be estimated, *Journal of the American Statistical Asscociation*, *61*, 1097 – 1129.

Hunter, D. R., and R. Li (2005), Variable selection using mm algorithms, *Ann. Statist.*, *33*, 1617–1642.

Huskova, M. (1998), Estimators in the location model with gradual changes, *Commentationes Mathematicae Universitatis Carolinae*, *39*, 147–157.

Johns, D., L. Stanek, K. Walker, S. Benromdhane, B. Hubbell, M. Ross, R. Devlin, D. Costa, and D. Greenbaum (2012), Practical advancement of multipollutant scientific and risk assessment approaches for ambient air pollution, *Environ Health Perspect*, *120*, 1238–1242, doi:10.1289/ehp.1204939.

Joseph, L., and D. Wolfson (1993), Maximum likelihood estimation in the multi-path change-point problem, *Annals of the Institute of Statistical Mathematics*, *45*, 511530.

Kidwell, C., and O. J.M. (2001), Development and evaluation of a prototype system for collecting sub-hourly ambient aerosol for chemical analysis, *Aerosol Science and Technology*, *35*(1), 596–601.

Kidwell, C., and O. J.M. (2004), Elemental analysis of sub-hourly ambient aerosol collections, *Aerosol Science and Technology*, *38*, 205–218.

Loh, P.-L., and M. J. Wainwright (2015), Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima, *ArXiv e-prints*, *27*(4), 538–557.

Lu, X., B. Nan, P. Song, and M. F. Sowers (2010), Longitudinal data analysis with event time as a covariate, *Stat Biosci*, *2*, 65 – 80.

Lund, R. B., and J. Reeves (2002), Detection of undocumented changepoints: A revision of the two-phase regression model, *J. Climate*, *15*, 2547 – 2554.

Lund, R. B., H. L. Hurd, P. Bloomfield, and R. L. Smith (1995), Climatological time series with periodic correlation, *J. Climate*, *8*, 2787 – 2809.

Mauderly, J., R. Burnett, M. Castillejos, H. Ozkaynak, J. Samet, D. Stieb, S. Vedal, and R. Wyzga (2010), Is the air pollution health research community prepared to support a multipollutant air quality management framework?, *Inhal Toxicol*, *22*, 1–19.

Meier, L., S. van de Geer, and P. Buhlmann (2008), The group lasso for logistic regression, *J Roy Stat Soc B*, *70*, 53–71, doi:10.1111/j.1467-9868.2007.00627.x.

Meinshausen, N., and P. Bhlmann (2006), High-dimensional graphs and variable selection with the lasso, *Ann. Statist.*, *34*, 1436–1462.

Meinshausen, N., and B. Yu (2009), Lasso-type recovery of sparse representations for high-dimensional data, *Ann. Statist.*, *37*, 246–270.

Min, J., D. Paek, S. Cho, and K. Min (2009), Exposure to environmental carbon monoxide may have a greater negative effect on cardiac autonomic function in people with metabolic syndrome, *Sci Total Environ*, *407(17)*, 4807–4811.

Morishita, M., G. Keeler, A. Kamal, J. Wagner, J. Harkema, and A. Rohr (2011), Identification of ambient $pm_{2.5}$ sources and analysis of pollution episodes in detroit, michigan using highly time-resolved measurements, *Atmospheric Environment*, *45*, 1627–1637.

Muggeo, V. M. R. (2003), Estimating regression models with unknown break-points, *Statistics in Medicine*, *22*(19), 3055–3071, doi:10.1002/sim.1545.

NCEP (2001), Third report of the ncep expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel iii), *NIH Publication No. 01-3670*.

Negahban, S. N., P. Ravikumar, M. J. Wainwright, and B. Yu (2012), A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers, *Statistical Science*.

Paatero, P., and U. Tapper (1994), Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values, *Environmetrics*, *5*, 111–126.

Poirer, D. J. (1973), Piecewise regression using cubic splines, *Journal of the American Statistical Asscociation*, *68*, 515 – 524.

Raftery, A. (1996), Approximate bayes factors and accounting for model uncertainty in generalised linear models, *Biometrika*, *83*(2), 251–266, doi:10.1093/biomet/83.2.251.

Raftery, A., D. Madigan, and J. Hoeting (1997), Bayesian model averaging for linear regression models, *J Am Stat Assoc*, *92*, 179–191.

Roberts, S., and M. Martin (2006), Using supervised principal components analysis to assess multiple pollutant effects, *Environmental Health Perspective*, *114(12)*, 1877–1882.

Robinson, D. E. (1964), Estimates for the points of intersection of two polynomial regressions, *Journal of the American Statistical Asscociation*, *59*, 214 – 224.

Rohr, A., A. Kamal, M. Morishita, B. Mukherjee, G. Keeler, J. Harkema, and J. Wagner (2011), Altered heart rate variability in spontaneously hypertensive rats is associated with specific particulate matter components in detroit, michigan, *Environmental Health Perspectives*, *119*, 474–480.

Rubin, D. (1976), Inference and missing data, *Biometrika*, *63*(3), 581–592.

Shao, J., and X. Deng (2012), Estimation in high-dimensional linear models with deterministic design matrices, *Ann. Statist.*, *40*(2), 812–831.

Sinisi, S., and M. van der Laan (2004), Deletion/substitution/addition algorithm in learning with applications in genomics, *Stat Appl Genet Mol Bio*, *3*.

Sowers, M. F. R., D. McConnell, B. Nan, S. Harlow, and J. F. Randolph Jr. (2008), Estradiol rates of change in relation to the final menstrual period in a population-based cohort of women, *J. Clin. Endocrinal Metab.*, *93(10)*, 3847 – 3852.

Sowers, M. F. R., D. McConnell, B. Nan, S. Harlow, and J. F. Randolph Jr. (2008b), Follicle stimulating hormone and its rate of change in defining menopause transition stages, *J. Clin. Endocrinal Metab.*, *93(10)*, 3958 – 3964.

Sun, Z., Y. Tao, S. Li, K. Ferguson, J. Meeker, S. Park, S. Batterman, and B. Mukherjee (2013), Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons, *Environmental Health*, *12*(1), 85, doi:10.1186/1476-069X-12-85.

Tishler, A., and I. Zang (1981), A new maximum likelihood algorithm for piecewise regression, *Journal of the American Statistical Asscociation*, *76*, 980–987.

van der Vaart, A., and J. Wellner (1996), *Weak Convergence and Empirical Processes: With Applications to Statistics*, New York: Springer.

van der Vaart, A., and J. Wellner (1999), Preservation theorems for glivenko-cantelli and uniform glivenko-cantelli classes, *Technical Report No. 361, Department of Statistics, University of Washington*.

Wagner, J., et al. (2014), Cardiovascular depression in rats exposed to inhaled particulate matter and ozone: effects of diet-induced metabolic syndrome, *Environ Health Perspect.*, *1*, 27–33.

Wang, H. (2009), Forward regression for ultra-high dimensional variable screening, *Journal of the American Statistical Asscociation*, *104*, 1512–1524.

Wang, H., R. Li, and C. Tsai (2007), Tuning parameter selectors for the smoothly clipped absolute deviation method, *Biometrika*, *94*, 553–568.

Wang, Z., H. Liu, and T. Zhang (2014), Optimal computational and statistical rates of convergence for sparse nonconvex learning problems, *Ann. Statist.*, *42*(6), 2164–2201.

Wold, S., M. Sjostrom, and L. Eriksson (2001), Pls-regression: a basic tool of chemometrics, *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 109 – 130, doi: http://dx.doi.org/10.1016/S0169-7439(01)00155-1, {PLS} Methods.

Yuan, M., and Y. Lin (2006), Model selection and estimation in regression with grouped variables, *J Roy Stat Soc B*, *68*, 49–67, doi:10.1111/j.1467-9868.2005.00532.x.

Zhang, C., and J. Huang (2008), The sparsity and bias of the lasso selection in high-dimensional linear regression, *Ann. Statist.*, *36*, 1567–1594.

Zhao, P., and B. Yu (2006), On model selection consistency of lasso, *J. Mach. Learn. Res.*, *7*, 2541–2563.

Zou, H. (2006), The adaptive lasso and its oracle properties, *J Am Stat Assoc*, *101*, 1418–1429, doi:10.1198/016214506000000735.

Zou, H., and T. Hastie (2005), Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society, Series B*, *62*, 301–320.

Zou, H., and H. Zhang (2009), On the adaptive elastic-net with a diverging number of parameters, *The Annals of Statistics*, *37*, 1733–1751.