ASYMPTOTIC INFERENCE FOR SEGMENTED REGRESSION MODELS

By

SHIYING WU

B.Sc., Beijing University, 1983

M.Sc., The University of British Columbia, 1988

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

DEPARTMENT OF STATISTICS

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

October 1992

Department of Statistics

The University of British Columbia
Vancouver, Canada

Date Oct 14 /92

Asymptotic inference for segmented regression models

Abstract

This thesis deals with the estimation of segmented multivariate regression models. A segmented regression model is a regression model which has different analytical forms in different regions of the domain of the independent variables. Without knowing the number of these regions and their boundaries, we first estimate the number of these regions by using a modified Schwarz' criterion. Under fairly general conditions, the estimated number of regions is shown to be weakly consistent. We then estimate the change points or "thresholds" where the boundaries lie and the regression coefficients given the (estimated) number of regions by minimizing the sum of squares of the residuals. It is shown that the estimates of the thresholds converge at the rate of $O_p(\ln^2 n/n)$, if the model is discontinuous at the thresholds, and $O_p(n^{-1/2})$ if the model is continuous. In both cases, the estimated regression coefficients and residual variances are shown to be asymptotically normal. It is worth noting that the condition required of the error distribution is local exponential boundedness which is satisfied by any distribution with zero mean and a moment generating function provided its second derivative is bounded near zero. As an illustration, a segmented bivariate regression model is fitted to real data and the relevance of the asymptotic results is examined through simulation studies.

The identifiability of the segmentation variable is also discussed. Under different conditions, two consistent estimation procedures of the segmentation variable are given.

The results are then generalized to the case where the noises are heteroscedastic and autocorrelated. The noises are modeled as moving averages of an infinite number of independently, identically distributed random variables multiplied by different constants in different regions. It is shown that with a slight modification of our assumptions, the estimated number of regions is still consistent. And the threshold estimates retain the convergence rate of $O_p(\ln^2 n/n)$ when the segmented regression model is discontinuous at the thresholds. The estimation procedures also give consistent estimates of the residual variances for each region. These estimates and the estimates of the regression coefficients are shown to be asymptotically normal. The consistent estimate of the segmentation variable is also given. Simulations are carried out for different model specifications to examine the performance of the procedures for different sample sizes.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I thank my supervisor, Dr. Jian Liu, for his inspiration, guidance, support and advice throughout the course of the work reported in this thesis.

I wish to express my deep gratitude to Professor James V. Zidek, for his guidance, encouragement, patience and valuable advice.

This thesis benefitted from the helpful comments of Professor Piet De Jong to whom I am indebted.

Professor John Petkau and Mr. Feifang Hu also made valuable comments.

Many thanks go to Dr. Harry Joe and Nancy E. Heckman for their encouragement and support during my stay at UBC.

Special thanks to Professor James V. Zidek, who provided boundless support throughout my graduate career at UBC.

The financial support from the Department of Statistics, University of British Columbia is acknowledged with great appreciation. I also acknowledge the support of the University of British Columbia through a University Graduate Fellowship.

Chapter 1

PROLOGUE

1.1 Introduction

This thesis deals with asymptotic estimation for segmented multivariate regression models. A segmented regression model is a regression model which has different analytical forms in different regions of the domain of the independent variables. This model may be useful when a response variable depends on the independent variables through a function whose form cannot be uniformly well approximated by a single finite Taylor expansion, and hence the usual linear regression models are not applicable. In such a situation, the possibility of regaining the simplicity of the Taylor expansion and added modeling flexibility is achieved by allowing the response to depend on these variables differently in different subregions of the domains of certain independent variables. For example, Yeh *et al* (1983) discuss the idea of an "anaerobic threshold". It is hypothesized that if a person's workload exceeds a certain threshold where his muscles cannot get enough oxygen, then the aerobic metabolic processes become anaerobic processes. This threshold is called "anaerobic threshold". In this case a model with two segments is suggested by the subject oriented theory. McGee and Carleton (1970) discuss another example where the dependent structure of the selling volume of a regional stock exchange on that of New York Stock Exchange and American Stock Exchange is thought to be changed by a change of govenment regulation. A model with four segments is considered appropriate in their analysis. Examples of this kind in various contexts are given by Sprent (1961), Dunicz (1969), Schulze (1984) and many others. In some situations, although a segmented model is considered suitable, the appropriate number of segments may not be known, as in the example mentioned above and the exchange rate problem we shall discuss in Chapter 5. Furthermore, in the case of multivariate regression, it may not be clear which of the independent variables relate to the change of the dependent structure, or, which independent variable can be best used as the segmentation variable.

In some problems where the independent variables are of low dimension, graphical approaches may be effective in determining the number of segments and which independent variable can best be chosen as the segmentation variable. However, if the independent variables are of high dimension, the interrelations of the independent variables may thwart such an approach. Therefore, an objective and automated approach is in order.

In this thesis, we develop procedures to estimate the model parameters, including the segmentation variable, the number of segments, the location of the thresholds, and other parameters in the model. Note that the word "threshold" is used to emphasize that the dependent structure changes when the segmentation variable exceeds certain values. The estimation procedures are based on least squares estimation and a modified version of Schwarz' (1978) criterion. These estimators are shown to be consistent under fairly mild conditions. In addition, asymptotic distributions are derived for the estimated regression coefficients and the estimated variance of the noises.

The procedures are then generalized to accommodate situations when the noise levels are different from segment to segment, and when the noise is autocorrelated. It is shown that the consistency of these estimators is retained. Simulated data sets are analyzed by the proposed procedures to show their performances for finite sample sizes, and the results seem satisfactory.

1.2 A review of segmented regression and related problems

One problem closely related to segmented regression is the change-point problem. A segmented regression problem reduces to a change-point problem if the regression functions are unknown constants and the boundaries of the segments are to be estimated. In general, a change-point problem refers to the problem of making inferences about the point in a sequence of random variables at which the law governing evolution of the process changes. As a matter of fact, part of the work in this thesis is greatly inspired by Yao's (1988) work on the change-point problem.

The segmented regression problem and change-point problem have attracted much attention since the 1950's. Shaban (1980) gives a rather complete list of references from the 1950's to 1970's. Among other authors, Quandt (1958) postulates a model of the form:

$$y_t = \beta_0^{(j)} + \beta_1^{(j)} x_t + \epsilon_t, \quad j = \begin{cases} 1, & \text{if } t < t^*, \\ 2, & \text{if } t \geq t^*, \end{cases} \quad t = 1, \cdots, N,$$

where t^* is unknown. Under the assumption that ϵ_t's are independent normal random variables, he obtains the maximum likelihood estimates for the parameters including t^*.

Robison (1964) considers a two-phase polynomial regression problem of the form:

$$y_t = \beta_0^{(j)} + \beta_1^{(j)} x_t + \beta_2^{(j)} x_t^2 + \cdots + \beta_k^{(j)} x_t^k + \epsilon_t, \qquad j = \begin{cases} 1, & \text{if } t < t^*; \\ 2, & \text{if } t \geq t^*. \end{cases}$$

Also assuming noises are independent normal variables, he obtains the maximum likelihood estimate and confidence interval for the change-point.

Adding to the model of Quandt (1958) the assumption that the model is everywhere continuous and the variances of the $\{\epsilon_t\}$ are identical, Hudson (1966) gives a concise method for calculating the overall least squares estimator of the intersection point of two intersecting regression lines. For the same problem, Hinkley (1969) derives an asymptotic distribution for the maximum likelihood estimate of the intersection which is claimed to be a better approximation to the finite sample distribution than the asymptotic normal distribution of Feder and Sylwester (1968).

For the change-point problem, Hinkley (1970) derives the asymptotic distribution of the maximum likelihood estimate of the change-point. He assumes that exactly one change occurs and that the means of the two submodels are known. He also gives the asymptotic distribution when these means are unknown, and the noises are assumed to be identically, independently distributed normal random variables ("iid normal" hereafter). As Hinkley notes, the maximum likelihood estimate is not consistent and the asymptotic result is not good for small samples when the two means are unknown.

In all of these problems, the number of change points is assumed to be exactly one. For problems where the number of change-points may be more than one, Quandt (1958, p880) concludes "The exact number of switches must be assumed to be known".

McGee and Carleton (1970) treat the estimation problem for cases where more than one change may occur. Their model is:

$$y_t = \beta_0^{(j)} + \beta_1^{(j)} x_{1t} + \cdots + \beta_k^{(j)} x_{kt} + \epsilon_t, \quad if \ \ t \in [\tau_{j-1}, \tau_j),$$

where $1 \leq \tau_1 < \cdots < \tau_L < \tau_{L+1} = N$ and the $\{\epsilon_t\}$ are iid $N(0, \sigma^2)$. Note that L and the τ_j's are unknown. Constrained by the computing power available at that time (1970), they propose a estimation method which essentially combines least squares estimation with hierarchical clustering. While being computationally efficient, their method is suboptimal (resulting from the use of hierarchical clustering), subjective (in terms of choice of L) and lacking theoretical justification.

Goldfeld and Quandt (1972, 1973a) discuss the so-called switching regression model specified as follows:

$$y_t = \begin{cases} x_t'\beta_1 + u_{1t}, & \text{if } \pi' z_t \le 0; \\ x_t'\beta_2 + u_{2t}, & \text{if } \pi' z_t > 0. \end{cases}$$

Here $z_t = (z_{1t}, \cdots, z_{kt})'$ are the observations on some exogenous variables (including, possibly, some or all of the regressors), $\pi = (\pi_1, \cdots, \pi_k)'$ is an unknown parameter, and the $\{u_{it}\}$ are independent normal random variables with zero means and variances, σ_i^2, $i = 1, 2$. The parameters, $\beta_1, \beta_2, \sigma_1^2, \sigma_2^2$ and π are to be estimated. They define $d(z_t) = \mathbf{1}_{(\pi' z_t > 0)}$ and reexpress the model as

$$y_t = x_t'[(1 - d(z_t))\beta_1 + d(z_t)\beta_2] + (1 - d(z_t))u_{1t} + d(z_t)u_{2t}.$$

For estimation the "D-method" is proposed: $d(z_t)$ is replaced by

$$\int_{-\infty}^{\pi' z_t} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\{-\frac{v}{2\sigma^2}\} dv$$

and the maximum likelihood estimates for the parameters are obtained. As they point out, the D-method can be extended to the case of more than two regimes.

Gallent and Fuller (1973) consider the problem of estimating the parameters in a piecewise polynomial model with continuous derivative, where the join points are unknown. They reparametrize the model so that the Gauss-Newton method can be applied to obtain the least squares estimates. An F statistic is suggested for model selection (including the number of regimes) without theoretical justification.

Poirer (1973) relates spline models and piecewise regression models. Assuming the change points known, he develops tests to detect structural changes in the model and to decide whether certain of the model coefficients vanish.

Ertel and Fowlkes (1976) also point out that the regression models for linear spline and piecewise linear regression have many common elements. The primary difference between them is that in the linear spline case, adjacent regression lines are required to intersect at the change-points, while in the piecewise linear case, adjacent regression lines are fitted separately. He develops some efficient algorithms to obtain least squares estimates for these models.

Feder (1975a) considers a one-dimensional segmented regression problem; it is assumed that the function is continuous over the entire range of the covariate and the number of segments is known. Under certain additional assumptions, he shows that the least squares estimates of the regression coefficients of the model are asymptotically normally distributed. Note that the two assumptions that the function is continuous and that the number of segments is known are essential for his results.

For the simplest two segments regression problem with continuity assumption, Miao (1988) proposes a hypothesis test procedure for the existence of a change-point together with a confidence interval of the change-point, based on the theory of Gaussian processes.

Statistical hypothesis tests for segmented regression models are studied by many authors, among them are Quandt (1960), Sprent (1961), Hinkley (1969), Feder (1975b) and Worsley (1983). Bayesian methods for the problem are considered by Farley and Hinich (1970), Bacon and Watts (1971), Broemeling (1974), Ferreira (1975), Holbert and Broemeling (1977) and Salazar, Broemeling and Chi (1981). Quandt (1972), Goldfeld and Quandt (1972, 1973b) and Quandt and Ramsey (1978) treat the problem as a random mixture of two regression lines.

Closely related to the problem studied in this thesis, Yao (1988) studies the following change-point problem: a sequence of independent normally distributed random variables have a common variance, but their means change l times along the sequence, with l unknown. He adopts the Schwarz criterion for estimating l and proves that such an estimator is consistent. Yao noted that consistency need not obtain without the normality assumption.

Yao and Au (1989) consider the problem of estimating a step function, $g(t)$, over $t \in [0,1]$ in the presence of additive noise. They assume that $t_i = i/n$ ($i = 1, \cdots, n$) are fixed points and the noise has a sixth or higher moment, and derive limiting distributions for the least squares estimators of the locations and sizes of the jumps when the number of jumps is either known or bounded. The discontinuity of $g(t)$ at each change point makes the estimated locations of the jumps converge rapidly to their true values.

This thesis is primarily about situations like those described above, where the segmented regression model may be viewed as a partial explanation model tries to capture our impression that an abrupt change in the mechanism underlying the process. It is linked to other paradigms in modern regression theory as well. Much of this theory (see the references below, for example) is concerned with regression functions of say, y on x, which cannot be well approximated globally by the leading terms of its Taylor expansion, and hence by a global linear model. This has led to various approaches to "nonparametric regression" (see Friedman, 1991, for a recent survey).

One such approach is that of Cleveland (1979) when the dimension of x is 1; his results, which use a linear model in a moving local window, are extended to higher dimensions by Cleveland and Devlin (1988). Weerahandi and Zidek (1988) use a Taylor expansion explicitly to construct a locally weighted smoother, also when the dimension of x is 1; a different expansion is used at each x-value thereby avoiding the shortcomings of using a single global expansion.

However, difficulties confront local weighting methodologies like those described above as well as kernel smoothers and splines because of the "curse of dimensionality" which becomes progressively more serious as the dimension of x grows beyond 2. These difficulties are well described by Friedman (1991) who presents an alternative methodology called "multivariate adaptive regression splines," or "MARS."

MARS avoids the curse of dimensionality by partitioning x's domain into a data-determined, but moderate number of subdomains within which spline functions of low dimensional subvectors of x are fitted. By using splines of order exceeding 0, MARS can lead to continuous smoothers. In contrast, its forerunner, called "recursive partitioning" by Friedman, must be discontinuous, because a different constant is fitted in different subdomains. But, like MARS it avoids the curse of dimensionality because it depends locally on a small number (in fact, none) of the coordinates of x. Friedman (1991) attributes to Breiman and Meisel (1976), a natural extension of recursive partitioning wherein a linear function of x is fitted within each subdomain. However, it can encounter the curse of dimensionality when these subdomains are small and Friedman (1991) ascribes the lack of popularity of this extension to this feature.

However, the curse of dimensionality is relative. If the subdomains of x are large the "curse" becomes less problematical. And within such subdomains, the Taylor expansion leads to linear models like those used by Breiman and Meisel (1976) and here, as natural approximants; in contrast, splines seem somewhat ad hoc. And linear models have a long history of application in statistics.

1.3 New contributions and their relationship to previous work

In this thesis, we address the problem of making asymptotic inference for the following model:

$$y_t = \beta_{i0} + \sum_{j=1}^{p} \beta_{ij} x_{tj} + \sigma_i \epsilon_t, \ if \ x_{td} \in (\tau_{i-1}, \tau_i], i = 1, \ldots, l+1, \tag{1.1}$$

where $\mathbf{z}_t = (x_{t1}, \ldots, x_{tp})'$ is an observed random variable; ϵ_t is assumed to have zero mean and unit variance, while β_{ij}, τ_i, σ_i $(i = 1, \ldots, l+1,\ j = 0, 1, \ldots, p)$, l and d are unknown parameters. Our main contributions are as follows.

A sequence of procedures are proposed to estimate all these parameters, based on least squares estimation and our modified Schwarz' criterion. It is shown that under mild conditions, the estimator, $\hat{l}$, of l is consistent. Furthermore, a bound on the rate of convergence of $\hat{\tau}_i$ and the asymptotic normality for estimators of β_{ij}, σ_i $(i = l, \ldots, l+1,\ j = 0, 1, \ldots, p)$ are obtained under certain additional assumptions.

When the segmentation is related to a few highly correlated covariates, it may not be clear which covariate can best be chosen as the segmentation variable. In such a case, d will be treated as an unknown parameter to be estimated. A new concept of identifiability of d is introduced to formulate the problem precisely. We prove that the least squares estimate of d is consistent. In addition, we propose another consistent and computationally efficient estimate of d. All of these are achieved without the Gaussian assumption on the noises.

In many practical situations, it is necessary to assume that the noises are heteroscedastic and serially correlated. Our estimation procedures and the asymptotic results are generalized to such situations. Asymptotic theory for stationary processes are developed to establish consistency and asymptotic normality of the estimates.

Note that in Model (1.1) if $\beta_{ij} = 0$ for all $i = 1, \cdots, l+1$ and $j = 1, \cdots, p$, equation (1.1) reduces to the change-point problem discussed by Yao (1988), x_d being the explanatory variable controlling the allocation of measurements associated with different dependence structures. Although our formulation is somewhat different from that of Yao (1988) in that we introduce an explanatory variable to allocate response measurements, both formulations are essentially the same from the point of view of an experimental design. If the other covariates are all known functionals of x_d, as in segmented polynomial regressions, and l is known, (1.1) reduces to the case discussed by Feder (1975a).

Unlike all the above mentioned work on segmented regression except McGee and Carleton (1970), we assume that the number of segments is unknown, and that the noise may be dependent. In terms of estimating l, we generalize Yao's (1988) work on the change-point problem to a multiple segmented regression set-up. Furthermore, his conditions on the noises are relaxed in the sense that the ϵ_t's do not have to be (a) normally distributed (rather, they could follow any of the many distributions commonly used for noise); (b) identically distributed; and (c) independent. In terms of making asymptotic inference on the regression coefficients and the change points, we do not assume continuity of the underlying function which is essential for Feder's (1975a) results. We find that without the continuity assumption, the estimated change points converge to the true ones at a much faster rate than the rate given by Feder. Finally, a consistent estimator is obtained for d, an additional parameter not found in any of the previous work.

Our results also relate to MARS. In fact, our estimation procedure can be viewed as adaptive regression using a different method of partitioning than Breiman and Meisel (1976). By placing an upper bound on the number of partitions, we can avoid the difficulties caused by curse of dimensionality, of fitting our model to data in high dimensional space (but recognize that there are trade–offs involved). And we have adopted a different stopping criterion in partitioning x–space; it is based on ideas of model selection rather than testing and seems more appealing to us. Finally, and most importantly, we are able to provide a large sample theory for our methodology. This feature of our work seems important to us. Although the MARS methodology appears to be supported by the empirical studies of Friedman (1991), there is an inevitable concern about the general merits of any procedure when it lacks a theoretical foundation.

Interestingly enough, it can be shown that in some very special cases, our estimation procedures coincide with those of MARS in estimating the change points, if our stopping criterion were adopted in MARS. This seems to indicate that, with our techniques, MARS may be modified to possess certain optimalities (e.g. consistency) or suboptimalities for more general cases.

So in summary, with the estimation procedures proposed in this thesis we regain some of the simplicity of the (piecewise) Taylor expansion and attendant linear models, while retaining some of the virtues of added modeling flexibility possessed by nonparametric approaches. Our large sample theory gives us precise conditions under which our methodology would work well, given sufficiently large samples. And by restricting the number of x–subdomains sufficiently we avoid the curse of dimensionality. Partitioning for our methodology, is data–based like that of MARS.

1.4 Outline of the following chapters

This dissertation is organized as follows. In Chapter 2, the identifiability of the segmentation variable in the segmented regression model is discussed first. We introduce a concept of *identifiability* and demonstrate how the concept naturally arises from the problem. Then we give an equivalent condition which is crucial in establishing the consistency. Finally, we give a sequence of procedures to estimate all the parameters involved in a "basic" segmented regression model with uncorrelated and homoscedastic noise. These procedures are illustrated with an example.

The consistency of the estimates given in Chapter 2 is proved in Chapter 3. Conditions under which the procedures give consistent estimates are also discussed. For technical reasons, the consistency of estimates other than that of the segmentation variable is established first. The estimation problem is treated as a problem of model selection, with the models represented by the possible number of segments, assuming the segmentation variable is known. Schwarz' criterion is tuned to an order of magnitude that can distinguish systematic bias from random noise and is used to select models. Then, with the established theories, the consistency of the estimated segmentation variable is proved. Simulations with various model specifications are carried out to demonstrate the finite sample behavior of the estimators, which prove to be satisfactory.

Results given in Chapter 2 and Chapter 3 are generalized to the case where the noise levels in different segments are different. The noise often derives from factors that cannot be clearly specified and about which little is known. In many practical situations, like that of the economic example mentioned above, the noise may represent a variety of factors of different magnitudes, over different segments. Therefore a heteroscedastic specification of the noise is often necessary. To meet practical needs further, the noise term in the model is assumed to be autocorrelated. The estimation procedures given in Chapter 2 are modified to accommodate these necessities and presented in Chapter 4. It is shown that under a moving average specification of the noise, the estimates given by the procedures are consistent. Further, the parameters specified in the moving average model of the noise term can be estimated by the estimated residuals. Simulation results are given to shed light on the finite sample behavior of the estimates.

A summary of the results established in this thesis is given in Chapter 5. Future research is also discussed. One line of future research comes from the similarity between segmented regression and spline techniques. Our model can first be generalized to the case where there are more than one segmentation variables. Then an "oblique" threshold model can be considered. An oblique threshold is one made by a linear combination of explanatory variables. This is reasonable because often there is no reason to believe that the threshold has to be parallel to any of the axes. Finally, by partitioning the domain of the explanatory variables into polygons, an adaptive regression splines could be developed. This could serve as an alternative to Friedman's (1988) multivariate adaptive regression spline method, or MARS.

Chapter 2

ESTIMATION OF SEGMENTED REGRESSION MODELS

In this chapter, we consider a special case of model (1.1) where the $\{\sigma_i\}$ are all equal and the $\{\epsilon_t\}$ are independent and identically distributed. In this case, the model can be reformulated as follows. Let $(y_1, x_{11}, \ldots, x_{1p})$, ..., $(y_n, x_{n1}, \ldots, x_{np})$ be the independent observations of the response, y, and the covariates, $x_1, \ldots, x_p$. Let $\mathbf{x}_t = (1, x_{t1}, \ldots, x_{tp})'$ for $t = 1, \ldots, n$ and $\tilde{\beta}_i = (\beta_{i0}, \beta_{i1}, \ldots, \beta_{ip})'$, $i = 1, \ldots, l+1$. Then,

$$y_t = \mathbf{x}_t'\tilde{\beta}_i + \epsilon_t, \ if \ x_{td} \in (\tau_{i-1}, \tau_i], \ i = 1, \ldots, l+1, \ t = 1, \ldots, n, \tag{2.1}$$

where the $\{\epsilon_t\}$ are iid with mean zero and variance σ^2 and are independent of $\{\mathbf{x}_t\}$, $-\infty = \tau_0 < \tau_1 < \cdots < \tau_{l+1} = \infty$. The $\tilde{\beta}_i$, τ_i, $(i = 1, \ldots, l+1)$, l, d and σ^2 are unknown parameters. When $\beta_d = 0$, the segmentation variable x_{td} becomes an exogenous variable as considered by Goldfeld and Quandt (1972, 1973a).

A sequence of estimation procedures is given to estimate the parameters in model (2.1). The estimation is done in three steps. First, the segmentation variable or the parameter d is estimated, if it is not known *a priori*. Then, with d known or supposed known, if estimated, the number of structural changes l and the locations of structural changes τ_i's are estimated by a modified Schwarz' criterion. Finally, based on the estimated d, l and τ_i's, the $\tilde{\beta}_i$'s and σ^2 are estimated by ordinary least squares. It will be shown in the next chapter that all these estimators are consistent, under certain conditions.

It is obvious that to estimate d consistently, it has to be identifiable. In Section 2.1, we discuss the identifiability of d. Specifically, we introduce a concept of *identifiability* and give equivalent conditions, all illustrated by examples. These conditions will be used in the next chapter to provide the consistency of the estimator of d.

Our estimation procedures are given in Section 2.2. In particular, two procedures are given to estimate d under different conditions. The first one assumes less prior knowledge while the second one requires less computational effort. Based on the estimated d, the estimation procedures for other parameters are then given. Finally, all the procedures are illustrated by an example in which the dependence of gas consumption on the weight and horse power of different cars is examined. Some general remarks are made in Section 2.3.

In the sequel, either a superscript or a subscript 0 will be used to denote the true parameter values.

2.1 Identifiability of the segmentation variable

Although in some applications, the parameter d can be determined *a priori* from background knowledge about the problem of concern, it can be hard to determine d with reasonable certainty, due to a lack of background information. For instance, if the segmentation is related to a few highly correlated covariates, it may not be clear which one can best be chosen as the segmentation variable. Therefore, there is a need for a defensible choice of d based on the data. When the vector of covariates are of high dimension and d cannot be identified by graphical methods, a computational procedure is required. However, when some of the covariates are highly correlated, it may not be clear whether d can be uniquely identified. In the following, we discuss the exact meaning of being "identified" and give a set of conditions under which d

can be uniquely identified.

To simplify notation, let $\mathbf{x}$ have the same distribution as that of $\mathbf{x}_1$ and $R_j^0 = \{\mathbf{x} : x_{d^0} \in (\tau_{j-1}^0, \tau_j^0]\}$, $j = 1, \ldots, l^0 + 1$. And for any d, let $\{R_j^d\}_{j=1}^{l+1}$ be a partition of $\mathbf{R}^p$ where $R_j^d = \{\mathbf{x} : x_d \in (\tau_{j-1}, \tau_j]\}$, $-\infty = \tau_0 < \tau_1 < \cdots < \tau_l < \tau_{l+1} = \infty$. Let L be a known upper bound on the number of thresholds. Intuitively speaking, d^0 is identifiable if for any $d \neq d^0$, and any partition $\{R_j^d\}_{j=1}^{L+1}$, there is at least one region, say R_r^d, on which the model exhibits clear nonlinearity.

Note that L is involved. Indeed, the identifiability of d^0 does depend on L when the domain of $\mathbf{x}$ takes a certain special form. This can be easily seen in the following two examples.

Example 1 $\mathbf{x}$ is uniformly distributed over the shaded area in Figure 2.1,

$$Y = \mathbf{1}_{(x_1 > 1)} + \epsilon,$$

where $\mathbf{1}_{(\cdot)}$ is an indicator function. And

$R_1^0 = \{\mathbf{x} : x_1 \in (-\infty, 1]\}$, $R_2^0 = \{\mathbf{x} : x_1 \in (1, \infty)\}$.

For $L = 1$, no threshold on x_2 can make the model piecewise linear over its domain. The only possible threshold which makes the model piecewise linear is $\tau_1 = 1$ as defined in the model. For $L = 2$, however, $\tau_1 = -1$, $\tau_2 = 1$ also make the model piecewise linear over its domain. Hence either x_1 or x_2 can be used as the threshold variable. ¶

The same phenomenon can also be seen in the next example.

Example 2 $\mathbf{x}$ is uniformly distributed with probabilities concentrated at the 8 points as specified in Figure 2.2,

$$Y = \mathbf{1}_{(x_1 > 0)} \cdot x_2 + \epsilon.$$

For $L = 1$, no threshold on x_2 can make the model piecewise linear over its domain. For $L = 2$, however, $\tau_1 = -1/2$, $\tau_2 = 1/2$ make the model piecewise linear over its domain. Hence either x_1 or x_2 can be used as the threshold variable. ¶

Sometimes, but not always, one cannot determine whether or not the model is linear on R_r^d unless the model can be uniquely determined on both $R_r^d \cap R_i^0$ and $R_r^d \cap R_j^0$ for a pair of adjacent i, j. In Example 2, if $R_1^d = \{\mathbf{x} : x_2 \leq 0\}$, dropping the point $(-1, -1)$ makes the model linear on R_1^d. Furthermore, since in model (2.1) we did not exclude the possibility of $\beta_i = \beta_j$ for nonadjacent i, j, to ensure the detection of nonlinearity on R_r^d, the model has to be uniquely determined on $R_r^d \cap R_i^0$ and $R_r^d \cap R_j^0$ for at least one pair of adjacent i, j. To this end, we need

$$\frac{1}{n}\sum_{t=1}^{n} \mathbf{x}_t \mathbf{x}_t' \mathbf{1}_{(\mathbf{x}_t \in R_r^d \cap R_{k+i}^0)} \tag{2.2}$$

be positive definite for $i = 1, 2$ and some $k \in \{0, \cdots, l^0 - 1\}$.

Asymptotically, we need (2.2) to hold with probability approaching 1 as n becomes large, and its LHS should not gradually degenerate to a singular matrix. This in turn can be stated as follow:

For any set A, let $\lambda(A)$ be the smallest eigenvalue of $E[\mathbf{x}\mathbf{x}'\mathbf{1}_{(\mathbf{x} \in A)}]$. Define $\lambda(\{R_j^d\}_{j=1}^{L+1}) = \max_{j,k} \min_{i=1,2}\{\lambda(R_j^d \cap R_{k+i}^0)\}$. We will need d^0 to be *identifiable*, defined as follows:

Definition 2.1 d^0 is identifiable w.r.t. L if for every $d \neq d^0$,

$$\Lambda = \inf_{\{R_j^d\}_{j=1}^{L+1}} \lambda(\{R_j^d\}_{j=1}^{L+1}) > 0, \tag{2.3}$$

where the inf is taken over all possible partitions of the form $\{R_j^d\}_{j=1}^{L+1}$.

If $l^0 = 1$, then $k = 0$ and $\lambda(\{R_j^d\}_{j=1}^{L+1}) = \max_j \min_{i=1,2}\{\lambda(R_j^d \cap R_i^0)\}$. Now, let us examine the identifiability of d^0 in the two examples given above.

Example 1 (continued) d^0 is not identifiable w.r.t. $L = 2$.

Since for $d = 2$, and $(\tau_1, \tau_2) = (-1, 1)$, either $P(R_j^d \cap R_1^0) = 0$ or $P(R_j^d \cap R_2^0) = 0$ for all $j = 1, 2, 3$.

d^0 is identifiable w.r.t. $L = 1$. Since for any τ_1, there exists $r \in \{1, 2\}$ such that $E[\mathbf{x}\mathbf{x}'\mathbf{1}_{(\mathbf{x} \in R_r^d \cap R_i^0)}]$ is positive definite, for $i = 1, 2$. ¶

Example 2 (continued) d^0 is not identifiable w.r.t. $L = 2$.

Let $d = 2$. If $(\tau_1, \tau_2) = (-0.5, 0.5)$ then each of $R_j^d \cap R_i^0$ will contain no more than two points with positive masses, $i = 1, 2$, $j = 1, 2, 3$. Hence $E[\mathbf{x}\mathbf{x}'\mathbf{1}_{(\mathbf{x} \in R_j^d \cap R_i^0)}]$ will be degenerate for all i, j.

d^0 is identifiable w.r.t. $L = 1$. Since for any τ_1 and $i = 1, 2$, there exists $r \in \{1, 2\}$ such that $R_r^d \cap R_i^0$ contain at least 3 points, with positive masses, which are not collinear. Hence $E\{\mathbf{x}\mathbf{x}'\mathbf{1}_{(\mathbf{x} \in R_r^d \cap R_i^0)}\}$ is positive definite. Because we have effectively just 4 choices of τ_1, the eigenvalues of $E\{\mathbf{x}\mathbf{x}'\mathbf{1}_{(\mathbf{x} \in R_r^d \cap R_i^0)}\}$, $i = 1, 2$, are positive. ¶

In more complicated cases, the identifiability condition may not be easy to verify. An equivalent condition is given in the theorem below. This theorem is essential in showing that the two methods of estimating d^0 given in the next section are consistent.

Theorem 2.1 *The following conditions are equivalent:*

(i) d^0 *is identifiable w.r.t.* L,

(ii) for any $d \neq d^0$, *there exist sets* $\{A_j^d\}_{j=1}^{L+1}$ *of the form* $A_j^d = \{\mathbf{x} : a_j \leq x_d \leq b_j\}$ *such that*

(a) $\lambda(A_s^d \cap R_{k+i}^0) > 0$ *for some* $0 \leq k \leq l^0 - 1$ *and all* $i = 1, 2$, $s = 1, \cdots, L + 1$, *and*

(b) for any partition $\{R_j^d\}_{j=1}^{L+1}$, $A_s^d \subset R_r^d$ *for some* $r, s \in \{1, \cdots, L + 1\}$. ¶

Before proving the theorem, let us find A_j^d's in the two examples given above. Assume, arbitrarily, $d = 2$. In Example 1, let $A_1^d = \{\mathbf{x} : -2 \leq x_2 \leq -0.5\}$ and $A_2^d = \{\mathbf{x} : 0.5 \leq x_2 \leq 2\}$. Then, A_1^d and A_2^d satisfy (ii) in Theorem 2.1. In Example 2, $A_1^d = \{\mathbf{x} : -1 \leq x_2 \leq 0\}$ and $A_2^d = \{\mathbf{x} : 0 \leq x_2 \leq 1\}$. Note that in this case, $A_1^d \cap A_2^d = \{0\}$; the sets overlap.

For any measurable set C in $\mathbf{R}^1$, let

$$\lambda^d(C) = \min_{i=1,2} \lambda(\{\mathbf{x} : x_d \in C\} \cap R_i^0).$$

Lemma 2.1 *$\lambda^d([a,u])$ is right continuous in u. $\lambda^d([u,b])$ is left continuous in u.*

Also, $\lim_{b \to -\infty} \lambda^d((-\infty, b]) = 0$, $\lim_{a \to \infty} \lambda^d([a, +\infty)) = 0$ *and* $\lambda^d(\{a\}) = 0$.

Proof Let $A = \{\mathbf{x} : a \leq x_d \leq u\} \cap R_1^0$, $A_\delta = \{\mathbf{x} : u < x_d \leq u + \delta\} \cap R_1^0$ and $A_+ = \{\mathbf{x} : a \leq x_d \leq u + \delta\} \cap R_1^0$. Then $A_+ = A \cup A_\delta$. Let $\mathbf{a}$ be the normalized eigenvector corresponding to $\lambda(A)$, the smallest eigenvalue of $E[\mathbf{x}\mathbf{x}'\mathbf{1}_{(\{\mathbf{x} \in A\})}]$. Then

$$\begin{aligned}
\lambda(A) &= \mathbf{a}'E[\mathbf{x}\mathbf{x}'\mathbf{1}_{(\{\mathbf{x} \in A\})}]\mathbf{a} \\
&= \mathbf{a}'E[\mathbf{x}\mathbf{x}'\mathbf{1}_{(\{\mathbf{x} \in A_+\})}]\mathbf{a} - \mathbf{a}'E[\mathbf{x}\mathbf{x}'\mathbf{1}_{(\{\mathbf{x} \in A_\delta\})}]\mathbf{a} \\
&\geq \lambda(A_+) - \mathbf{a}'E[\mathbf{x}\mathbf{x}'\mathbf{1}_{(\{\mathbf{x} \in A_\delta\})}]\mathbf{a} \\
&\geq \lambda(A_+) - tr(E[\mathbf{x}\mathbf{x}'\mathbf{1}_{(\{\mathbf{x} \in A_\delta\})}]) \\
&= \lambda(A_+) - E[\mathbf{x}'\mathbf{x}\mathbf{1}_{(\{\mathbf{x} \in A_\delta\})}].
\end{aligned}$$

By the dominated convergence theorem, $E[\mathbf{x}'\mathbf{x}\mathbf{1}_{(\{\mathbf{x} \in A_\delta\})}] = E[\mathbf{x}'\mathbf{x}\mathbf{1}_{(\{\mathbf{x}: u < x_d \leq u+\delta\} \cap R_1^0)}]$ converges to 0 as $\delta \to 0_+$. Therefore, $\lambda(A) \leq \lambda(A_+) \leq \lambda(A) + o(1)$ and $\lambda(E[\mathbf{x}\mathbf{x}'\mathbf{1}_{(\{\mathbf{x}: a \leq x_d \leq u\} \cap R_1^0)}])$ is right continuous in u. Replacing R_1^0 by R_2^0 in the above argument, we have that $\lambda(E[\mathbf{x}\mathbf{x}' \mathbf{1}_{(\{\mathbf{x}: a \leq x_d \leq u\} \cap R_2^0)}])$ is right continuous in u. Since $\lambda^d([a,u])$ is the minimum of the two right continuous functions, it is also right continuous.

Now, let $A = \{\mathbf{x} : u \leq x_d \leq b\} \cap R_1^0$, $A_\delta = \{\mathbf{x} : u - \delta \leq x_d < u\} \cap R_1^0$ and $A_- = \{\mathbf{x} : u - \delta \leq x_d \leq b\} \cap R_1^0$. Then $A_- = A \cup A_\delta$. Let $\mathbf{a}$ be the normalized eigenvector corresponding to $\lambda(A)$, the smallest eigenvalue of $E[\mathbf{x}\mathbf{x}'\mathbf{1}_{(\{\mathbf{x}\in A\})}]$. Then

$$\begin{aligned}
\lambda(A) &= \mathbf{a}'E[\mathbf{x}\mathbf{x}'\mathbf{1}_{(\{\mathbf{x}\in A\})}]\mathbf{a} \\
&= \mathbf{a}'E[\mathbf{x}\mathbf{x}'\mathbf{1}_{(\{\mathbf{x}\in A_-\})}]\mathbf{a} - \mathbf{a}'E[\mathbf{x}\mathbf{x}'\mathbf{1}_{(\{\mathbf{x}\in A_\delta\})}]\mathbf{a} \\
&\geq \lambda(A_-) - \mathbf{a}'E[\mathbf{x}\mathbf{x}'\mathbf{1}_{(\{\mathbf{x}\in A_\delta\})}]\mathbf{a} \\
&\geq \lambda(A_-) - tr(E[\mathbf{x}\mathbf{x}'\mathbf{1}_{(\{\mathbf{x}\in A_\delta\})}]) \\
&= \lambda(A_-) - E[\mathbf{x}'\mathbf{x}\mathbf{1}_{(\{\mathbf{x}\in A_\delta\})}].
\end{aligned}$$

By the dominated convergence theorem, $E[\mathbf{x}'\mathbf{x}\mathbf{1}_{(\{\mathbf{x}\in A_\delta\})}] = E[\mathbf{x}'\mathbf{x}\mathbf{1}_{(\{\mathbf{x}:u-\delta\leq x_d<u\}\cap R_1^0)}]$ converges to 0 as $\delta \to 0_+$. Therefore, $\lambda(A) \leq \lambda(A_-) \leq \lambda(A) + o(1)$ and $\lambda(E[\mathbf{x}\mathbf{x}'\mathbf{1}_{(\{\mathbf{x}:u\leq x_d\leq b\}\cap R_1^0)}])$ is left continuous in u. Replacing R_1^0 by R_2^0 in the above argument, we have that $\lambda(E[\mathbf{x}\mathbf{x}'\mathbf{1}_{(\{\mathbf{x}:u\leq x_d\leq b\}\cap R_2^0)}])$ is left continuous in u. Since $\lambda^d([u,b])$ is the minimum of the two left continuous functions, so it is also left continuous.

Observe that

$$0 \leq \lambda^d([a,+\infty)) \leq tr(E[\mathbf{x}\mathbf{x}'\mathbf{1}_{(\{\mathbf{x}:a\leq x_d<\infty\}\cap R_1^0)}]) \leq E[\mathbf{x}'\mathbf{x}\mathbf{1}_{(\{\mathbf{x}:a\leq x_d<\infty\}\cap R_1^0))}].$$

By the dominated convergence theorem, the RHS converges to 0 as $a \to \infty$. Thus

$$\lim_{a\to\infty} \lambda^d([a,+\infty)) = 0.$$

Similarly,

$$0 \leq \lambda^d(-\infty,b]) \leq tr(E[\mathbf{x}\mathbf{x}'\mathbf{1}_{(\{\mathbf{x}:-\infty<x_d\leq b\}\cap R_1^0)}]) \leq E[\mathbf{x}'\mathbf{x}\mathbf{1}_{(\{\mathbf{x}:-\infty<x_d\leq b\}\cap R_1^0)}].$$

By the dominated convergence theorem again, the RHS converges to 0 as $b \to -\infty$. Thus

$$\lim_{b\to-\infty} \lambda^d((-\infty,b]) = 0.$$

Since the $(d+1)$th row of the matrix $E[\mathbf{x}\mathbf{x}'\mathbf{1}_{(\{\mathbf{x}:x_d=a\}\cap R_1^0)}]$ is its first row multiplied by a, its rank is less than or equal to p and hence it is degenerate. So does the rank of $E[\mathbf{x}\mathbf{x}'\mathbf{1}_{(\{\mathbf{x}:x_d=a\}\cap R_2^0)}]$. Hence $\lambda^d(\{a\}) = 0$. ¶

Let $b_L^* = \sup\{b : \lambda^d([b, +\infty)) \geq \Lambda\}$ where $\Lambda > 0$ is given by Definition 2.1, $b_{L+1}^* = \infty$, and, recursively, $b_{j-1}^* = \sup\{b \leq b_j^* : \lambda^d([b, b_j^*]) \geq \Lambda\}$, $j = 2, \ldots, L$, where, by convention, $b_{j-1}^* = -\infty$ if $\{b \leq b_j^* : \lambda^d([b, b_j^*]) \geq \Lambda\} = \phi$.

Lemma 2.2 *Suppose d^0 is identifiable w.r.t. L. Let $b_0^* = -\infty$. Then*

(i) $-\infty = b_0^* < b_1^* < \ldots < b_L^* < b_{L+1}^* = \infty$, *and*

(ii) $\lambda^d((-\infty, b_1^*]) \geq \Lambda$.

Proof (i) Lemma 2.1 implies $\lim_{a \to \infty} \lambda^d([a, \infty)) = 0$, so $b_L^* < \infty$. And $b_L^* > -\infty$.

For if it were not, i.e., $b_L^* = -\infty$, then since $\lim_{b \to -\infty} \lambda^d((-\infty, b]) = 0$, there exists $\tau_1 \in (-\infty, \infty)$, such that $\lambda^d((-\infty, \tau_1]) < \Lambda$. In view of the definition of b_L^* and the assumption that $b_L^* = -\infty$, we have that $\lambda^d((\tau_1, \infty)) < \Lambda$. For any $\tau_2, \cdots, \tau_L$ such that $-\infty = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_L < \tau_{L+1} = \infty$, we have $\lambda^d((\tau_{j-1}, \tau_j]) < \Lambda$, $j = 1, \cdots, L+1$. This contradicts to the definition of Λ. So, $-\infty < b_L^* < \infty$.

Assume that $b_j^*, \cdots, b_L^*$ have been well defined and satisfy $-\infty < b_j^* < \cdots < b_L^* < \infty$. We will now show that $-\infty < b_{j-1}^* < b_j^*$.

By Lemma 2.1, $\lambda^d(\{a\}) = 0$ and $\lambda^d([u, b])$ is left continuous in u. Hence, $b_{j-1}^* < b_j^*$.

Suppose $b_{j-1}^* = -\infty$. Since $\lim_{b \to -\infty} \lambda^d((-\infty, b]) = 0$, there exists $\tau_{j-1} \in (-\infty, b_j^*)$ such that $\lambda^d((-\infty, \tau_{j-1}]) < \Lambda$. For this τ_{j-1}, let $\tau_0 = \infty$ and choose $\tau_1, \cdots, \tau_{j-2}$ such that $\infty = \tau_0 < \tau_1 < \cdots < \tau_{j-2} < \tau_{j-1}$. Then

$$\lambda^d((\tau_{k-1}, \tau_k]) \leq \lambda^d((-\infty, \tau_{j-1}]) < \Lambda, \qquad k = 1, \cdots, j-1.$$

Since $b_{j-1}^* = -\infty$, $\lambda^d([\tau_{j-1}, b_j^*]) < \Lambda$. By right continuity of $\lambda^d([a, \cdot])$, there exists $\delta_j > 0$ such that $\tau_j = b_j^* + \delta_j \in (b_j^*, b_{j+1}^*)$ and $\lambda^d([\tau_{j-1}, \tau_j]) < \Lambda$. Repeating this argument we can see that there exists $\delta_k > 0$, such that $\tau_k = b_k^* + \delta_k \in (b_k^*, b_{k+1}^*)$ and $\lambda^d([\tau_{k-1}, \tau_k]) < \Lambda$, where $k = j, \cdots, L$. By the definition of b_L^*, $\lambda^d([\tau_L, \infty)) < \Lambda$.

In summary, we have

$$\lambda^d((\tau_{k-1}, \tau_k]) \leq \lambda^d([\tau_{k-1}, \tau_k]) < \Lambda, \qquad k = 1, \cdots, L,$$

and $\lambda^d((\tau_L, \infty)) < \Lambda$. That is, the partition $\{R_j^d\}_{j=1}^{L+1}$, where $R_j^d = \{\mathbf{x} : x_d \in (\tau_{j-1}, \tau_j]\}$, satisfy $\min_{i=1,2} \lambda(R_j^d \cap R_i^0) = \lambda^d((\tau_{j-1}, \tau_j]) < \Lambda$, $j = 1, \cdots, L+1$. This again contradicts the definition of Λ. By induction, $-\infty < b_{j-1}^* < b_j^*$ for $j = 2, \cdots, L+1$. Thus, (i) is verified.

(ii) If not, $\lambda^d((-\infty, b_1^*]) < \Lambda$. Then, by the right continuity of $\lambda^d([a, \cdot])$, there exists $\delta_1 > 0$ such that $\tau_1 = b_1^* + \delta_1 < b_2^*$ and $\lambda^d((-\infty, \tau_1]) < \Lambda$. By the definition of b_1^*, $\lambda^d([\tau_1, b_2^*]) < \Lambda$ and hence there exists $\delta_2 > 0$, such that $\tau_2 = b_2^* + \delta_2 < b_3^*$ and $\lambda^d([\tau_1, \tau_2]) < \Lambda$.

By repeating this process we shall see that there exists $-\infty = \tau_0 < \tau_1 < \cdots < \tau_{L-1} < b_L^* < \tau_L = b_L^* + \delta_L < \tau_{L+1} = \infty$ such that $\lambda^d((\tau_{j-1}, \tau_j]) < \Lambda$, $j = 1, \cdots, L+1$.

This leads again to a contradiction to the definition of Λ. ¶

Proof of Theorem 2.1 Without loss of generality, $l^0 = 1$ is assumed. Suppose (ii) holds. The condition $\lambda(A_s^d \cap R_i^0) > 0$ for all s and i implies $\min_{i,s}\{\lambda(A_s^d \cap R_i^0)\} > 0$. Then, $\lambda(\{R_j^d\}_{j=1}^{L+1}) \geq \min_{i=1,2} \lambda(R_r^d \cap R_i^0) \geq \min_{i=1,2} \lambda(A_s^d \cap R_i^0) \geq \min_{i,s}\{\lambda(A_s^d \cap R_i^0)\}$. We conclude that d^0 is identifiable w.r.t. L by taking the infima in the last inequality.

Now assume (i) holds.

Let $A_j^d = \{\mathbf{x} : x_d \in [b_{j-1}^*, b_j^*]\}$, where b_j^* is defined in Lemma 2.2, $j = 1, \cdots, L+1$. By Lemma 2.2, $-\infty = b_0^* < b_1^* < \cdots < b_L^* < b_{L+1}^* = \infty$, and $\lambda^d((-\infty, b_1^*]) \geq \Lambda$. By the definition of b_j^*'s, $\lambda^d([u, b_j^*]) \geq \Lambda$ for all $u < b_{j-1}^*$, $j = 2, \cdots, L+1$. By Lemma 2.1, $\lambda^d([u, b])$ is left continuous in u. Hence, $\lambda^d([b_{j-1}^*, b_j^*]) \geq \Lambda$, $j = 2, \cdots, L+1$. By the definition of $\lambda^d(\cdot)$, $\lambda(A_1^d \cap R_i^0) = \lambda(\{\mathbf{x} : x_d \in (-\infty, b_1^*]\} \cap R_i^0) \geq \lambda^d((-\infty, b_1^*]) \geq \Lambda$, and $\lambda(A_s^d \cap R_i^0) = \lambda(\{\mathbf{x} : x_d \in [b_{s-1}^*, b_s^*]\} \cap R_i^0) \geq \lambda^d([b_{s-1}^*, b_s^*]) \geq \Lambda$, $s = 2, \cdots, L+1$. That is, $\{A_j^d\}_{j=1}^{L+1}$ satisfy (a) in Theorem 2.1 (ii).

It remains to show that for any $\{R_j^d\}_{j=1}^{L+1}$, where $R_j^d = \{\mathbf{x} : x_d \in (\tau_{j-1}, \tau_j]\}$, there exists $r, s \in \{1, \cdots, L+1\}$ such that $R_r^d \subset A_s^d$. We shall show it by sequential exhaustive argument. If $R_1^d \not\supset A_1^d$ then $\tau_1 < b_1^*$. If $R_2^d \not\supset A_i^d$, $i = 1, 2$, then $\tau_2 < b_2^*$. If $R_3^d \not\supset A_i^d$, $i = 1, 2, 3$, then $\tau_3 < b_3^*$. $\cdots$. If $R_L^d \not\supset A_i^d$, $i = 1, \cdots, L$, then, $\tau_L < b_L^*$ and hence $R_{L+1}^d \supset A_{L+1}^d$.

This completes the proof of Theorem 2.1. ¶

Corollary 2.2 Suppose the distribution of $\mathbf{z}_1 = (x_{11}, \ldots, x_{1p})'$ has support $(a_1, b_1) \times \cdots \times (a_p, b_p)$, where $-\infty \leq a_i < b_i \leq \infty$, $i = 1, \ldots, p$. Then for any integer $L \geq l^0$, d^0 is identifiable w.r.t. L.

Proof For any $d \neq d^0$, any $L+1$ mutually exclusive subsets of the form $\{\mathbf{x} : x_d \in [\alpha, \eta]\}$, where $\alpha < \eta$ and $[\alpha, \eta] \subset (a_d, b_d)$, will serve as the $\{A_j^d\}_{j=1}^{L+1}$ in Theorem 2.1. Hence the identifiability of d^0 follows. ¶

Corollary 2.3 Suppose the support of distribution of $\mathbf{z}_1 = (x_{11}, \ldots, x_{1p})'$ is a convex subset of $\mathbf{R}^p$. Then for any integer $L \geq l^0$, d^0 is identifiable w.r.t. L.

Proof Since the support of distribution of $\mathbf{z}_1$ is convex, it contains a subset of the form $(a_1, b_1) \times \ldots \times (a_p, b_p)$, where $-\infty \leq a_i < b_i \leq \infty$, $i = 1, \ldots, p$. For any $d \neq d^0$, any $L+1$ mutually exclusive subsets of the form $\{\mathbf{x} : x_d \in [\alpha, \eta]\}$, where $\alpha < \eta$ and $[\alpha, \eta] \subset (a_d, b_d)$, will serve as the $\{A_j^d\}_{j=1}^{L+1}$ in Theorem 2.1. ¶

2.2 Estimation procedures

The least squares criterion is used to select d. The idea is simple. Suppose that d^0 is identifiable and that a wrong d were chosen as the threshold variable. Then for sufficiently large n, on at least one of the R_j^d's, say R_r^d, the model exhibits nonlinearity, resulting in a large sum of squared errors on R_r^d. Hence, the total sum of squared residuals is large. In contrast, if d^0 were chosen, by adjusting the $\hat{\tau}_j$'s, the model on each $\{\mathbf{x} : \hat{\tau}_{j-1} < x_{d^0} \le \hat{\tau}_j\}$ would be roughly linear, resulting in a smaller total sum of squared errors. Therefore, $\hat{d}$ should be chosen as the d resulting in the smallest total sum of squared errors. To simplify the implementation of this idea, let

$$X_n := \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}, \; Y_n := \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \tilde{\epsilon}_n := \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

$$I_n(A) := diag(\mathbf{1}_{(\mathbf{x}_1 \in A)}, \ldots, \mathbf{1}_{(\mathbf{x}_n \in A)}), \; A \subset R^{p+1}$$

$$X_n(A) := I_n(A) X_n,$$

$$H_n(A) := X_n(A)[X_n'(A) X_n(A)]^- X_n'(A),$$

$$S_n(A) := Y_n'(I_n(A) - H_n(A)) Y_n,$$

and

$$T_n(A) := \tilde{\epsilon}_n' H_n(A) \tilde{\epsilon}_n,$$

where in general for any matrix M, M^- denotes a generalized inverse. Note that $X_n(A)$, $H_n(A)$ and $S_n(A)$ are, respectively, the covariates, "hat matrix" and the sum of squared residual errors from fitting a linear model based on just the observations in A.

Finally, for any $\{R_j^d\}_{j=1}^{l+1}$ define the total sum of squares over different regions as

$$S_n^d(\tau_1, \ldots, \tau_l) := \sum_{i=1}^{l+1} S_n(R_j^d).$$

The first method for estimating d^0 is given below.

Method 1 Suppose d^0 is identifiable w.r.t. L. Choose d to minimize the sum of squared errors. More precisely, let $S_n^d := S_n^d(\hat{\tau}_1^d, \ldots, \hat{\tau}_L^d)$, where $\hat{\tau}_1^d < \cdots < \hat{\tau}_L^d$ minimize $S_n^d(\tau_1, \ldots, \tau_L)$ over all

$(\tau_1, \ldots, \tau_L)$. Select $\hat{d}$ such that $S_n^{\hat{d}} \leq S_n^d$ for $d = 1, \ldots, p$. Should multiple minimizers occur, we define $\hat{d}$ to be the smallest of them.

Remark When calculating $S_n(R_j^d)$, at least p data points must be in R_j^d to ensure the regression coefficients on that segment are uniquely determined.

This method requires intensive computation. As Feder (1975a) and other authors note, $S_n^d(\tau_1, \cdots, \tau_L)$ may not be differentiable at the true change points. So to minimize $S_n^d(\tau_1, \cdots, \tau_L)$, one has to search all $(\tau_1, \cdots, \tau_L)$. Fortunately, we can do this by restricting ourselves to the finite set $\{x_{1d}, \cdots, x_{nd}\}$, without loss of generality. Even so, exhausting all $(\tau_1^d, \cdots, \tau_L^d)$ for any d needs $\binom{n}{L} \times (L+1)$ linear fits. Although a method more efficient than actually doing the $\binom{n}{L} \times (L+1)$ fits exists, there is still a lot of work for any $L \geq 3$ and large n. So, under stronger conditions, we give another more efficient method. This method is based on the following idea. Suppose $\mathbf{z}_1 = (x_{11}, \ldots, x_{1p})'$ is a continuous random vector and the support of its distribution is $(a_1, b_1) \times \ldots \times (a_p, b_p)$, where $-\infty \leq a_i < b_i \leq \infty$, $(i = 1, \cdots, p)$. Then for any d we can partition (a_d, b_d) into $2L + 2$ disjoint intervals such that there are an equal number of observations in each of the intervals. For any $d \neq d^0$, on all these intervals the model will exhibit nonlinearity and hence the linear fits will result in larger sum of squared errors. If $d = d^0$, then there are at least $L + 1$ intervals that are entirely embedded in one of the $(\tau_{j-1}^0, \tau_j^0]$'s. Hence, on those intervals, the model is linear and the sum of squared errors from linear fits are smaller. Thus, the total of the smallest $L + 1$ sums of squared errors for $d = d^0$ is expected to be smaller than that for $d \neq d^0$. It is easy to see that the above argument holds as long as the number of partitions is no less than $L + 2$. The practical advantages of choosing a number larger than $L + 2$ will be discussed in Section 3.2. We summarize the above discussion as follows:

Method 2 Suppose $\mathbf{z}_1 = (x_{11}, \ldots, x_{1p})'$ is a continuous random vector and the support of its distribution is $(a_1, b_1) \times \ldots \times (a_p, b_p)$, where $-\infty \leq a_i < b_i \leq \infty$, $i = 1, \ldots, p$. Let r_j^d be the $[100 \times j/(2L+2)]$th percentile of x_{td}'s, $\hat{R}_j^d = \{\mathbf{x}_1 : x_{1d} \in (r_{j-1}^d, r_j^d]\}$, $j = 1, \ldots, 2L+2$. Select $\hat{d}$, so that

$$\tilde{S}_n^{\hat{d}} \leq \tilde{S}_n^d,$$

for all $d = 1, \cdots, p$, where

$$\tilde{S}_n^d := \sum_{i=1}^{L+1} S_n(\hat{R}_{(i)}^d)$$

and $S_n(\hat{R}_{(i)}^d)$ is the ith smallest of $S_n(\hat{R}_1^d), \cdots, S_n(\hat{R}_{2L+2}^d)$.

Remark For any d, Method 2 requires only $2L+2$ linear fits (independent of n). The computational effort is significantly reduced compared with Method 1.

Now, with d^0 estimated above, we can assume that d^0 is known and estimate other parameters. For simplicity, we shall drop the superscript, d, on S_n^d and τ_j^d's in the rest of this section.

First we estimate l^0 and the thresholds, $\tau_1^0, \ldots, \tau_{l^0}^0$, by minimizing the modified Schwarz' criterion (Schwarz, 1978),

$$MIC(l) := \ln[S(\hat{\tau}_1, \ldots, \hat{\tau}_l)/(n - p^*)] + p^* \frac{c_0 (\ln n)^{2+\delta_0}}{n}, \tag{2.4}$$

for some constants $c_0 > 0, \delta_0 > 0$. In equation (2.4), $p^* = (l+1)p + l \approx (l+1)(p+1)$ is the total number of fitted parameters, and for any fixed l, $\hat{\tau}_1, \ldots, \hat{\tau}_l$ are the least squares estimates which minimize $S_n(\tau_1, \ldots, \tau_l)$ subject to $-\infty = \tau_0 < \tau_1 < \cdots < \tau_{l+1} = \infty$.

Recall that Schwarz' criterion (SC) is defined by

$$SC(l) = ln[S(\hat{\tau}_1, \ldots, \hat{\tau}_l)/(n - l)] + l \frac{2ln(n)}{n}. \tag{2.5}$$

We can see that the distinction between $MIC(l)$ and $SC(l)$ lies in the severity of the penalty for overspecification. And a severer penalty is essential for the correct specification of a non-Gaussian, segmented regression model, since $SC(l)$ is derived under Gaussian assumption (c.f., Yao, 1988). Both criteria are sometimes referred as penalized least squares.

With estimates, $\hat{l}$ of l^0, and $\hat{\tau}_i$ for τ_i^0, $i = 1, \ldots, \hat{l}$ available, we then estimate the other regression parameters $\{\tilde{\beta}_i^0\}$ and the residual variance σ_0^2 by the ordinary least squares estimates,

$$\hat{\tilde{\beta}}_i = [X_n'(\hat{R}_i)X_n(\hat{R}_i)]^{-}X_n'(\hat{R}_i)\mathbf{Y}_n, \ i = 1, \ldots, \hat{l}+1,$$

and

$$\hat{\sigma}^2 = S_n(\hat{\tau}_1, \ldots, \hat{\tau}_l)/(n - \hat{p^*}),$$

where $\hat{R}_i = \{\mathbf{x} : \hat{\tau}_{i-1} < x_{d^0} \leq \hat{\tau}_i\}$, $\hat{p^*} = (\hat{l}+1)p + \hat{l}$. Under regularity conditions essential for the identifiability of the regression parameters, we shall see in Chapter 3 that the ordinary least squares estimates $\hat{\tilde{\beta}}_j$ will be unique with probability approaching 1, for $j = 1, \ldots, \hat{l}+1$, as $n \to \infty$.

While for a really large sample size, we do not expect the choice of δ_0 and c_0 to be crucial, for small to moderate sample sizes, this choice does influence the model selection. Below, we briefly discuss the choice of c_0 and δ_0.

In general, when selecting models, a relatively large penalty term would be preferable for the models that can be easily identified. This is because a larger penalty will greatly reduce the probability of overestimation while not risking underestimation too much. However, if the model is difficult to identify (e.g., a continuous model with $||\tilde{\beta}_{j+1} - \tilde{\beta}_j||$ small), the penalty should not be too large since the risk of underestimation is now high.

Another factor influencing the choice of the penalty is the error distribution. A distribution with heavy tails is likely to generate extreme values, making it look as though a change in response has occurred. To counter this effect, one needs a heavier penalty. In fact, if ϵ_t has only finite order moments, a penalty of order $n^{\alpha-1}$ for some $\alpha > 0$ is needed to make the estimation of l^0 consistent.

Given that the best criterion is model dependent and no uniformly optimal choice can be made, the following considerations guide us to a reasonable choice of δ_0 and c_0:

(1) From the proof of Lemma 3.2 in Section 3.1, we shall see that it is possible that the exponent $2 + \delta_0$ in the penalty term of MIC may be further reduced, while keeping the model selection procedure consistent. And since the Schwarz' criterion (where the exponent is 1) is obtained by maximizing the posterior likelihood in a model selection paradigm and is widely used in model selection problems, it may be used as a baseline reference. Adopting such a view, δ_0 should be small to reduce the potential risk of underestimation when the noise is normal and n is not large.

(2) For a small sample, it is practically difficult to distinguish normal and double exponential noise, or t distributed noise. And, hence, one would not expect the choice of SC or any other reasonable criterion to make a drastic difference.

(3) As Yao (1988) noted for large samples, SC tends to overestimate l^0 if the noise is not normal. We observe such overestimation in our simulations under different model specifications when $n = 50$ (see Section 3.3).

Based on (1), we should choose a small δ_0. And by (2), with δ_0 chosen, we can choose some moderate n_0, and solve for c_0 by forcing MIC equal to SC at n_0. By (3), $n_0 < 50$ seems desirable. In the simulation reported in the next section, we (arbitrarily) choose δ_0 to be 0.1 (which is considered to be small). With such a δ_0, we arbitrarily choose $n_0 = 20$ and solve for c_0. We get $c_0 = 0.299$.

In summary, since the "best" selection of the penalty is model dependent for finite samples, no optimal pair of (c_0, δ_0) can be recommended. On the other hand, our choice of $\delta_0 = 0.1$ and $c_0 = 0.299$ performs reasonably well for most of the cases we experimented with in our simulation. The simulation results are reported in Section 3.3. Further study is needed on the choice of δ_0 and c_0 under different assumptions.

A data set used in Henderson and Velleman (1981) is analyzed below to illustrate the estimation procedures proposed above. The data consist of measurements of three variables, miles per gallon (y), weight (x_1) and horse power (x_2), on thirty eight 1978-79 model automobiles. The dependence of y on x_1 and x_2 is of interest. Graphs of the data show a certain nonlinear dependence structure between y and x_1 (see Figure 2.3).

Suppose we want to fit a model of the form (2.1). In this case, it becomes

$$y_t = \beta_{i0} + \beta_{i1}x_{t1} + \beta_{i2}x_{t2} + \epsilon_t, \ if \ x_{td} \in (\tau_{i-1}, \tau_i], i = 1, \ldots, l+1, \tag{2.6}$$

where ϵ_t is assumed to have zero mean and variance σ^2. To demonstrate the use of two methods of estimating d^0, let us ignore the information given by Figure 2.3 (which suggests $d^0 = 1$ and $l^0 = 1$) and estimate d^0 by calculation.

First, we (arbitrarily) choose $L = 2$ and apply Method 1. We get $S_n^1 = 120.0$ and $S_n^2 = 136.0$. Hence $\hat{d} = 1$ is chosen by Method 1. With $L = 2$ we get on applying Method 2, $S_n^1 = 14.6$ and $S_n^2 = 15.3$. Thus, $\hat{d} = 1$ is also chosen by Method 2. Both methods agree with the casual observation made above about Figure 2.3.

Next, with $d = 1$, we calculate and compare $MIC(l)$ for $l = 0, 1, 2$ to estimate l^0. For illustrative purposes, the constants c_0 and δ_0 in the penalty term of MIC are chosen as 0.2 and 0.05 respectively, to enable the piecewise model to remain competitive for this small sample example. The MIC values for $l = 0, 1, 2$ are 2.28, 2.11 and 2.31 respectively. Thus $l = 1$ is chosen by the criterion. Then with $l = 1$, $\hat{\tau}_1 = 2.7$ is obtained. With these estimates, the estimated coefficients are $(\hat{\beta}_{10}, \hat{\beta}_{11}, \hat{\beta}_{12}) = (48.82, -5.23, -0.08)$, $(\hat{\beta}_{20}, \hat{\beta}_{21}, \hat{\beta}_{22}) = (30.76, -1.84, -0.05)$ and $\hat{\sigma}^2 = 4.90$.

Finally, treating the MIC as a general model selection criterion rather than a tool for finding l^0, two more competing models are fitted to the data. These are

$$y_t = \beta_0 + \beta_1 x_{t1} + \epsilon_t, \tag{2.7}$$

and

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t1}^2 + \beta_3 x_{t2} + \epsilon_t. \tag{2.8}$$

From Figure 2.3, both models seem appealing. The MIC values for these two models are 2.24 and 2.12. Thus, the segmented model is chosen as the "best". Needless to say, it is only the "best" among the few models considered; further model reduction may be possible.

2.3 General remarks

In Section 2.1, we have discussed the identifiability of d^0. It can be seen from Corollary 2.3 that in many regression problems, d^0 can be treated as identifiable w.r.t. any $L \geq l^0$. But, it is important to realize that d^0 is not always uniquely identifiable and to know when it is not uniquely identifiable, in an asymptotic sense. It is also important to bear in mind the question of identifiability in a design problem. The results in Section 2.1 have provided an answer to these questions. Moreover, these results not only provide a foundation for estimating d^0 in model (2.1) for continuous covariates, but they also address the same problem when the covariates are discrete or ordered categorical. For example, one may want to know which of the two covariates, the dose of certain drug or age group, alters the dependent structure of blood pressure on the two. In this case, the identification of d^0 is important even when the change point is not uniquely defined.

As in the example of automobiles, the MIC we proposed in the last section should be treated as a method of model selection, and not merely as a tool of estimating d^0. In fact, in the case when d^0 is only identifiable w.r.t. some number less than the known L, d^0 and l^0 can be jointly estimated by minimizing MIC over all the combinations of $d(\leq p)$ and $l(\leq L)$. In the next chapter, the consistency of these estimates, under certain conditions, will be shown.

From a much broader perspective, our estimation procedures can be seen as a general adaptive model fitting technique. The upper bound L on the number of segments is imposed to ensure computational feasibility and to avoid the "curse of dimensionality"; in other words, L ensures there are sufficient data to enable each piece of the model to be well estimated even when the covariate is a vector of high dimension. With this upper bound, the number of segments and the boundaries of each segment are selected by the data. It will be shown in the next chapter that these estimates are also consistent.

Chapter 3

ASYMPTOTIC RESULTS

FOR ESTIMATORS OF SEGMENTED REGRESSION MODELS

In this chapter, asymptotic results for the estimators given in the last chapter are proved. The exact conditions under which these results hold are stated and explained. It will be seen that these conditions seem realistic for many practical problems. More importantly, the techniques we use in this chapter constitute a foundation for the generalizations in Chapter 4 of Model (2.1). In some cases the parameter d^0 is known *a priori*, in such cases the notation required for presenting the proof of our results is relatively simple, and so we first prove the results for these cases. In Section 3.1 we establish the consistency of the estimated number of segments, the estimated thresholds and the estimated regression coefficients. Then, for the discontinuous model, an upper bound is given for the rate of convergence of the estimated change points. The asymptotic normality of the estimated regression coefficients and of the estimated variance of the noise is also established. In Section 3.2 we move to the case of unknown d^0 and prove the consistency of the two estimators of d^0 given in Section 2.2. It will be easy to see that the results proved in Section 3.1 still hold if d^0 is replaced by its consistent estimate. In Section 3.3, the finite sample behavior of these estimators is investigated by simulation for various models and noise distributions. Some general remarks are made in Section 3.4. The asymptotic normality of the various estimates for the continuous model is established in Section

3.5.

3.1 Asymptotic results when the segmentation variable is known

In this section, the parameter d in model (2.1) is assumed known. Consequently, we can simplify the notation at the beginning of Section 2.2. For any $-\infty \leq \alpha < \eta \leq \infty$, let

$$I_n(\alpha, \eta) := diag(\mathbf{1}_{(x_{1d} \in (\alpha, \eta])}, \ldots, \mathbf{1}_{(x_{nd} \in (\alpha, \eta])}),$$

$$X_n(\alpha, \eta) := I_n(\alpha, \eta) X_n,$$

and

$$H_n(\alpha, \eta) := X_n(\alpha, \eta)[X_n'(\alpha, \eta) X_n(\alpha, \eta)]^{-} X_n'(\alpha, \eta),$$

where in general for any matrix A, A^{-} will denote a generalized inverse while $\mathbf{1}_{(\cdot)}$ represents the indicator function. Similarly, let

$$Y_n(\alpha, \eta) := I_n(\alpha, \eta) Y_n, \ \tilde{\epsilon}_n(\alpha, \eta) := I_n(\alpha, \eta) \tilde{\epsilon}_n,$$

$$S_n(\alpha, \eta) := Y_n'[I_n(\alpha, \eta) - H_n(\alpha, \eta)] Y_n,$$

$$S_n(\tau_1, \ldots, \tau_l) := \sum_{i=1}^{l+1} S_n(\tau_{i-1}, \tau_i), \tau_0 := -\infty, \tau_{l+1} := \infty,$$

and

$$T_n(\alpha, \eta) := \tilde{\epsilon}_n' H_n(\alpha, \eta) \tilde{\epsilon}_n.$$

Observe that $S_n(\alpha, \eta)$ is just the error sum of squares when a linear model is fitted over the "threshold" interval $(\alpha, \eta]$. Also, let the forecast of Y_n on the interval $(\alpha, \eta]$, $\hat{Y}_n(\alpha, \eta)$, be defined by

$$\hat{Y}_n(\alpha, \eta) := H_n(\alpha, \eta) Y_n.$$

Then, in terms of true parameters, (2.1) can be rewritten in the vector form,

$$Y_n = \sum_{i=1}^{l^0+1} X_n(\tau_{i-1}^0, \tau_i^0) \tilde{\beta}_i^0 + \tilde{\epsilon}_n. \tag{3.1}$$

To establish the asymptotic theory for the estimation problems of Model (3.1), some assumptions have to be made. First, we assume an upper bound, L, of l^0 can be specified. This is because in practice, the sample size n is always finite and hence any l^0 that can be effectively identified is always bounded. We also assume the segmentation does occur at every true threshold, i.e., $\tilde{\beta}_j^0 \neq \tilde{\beta}_{j+1}^0$, $j = 1, \ldots, l^0$, so that these parameters are uniquely defined. The covariates $\{\mathbf{x}_t\}$ are assumed to be strictly stationary, ergodic random sequence. Further, $\{\mathbf{x}_t\}$ and the errors sequence $\{\epsilon_t\}$ are assumed independent. These are the basic assumptions underlying our analysis.

To simplify the problem further, we assume in this chapter that the errors $\{\epsilon_t\}$ are iid random variables with mean zero and variance σ_0^2. In addition, a *local exponential boundedness condition* is placed on the distribution of the errors $\{\epsilon_t\}$. A random variable Z is said to be *locally exponentially bounded* if there exist two positive constants, c_0 and T_0, such that

$$E(e^{uZ}) \leq e^{c_0 u^2}, \ for \ \ every \ |u| \leq T_0. \tag{3.2}$$

The above assumptions are summarized in

Assumption 3.0*:*

The covariates $\{\mathbf{x}_t\}$ and the errors $\{\epsilon_t\}$ are independent, where the $\{\mathbf{x}_t\}$ are strictly stationary and ergodic with $E(\mathbf{x}_1'\mathbf{x}_1) < \infty$, $\{\epsilon_t\}$ are iid with a locally exponentially bounded distribution having mean zero and variance σ_0^2. For the number of threshold l^0, there exists a known L such that $l^0 \leq L$. Also, for any $j = 1, \cdots, l^0$, $\tilde{\beta}_j^0 \neq \tilde{\beta}_{j+1}^0$.

Remark The local exponential boundedness condition is satisfied by any distribution with zero mean and a moment generating function with second derivative bounded around zero. Many distributions commonly used as error distributions such as those in the symmetrized exponential family are of this type, and hence all the theorems in this chapter will commonly apply.

The next assumption is required to identify the number of thresholds l^0 consistently.

Assumption 3.1

There exists $\delta \in (0, \min_{1\leq j\leq l^0}(\tau_{j+1}^0 - \tau_j^0)/2)$ such that both $E\{\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_i^0-\delta,\tau_i^0])}\}$ and $E\{\mathbf{x}_1\mathbf{x}_1'$ $\mathbf{1}_{(x_{1d}\in(\tau_i^0,\tau_i^0+\delta])}\}$ are positive definite for each of the true thresholds $\tau_1^0, \ldots, \tau_{l^0}^0$.

Under *Assumption 3.1*, the design matrix $X_n(\alpha, \eta)$ has full column rank a.s. as $n \to \infty$ for every open interval (α, η) containing at least one of $(\tau_i^0 - \delta, \tau_i^0 + \delta]$, $i = 1, \ldots, l^0$. So $\hat{\tilde{\beta}}(\alpha, \eta) = [X_n'(\alpha,\eta)X_n(\alpha,\eta)]^- X_n'(\alpha,\eta)Y_n$ will be unique with probability tending to 1 as $n \to \infty$.

It is easy to see that *Assumption 3.1* is satisfied if and only if the conditional covariances of $\mathbf{z}_1 = (x_{11}, \ldots, x_{1p})'$, $Cov(\mathbf{z}_1|x_{1d} \in (\tau_i^0 - \delta, \tau_i^0])$ and $Cov(\mathbf{z}_1|x_{1d} \in (\tau_i^0, \tau_i^0 + \delta])$, $(i = 1, \ldots, l^0)$, are both positive definite. *Assumption 3.1* means that the model can be uniquely determined over each of $\{\mathbf{x}_1 : x_{1d} \in (\tau_i^0 - \delta, \tau_i^0]\}$ and $\{\mathbf{x}_1 : x_{1d} \in (\tau_i^0, \tau_i^0 + \delta]\}$, $i = 1, \ldots, l^0$. The remark immediately after the proof of Theorem 3.1 will show that this assumption can be slightly relaxed.

To estimate the thresholds consistently, we need

Assumption 3.2

For any sufficiently small $\delta > 0$, $E\{\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_i^0-\delta,\tau_i^0])}\}$ and $E\{\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_i^0,\tau_i^0+\delta])}\}$ are positive definite, $i = 1, \cdots, l^0$. Also, $E(\mathbf{x}_1'\mathbf{x}_1)^u < \infty$ for some $u > 1$.

Obviously, *Assumption 3.2* implies *Assumption 3.1*.

If Model (3.1) is discontinuous at τ_j^0 for some $j = 1, \cdots, l^0$, it will be shown that the least squares estimate $\hat{\tau}_j$ converges to τ_j^0 at the rate no slower than $O_p(\ln^2 n/n)$, under the following assumption:

Assumption 3.3

(A.3.3.1) The covariates $\{\mathbf{x}_t\}$ *are iid random variables. Also,* $E(\mathbf{x}_1'\mathbf{x}_1)^u < \infty$ *for some* $u > 2$.

(A.3.3.2) Within some small neighborhoods of the true thresholds, x_{1d} *has a positive and continuous probability density function* $f_d(\cdot)$ *with respect to the one dimensional Lebesgue measure.*

(A.3.3.3) There exists one version of $E[\mathbf{x}_1\mathbf{x}_1'|x_{1d} = x]$ *which is continuous within some neighborhoods of the true thresholds and that version has been adopted.*

Remark *Assumptions (A.3.3.2)-(A.3.3.3)* are satisfied if $\mathbf{z}_1 = (x_1, \cdots, x_p)$ has a joint distribution in canonical form from the exponential family.

Note that *Assumptions 3.1-3.3* are made on the distribution of $\{\mathbf{x}_t\}$. When $\{\mathbf{x}_t\}$ are nonrandom, one may assume the empirical distribution function of $\{\mathbf{x}_t\}$ converges to a distribution function satisfying these assumptions.

Now, the main results of this section are presented in the next five theorems. Their proofs are given in the sequel.

Theorem 3.1 *Assume for the segmented linear regression model (3.1) that Assumptions 3.0 and 3.1 are satisfied. Then* $\hat{l}$*, the minimizer of (2.4), converges to* l^0 *in probability as* $n \to \infty$.

Remark In the nonlinear minimization of $S(\tau_1, \ldots, \tau_l)$, the possible values of $\tau_1 < \ldots < \tau_l$ may be limited to $\{x_{1d}, \ldots, x_{nd}\}$. This restriction induces no loss of generality.

Theorems 3.2 and 3.3 show that the estimates $\hat{\tau}$, $\hat{\tilde{\beta}}_j'$s and $\hat{\sigma}^2$ are consistent.

Theorem 3.2 *Assume for the segmented linear regression model (3.1) that Assumptions 3.0* *and 3.2 are satisfied. Then*

$$\hat{\tau} - \tau^0 = o_p(1),$$

where $\tau^0 = (\tau_1^0, \ldots, \tau_{l^0}^0)$ *and* $\hat{\tau} = (\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{l}})$ *is the least squares estimate of* τ^0 *based on* $l = \hat{l}$, *and* $\hat{l}$ *is a minimizer of* $MIC(l)$ *subject to* $l \leq L$.

Theorem 3.3 *If the marginal cdf* F_d *of* x_{1d} *satisfies the Lipschitz condition* $|F_d(x') - F_d(x'')| \leq C|x' - x''|$ *for some constant* C *in a small neighborhood of* $x_{1d} = \tau_j^0$ *for every* j, *then under the conditions of Theorem 3.2, the least squares estimates* $(\hat{\tilde{\beta}}_j,\ j = 1, \ldots, \hat{l})$ *based on the estimates* $\hat{l}$ *and* $\hat{\tau}_j$*'s as defined in Section 2.2 are consistent.*

The next two theorems show that if Model (3.1) is discontinuous at τ_j^0 for some $j = 1, \cdots, l^0$, then the threshold estimate $\hat{\tau}_j$ converges to the true thresholds τ_j^0 at the rate of $O_p(ln^2 n/n)$, and the least squares estimates of $\tilde{\beta}_j^0$ and σ_0^2 based on the estimated thresholds are asymptotically normal.

Theorem 3.4 *Suppose for the segmented linear regression model (3.1) that Assumptions 3.0, 3.2 and 3.3 are satisfied. For any* $j \in \{1, \cdots, l^0\}$ *such that* $P(\mathrm{x}_1'(\tilde{\beta}_{j+1}^0 - \tilde{\beta}_j^0) \neq 0 | x_d = \tau_j^0) > 0$,

$$\hat{\tau}_j - \tau_j^0 = O_p(\frac{\ln^2 n}{n}).$$

Let $\hat{\tilde{\beta}}_j$ and $\hat{\sigma}^2$ be the least squares estimates of $\tilde{\beta}_j^0$ and σ_0^2 based on the estimates $\hat{l}$ and $\hat{\tau}_j$'s as defined in Section 2.2, $j = 1, \ldots, l^0 + 1$.

Theorem 3.5 *Suppose for the segmented linear regression model (3.1) that Assumptions 3.0, 3.2 and 3.3 are satisfied. If* $P(\mathrm{x}_1'(\tilde{\beta}_{j+1}^0 - \tilde{\beta}_j^0) \neq 0 | x_d = \tau_j^0) > 0$ *for all* $j = 1, \cdots, l^0$, *then* $\sqrt{n}(\hat{\tilde{\beta}}_j - \tilde{\beta}_j^0)$ *and* $\sqrt{n}[\hat{\sigma}^2 - \sigma_0^2]$ *converge in distribution to normal distributions with finite variances,* $j = 1, \ldots, l^0 + 1$.

Remark The asymptotic variances can be computed by first treating l^0 and τ_j^0, $(j = 1, \ldots, l^0)$, as known so that the usual "estimates" of the variances of the estimates of the regression coefficients and residual variance can then be written down explicitly by substituting $\hat{l}$ and $\hat{\tau}_j$ for l^0 and τ_j^0, $(j = 1, \ldots, l^0)$, in these variance "estimates". For example, the asymptotic covariance matrix for $\hat{\hat{\beta}}_j$ is $\sigma_0^2 G_j^{-1}$, where $G_j = E[\mathbf{x}_1\mathbf{x}_1' \mathbf{1}_{(x_{1d} \in (\tau_{j-1}^0, \tau_j^0])}]$.

The proof of Theorem 3.1 is motivated by the following idea. If the model is overfitted $(l^0 < l \leq L)$, the reduction in the mean square error will be bounded in probability by a positive sequence tending to zero. In fact, this turns out to be $O_p(\ln^2 n/n)$. On the other hand, if the model is underfitted $(l < l^0)$, the inflation in the mean square error will be of order $O_p(1)$. Hence, by setting the penalty term in MIC equal to a quantity of order bigger than $O_p(\ln^2 n/n)$ but still tending to 0, we can avoid both overfitting and underfitting. This idea is formulated in a series of lemmas.

The result of Lemma 3.1 is a consequence of the local exponential boundedness assumption, which gives the added flexibility of modeling with non-Gaussian noises. Using the properties of the hat matrix $H_n(x_{sd}, x_{td})$, Lemma 3.2 establishes a uniform bound of $T_n(\alpha, \eta)$ for all $\alpha < \eta$. With this lemma, we show in Proposition 3.1 that the mean squared residuals differs from the mean squared pure errors only by $O_p(\ln^2 n/n)$, which in sequel motivates the choice of the penalty term in our MIC. Given Lemma 3.2 and Proposition 3.1, the results of Lemmas 3.3 and 3.4 are more or less expected.

Lemma 3.1 *Let $Z_1, \ldots, Z_k$ be i.i.d. locally exponentially bounded random variables, i.e., $E(e^{uZ_1}) \leq e^{c_0 u^2}$ for $|u| \leq T_0$, where T_0 and $c_0 \in (0, \infty)$. Let $S_k = \sum_{i=1}^k a_i Z_i$, where the a_i's are* *constants. Then for any $t_0 > 0$ satisfying $|t_0 a_i| \le T_0$, $i \le k$,*

$$P\{|S_k| \ge x\} \le 2e^{-t_0 x + c_0 t_0^2 \sum_{i=1}^k a_i^2}. \tag{3.3}$$

Proof It follows from Markov's inequality that for the hypothesized t_0,

$$P\{S_k \ge x\} = P\{e^{t_0 S_k} \ge e^{t_0 x}\} \le e^{-t_0 x} E(e^{t_0 S_k}) = e^{-t_0 x} E(e^{t_0 \sum_{i=1}^k a_i Z_i}) \le e^{-t_0 x} e^{c_0 t_0^2 \sum_{i=1}^k a_i^2},$$

and to conclude the proof of (3.3),

$$P\{S_k \le -x\} = P\{-S_k \ge x\} \le e^{-t_0 x} e^{c_0 t_0^2 \sum_{i=1}^k a_i^2}. \qquad \P$$

Lemma 3.2 *Assume for the segmented linear regression model (3.1) that Assumption 3.0 is satisfied. Let $T_n(\alpha, \eta), -\infty \le \alpha < \eta \le \infty$, be defined as in the beginning of this section. Then*

$$P\{\sup_{\alpha<\eta} T_n(\alpha, \eta) \ge \frac{9p_0^3}{T_0^2} \ln^2 n\} \to 0, \qquad \text{as } n \to 0, \tag{3.4}$$

where p_0 is the true order of the model and T_0 is the constant associated with the local exponential boundedness condition for the $\{\epsilon_t\}$.

Proof Conditioning on X_n, we have

$$P\{\sup_{\alpha<\eta} T_n(\alpha, \eta) \ge \frac{9p_0^3}{T_0^2} \ln^2 n \mid X_n\} = P\{\max_{x_{sd}<x_{td}} \tilde{\epsilon}_n' H_n(x_{sd}, x_{td}) \tilde{\epsilon}_n \ge \frac{9p_0^3}{T_0^2} \ln^2 n \mid X_n\}$$

$$\le \sum_{x_{sd}<x_{td}} P\{\tilde{\epsilon}_n' H_n(x_{sd}, x_{td}) \tilde{\epsilon}_n \ge \frac{9p_0^3}{T_0^2} \ln^2 n \mid X_n\}.$$

Since $H_n(x_{sd}, x_{td})$ is nonnegative definite and idempotent, it can be decomposed as $H_n(x_{sd}, x_{td}) = W'\Lambda W$, where W is orthogonal and $\Lambda = diag(1, \cdots, 1, 0, \cdots, 0)$ with $p := rank(H_n(x_{sd}, x_{td})) = rank(\Lambda) \le p_0$. Set $Q = (I_p, \mathbf{0})W$. Then Q has full row rank p. Let $Q' = (\mathbf{q}_1, \cdots, \mathbf{q}_p)$ and $U_l = \mathbf{q}_l' \tilde{\epsilon}_n$, $l = 1, \ldots, p$. Then

$$\tilde{\epsilon}_n' H_n(x_{sd}, x_{td}) \tilde{\epsilon}_n = \tilde{\epsilon}_n' Q' Q \tilde{\epsilon}_n = \sum_{l=1}^p U_l^2.$$

Since $p \leq p_0$ and

$$\begin{aligned}
&P\{\sum_{l=1}^{p} U_l^2 \geq \frac{9p_0^3}{T_0^2}\ln^2 n|X_n\} \\
\leq &P\{\sum_{l=1}^{p} U_l^2 \geq p\frac{9p_0^2}{T_0^2}\ln^2 n|X_n\} \\
\leq &P\{U_l^2 \geq \frac{9p_0^2}{T_0^2}\ln^2 n \ for\ some\ l|X_n\} \\
\leq &\sum_{l=1}^{p} P\{U_l^2 \geq \frac{9p_0^2}{T_0^2}\ln^2 n|X_n\},
\end{aligned}$$

it suffices to show, for any l, that

$$\sum_{x_{sd}<x_{td}} P\{U_l^2 \geq \frac{9p_0^2}{T_0^2}\ln^2 n \mid X_n\} \to 0, \qquad as\ n \to 0.$$

Noting that $p = trace(H_n(x_{sd}, x_{td})) = \sum_{l=1}^{p} \parallel \mathbf{q}_l \parallel^2$, we have $\parallel \mathbf{q}_l \parallel^2 = \mathbf{q}_l'\mathbf{q}_l \leq p \leq p_0$, $l = 1, \ldots, p$. By Lemma 3.1, with $t_0 = T_0/p_0$ we have

$$\sum_{x_{sd}<x_{td}} P\{|U_l| \geq 3p_0 \ln n/T_0 \mid X_n\} \leq \sum_{x_{sd}<x_{td}} 2\exp(-\frac{T_0}{p_0}\cdot\frac{3p_0}{T_0}\ln n)\exp(c_0(T_0/p_0)^2 p_0)$$

$$\leq n(n-1)/n^3 \exp(c_0 T_0^2/p_0) \to 0,$$

as $n \to \infty$, where c_0 is the constant specified in Lemma 3.1. Finally, by appealing to the dominated convergence theorem we obtain the desired result without conditioning. ¶

Proposition 3.1 *Consider the segmented regression model 3.1.*

(i) For any j and $(\alpha, \eta] \subset (\tau_{j-1}^0, \tau_j^0]$,

$$S_n(\alpha, \eta) = \tilde{\epsilon}_n'(\alpha, \eta)\tilde{\epsilon}_n(\alpha, \eta) - T_n(\alpha, \eta).$$

(ii) Suppose Assumption 3.0 is satisfied. Let $m \geq 1$. Then uniformly for all $(a_1, \cdots, a_m)$ such that $-\infty < a_1 < \cdots < a_m < \infty$,

$$S_n(\xi_1, \cdots, \xi_{m+l^0}) = \sum_{i=1}^{m+l^0+1} S_n(\xi_{i-1}, \xi_i) = \tilde{\epsilon}_n'\tilde{\epsilon}_n + O_p(\ln^2 n),$$

where $\xi_0 = -\infty$, $\xi_{m+l^0+1} = \infty$, and $\{\xi_1, \cdots, \xi_{m+l^0}\}$ is the set $\{\tau_1^0, \cdots, \tau_{l^0}^0, a_1, \cdots, a_m\}$ after ordering its elements.

Proof: (i) Observe that

$$\begin{aligned}
S_n(\alpha,\eta) =& Y_n'(I_n(\alpha,\eta) - H_n(\alpha,\eta)Y_n \\
=& (X_n(\alpha,\eta)\tilde{\beta}_j^0 + \tilde{\epsilon}_n(\alpha,\eta))'(X_n(\alpha,\eta)\tilde{\beta}_j^0 + \tilde{\epsilon}_n(\alpha,\eta)) \\
& - (X_n(\alpha,\eta)\tilde{\beta}_j^0 + \tilde{\epsilon}_n(\alpha,\eta))'H_n(\alpha,\eta)(X_n(\alpha,\eta)\tilde{\beta}_j^0 + \tilde{\epsilon}_n(\alpha,\eta)) \\
=& \tilde{\beta}_j^{0\prime} X_n'(\alpha,\eta)X_n(\alpha,\eta)\tilde{\beta}_j^0 + 2\tilde{\epsilon}_n'(\alpha,\eta)X_n(\alpha,\eta)\tilde{\beta}_j^0 + \tilde{\epsilon}_n'(\alpha,\eta)\tilde{\epsilon}_n(\alpha,\eta) \\
& - [\tilde{\beta}_j^{0\prime} X_n'(\alpha,\eta)H_n(\alpha,\eta)X_n(\alpha,\eta)\tilde{\beta}_j^0 \\
& + 2\tilde{\epsilon}_n'(\alpha,\eta)H_n(\alpha,\eta)X_n(\alpha,\eta)\tilde{\beta}_j^0 + \tilde{\epsilon}_n'(\alpha,\eta)H_n(\alpha,\eta)\tilde{\epsilon}_n(\alpha,\eta)].
\end{aligned}$$

Noting that $H_n(\alpha,\eta)$ is idempotent and

$$X_n'(\alpha,\eta)H_n(\alpha,\eta)X_n(\alpha,\eta) = X_n'(\alpha,\eta)X_n(\alpha,\eta),$$

we have

$$\begin{aligned}
&(X_n(\alpha,\eta) - H_n(\alpha,\eta)X_n(\alpha,\eta))'(X_n(\alpha,\eta) - H_n(\alpha,\eta)X_n(\alpha,\eta)) \\
=& X_n'(\alpha,\eta)(I_n(\alpha,\eta) - H_n(\alpha,\eta))X_n(\alpha,\eta) \\
=& X_n'(\alpha,\eta)X_n(\alpha,\eta) - X_n'(\alpha,\eta)X_n(\alpha,\eta) = \mathbf{0}
\end{aligned}$$

and hence $X_n(\alpha,\eta) = H_n(\alpha,\eta)X_n(\alpha,\eta)$. Therefore

$$\begin{aligned}
S_n(\alpha,\eta) =& \tilde{\epsilon}_n'(\alpha,\eta)\tilde{\epsilon}_n(\alpha,\eta) - \tilde{\epsilon}_n'(\alpha,\eta)H_n(\alpha,\eta)\tilde{\epsilon}_n(\alpha,\eta) \\
=& \tilde{\epsilon}_n'(\alpha,\eta)\tilde{\epsilon}_n(\alpha,\eta) - T_n(\alpha,\eta).
\end{aligned}$$

(ii) By (i),

$$\begin{aligned}
& S_n(\xi_1, \cdots, \xi_{m+l^0}) \\
=& \sum_{i=1}^{m+l^0+1} S_n(\xi_{i-1}, \xi_i) \\
=& \sum_{i=1}^{m+l^0+1} [\tilde{\epsilon}_n'(\xi_{i-1}, \xi_i)\tilde{\epsilon}_n(\xi_{i-1}, \xi_i) - T_n(\xi_{i-1}, \xi_i)] \\
=& \tilde{\epsilon}_n'\tilde{\epsilon}_n - \sum_{i=1}^{m+l^0+1} T_n(\xi_{i-1}, \xi_i).
\end{aligned}$$

Note that each of $(\xi_{i-1}, \xi_i]$ is contained in one of $(\tau_{j-1}^0, \tau_j^0]$, $j = 1, \cdots, l^0 + 1$. By Lemma 3.2, $\sum_{i=1}^{m+l^0+1} T_n(\xi_{i-1}, \xi_i) \le (m + l^0 + 1) \sup_{\alpha<\eta} T_n(\alpha < \eta) = O_p(\ln^2 n)$. ¶

Lemma 3.3 *Under the condition of Theorem 3.1, there exists* $\delta \in (0, \min_{1\le j\le l^0}(\tau_{j+1}^0 - \tau_j^0)/2)$ *such that for* $r = 1, \ldots, l^0$,

$$[S_n(\tau_r^0 - \delta, \tau_r^0 + \delta) - S_n(\tau_r^0 - \delta, \tau_r) - S_n(\tau_r^0, \tau_r^0 + \delta)]/n \xrightarrow{a.s.} C_r \tag{3.5}$$

for some $C_r > 0$ *as* $n \to \infty$.

Proof It suffices to prove the result when $l^0 = 1$. For notational simplicity, we omit the subscripts and superscripts 0 in this proof. For the δ in *Assumption 3.1*, let $X_1^* = X_n(\tau_1 - \delta, \tau_1)$, $X_2^* = X_n(\tau_1, \tau_1 + \delta)$, $X^* = X_n(\tau_1 - \delta, \tau_1 + \delta) = X_1^* + X_2^*$, $\tilde{\epsilon}_1^* = \tilde{\epsilon}_n(\tau_1 - \delta, \tau_1)$, $\tilde{\epsilon}_2^* = \tilde{\epsilon}_n(\tau_1, \tau_1 + \delta)$, $\tilde{\epsilon}^* = \tilde{\epsilon}_1^* + \tilde{\epsilon}_2^*$ and $\hat{\tilde{\beta}} = (X^{*\prime}X^*)^- X^{*\prime}Y_n$. As in ordinary regression, we have

$$\begin{aligned}
&S_n(\tau_1 - \delta, \tau_1 + \delta)\\
=&\|X_1^*\tilde{\beta}_1 + X_2^*\tilde{\beta}_2 + \tilde{\epsilon}^* - X^*\hat{\tilde{\beta}}\|^2\\
=&\|X_1^*(\tilde{\beta}_1 - \hat{\tilde{\beta}}) + X_2^*(\tilde{\beta}_2 - \hat{\tilde{\beta}}) + \tilde{\epsilon}^*\|^2\\
=&\|X_1^*(\tilde{\beta}_1 - \hat{\tilde{\beta}})\|^2 + \|X_2^*(\tilde{\beta}_2 - \hat{\tilde{\beta}})\|^2 + \|\tilde{\epsilon}^*\|^2 + 2\tilde{\epsilon}^{*\prime}X_1^*(\tilde{\beta}_1 - \hat{\tilde{\beta}}) + 2\tilde{\epsilon}^{*\prime}X_2^*(\tilde{\beta}_2 - \hat{\tilde{\beta}}).
\end{aligned}$$

It then follows from the strong law of large numbers for stationary ergodic stochastic processes that as $n \to \infty$,

$$\frac{1}{n}X^{*\prime}X^* = \frac{1}{n}\sum_{t=1}^{n} \mathbf{x}_t\mathbf{x}_t'\mathbf{1}_{(x_{td}\in(\tau_1-\delta,\tau_1+\delta])} \xrightarrow{a.s.} E\{\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_1-\delta,\tau_1+\delta])}\} > 0,$$

$$\frac{1}{n}X_j^{*\prime}X_j^* \xrightarrow{a.s.} \begin{cases} E\{\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_1-\delta,\tau_1])}\} > 0, & \text{if j=1,}\\ E\{\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_1,\tau_1+\delta])}\} > 0, & \text{if j=2,} \end{cases}$$

and

$$\frac{1}{n}X^{*\prime}Y_n \xrightarrow{a.s.} E\{y_1\mathbf{x}_1\mathbf{1}_{(x_{1d}\in(\tau_1-\delta,\tau_1+\delta])}\}.$$

Therefore,

$$\hat{\tilde{\beta}} \xrightarrow{a.s.} \{E\{\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_1-\delta,\tau_1+\delta])}\}\}^{-1}E\{y_1\mathbf{x}_1\mathbf{1}_{(x_{1d}\in(\tau_1-\delta,\tau_1+\delta])}\} =: \tilde{\beta}^*.$$

Similarly, it can be shown that

$$\frac{1}{n}||X_j^*(\tilde{\beta}_j - \hat{\tilde{\beta}})||^2 \xrightarrow{a.s.} \begin{cases} (\tilde{\beta}_1 - \tilde{\beta}^*)'E(\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_1-\delta,\tau_1])})\cdot(\tilde{\beta}_1 - \tilde{\beta}^*), & \text{if} \quad \text{j=1,} \\ (\tilde{\beta}_2 - \tilde{\beta}^*)'E(\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_1,\tau_1+\delta])})\cdot(\tilde{\beta}_2 - \tilde{\beta}^*), & \text{if} \quad \text{j=2,} \end{cases}$$

$$\frac{1}{n}\tilde{\epsilon}^{*\prime}X_j^*(\tilde{\beta}_j - \hat{\tilde{\beta}}) \xrightarrow{a.s.} 0, \; for \; j = 1, 2,$$

and

$$\frac{1}{n}||\tilde{\epsilon}^*||^2 \xrightarrow{a.s.} \sigma^2 P\{x_{1d} \in (\tau_1 - \delta, \tau_1 + \delta]\}.$$

Thus as $n \to \infty$, $\frac{1}{n}S_n(\tau_1 - \delta, \tau_1 + \delta)$ has a finite limit, this limit being given by

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{n}S_n(\tau_1 - \delta, \tau_1 + \delta) \\ =&(\tilde{\beta}_1 - \tilde{\beta}^*)'E(\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_1-\delta,\tau_1])})\cdot(\tilde{\beta}_1 - \tilde{\beta}^*) + (\tilde{\beta}_2 - \tilde{\beta}^*)'E(\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_1,\tau_1+\delta])})\cdot(\tilde{\beta}_2 - \tilde{\beta}^*) \\ &+ \sigma^2 P\{x_{td} \in (\tau_1 - \delta, \tau_1 + \delta]\}. \end{aligned}$$

It remains to show that $\frac{1}{n}S_n(\tau_1 - \delta, \tau_1)$ and $\frac{1}{n}S_n(\tau_1, \tau_1 + \delta)$ converge to $\sigma^2 P\{x_{1d} \in (\tau_1 - \delta, \tau_1]\}$ and $\sigma^2 P\{x_{1d} \in (\tau_1, \tau_1 + \delta]\}$, respectively, and either $(\tilde{\beta}_1 - \tilde{\beta}^*)'E(\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_1-\delta,\tau_1])})(\tilde{\beta}_1 - \tilde{\beta}^*) > 0$ or $(\tilde{\beta}_2 - \tilde{\beta}^*)'E(\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_1,\tau_1+\delta])})(\tilde{\beta}_2 - \tilde{\beta}^*) > 0$. The latter is a direct consequence of the assumed conditions while the former can be shown again by the law of large numbers. To this end, we first write $S_n(\tau_1 - \delta, \tau_1)$ in the following form (bearing in mind that l^0 is assumed to be 1 in the proof),

$$S_n(\tau_1 - \delta, \tau_1) = \tilde{\epsilon}_1^{*\prime}\tilde{\epsilon}_1^* - T_n(\tau_1 - \delta, \tau_1)$$

using Proposition 3.1 (i). By the strong law of large numbers,

$$\begin{aligned} &\frac{1}{n}\tilde{\epsilon}_1^{*\prime}\tilde{\epsilon}_1^* \xrightarrow{a.s.} E[\epsilon_1^2\mathbf{1}_{(x_{1d}\in(\tau_1-\delta,\tau_1])}] = \sigma^2 P\{x_{1d} \in (\tau_1 - \delta, \tau_1]\}, \\ &\frac{1}{n}\tilde{\epsilon}_1^* X_1^* \xrightarrow{a.s.} E[\epsilon_1\mathbf{x}_1\mathbf{1}_{(x_{1d}\in(\tau_1-\delta,\tau_1])}] = \mathbf{0}, \end{aligned}$$

and $W = \lim_{n\to\infty} X_1^{*\prime} X_1^*$ is positive definite under the assumption. Therefore,

$$T_n(\tau_1 - \delta, \tau_1) = (\frac{1}{n}\tilde{\epsilon}_1^{*\prime} X_1^*)(\frac{1}{n} X_1^{*\prime} X_1^*)^-(\frac{1}{n} X_1^{*\prime}\tilde{\epsilon}_1^*) \xrightarrow{a.s.} \mathbf{0} W^{-1}\mathbf{0} = 0$$

and hence $\frac{1}{n}S_n(\tau_1 - \delta, \tau_1) \xrightarrow{a.s.} \sigma^2 P\{x_{1d} \in (\tau_1 - \delta, \tau_1]\}$. The same argument can also be used to show that $\frac{1}{n}S_n(\tau_1, \tau_1 + \delta) \xrightarrow{a.s.} \sigma^2 P\{x_{1d} \in (\tau_1, \tau_1 + \delta]\}$. This completes the proof. ¶

Lemma 3.4 *Under the condition of Theorem 3.1, we have*

(i) for every $l < l^0$, $P\{\hat{\sigma}_l^2 > \sigma_0^2 + C\} \to 1$, as $n \to \infty$ for some $C > 0$, and

(ii) for every l such that $l^0 \le l \le L$, where L is an upper bound of l^0,

$$0 \le \frac{1}{n}\tilde{\epsilon}_n' \tilde{\epsilon}_n - \hat{\sigma}_l^2 = O_p(ln^2(n)/n), \tag{3.6}$$

where $\hat{\sigma}_l^2 = \frac{1}{n}S_n(\hat{\tau}_1, \ldots, \hat{\tau}_l)$ is the estimated σ_0^2 when the number of true thresholds is assumed to be l.

Proof (i) Since $l < l^0$, for the $\delta \in (0, \min_{1\le j\le l^0}(\tau_{j+1}^0 - \tau_j^0)/2)$ in *Assumption 3.1*, there exists $1 \le r \le l^0$, such that $(\hat{\tau}_1, \ldots, \hat{\tau}_l) \in A_r := \{(\tau_1, \ldots, \tau_l) : |\tau_s - \tau_r^0| > \delta, \ \textit{for all } s = 1, \ldots, l\}$. Hence, if we can show that for each r, $1 \le r < l^0$, with probability approaching 1,

$$\min_{(\tau_1,\cdots,\tau_l)\in A_r} S_n(\tau_1, \cdots, \tau_l)/n > \sigma_0^2 + C_r,$$

for some $C_r > 0$, then by choosing $C := \min_{1\le r\le l^0}\{C_r\}$, we will have proved the desired result.

For any $(\tau_1, \cdots, \tau_l) \in A_r$, let $\xi_1 \le \cdots \le \xi_{l+l^0+1}$ be the ordered set $\{\tau_1, \ldots, \tau_l, \tau_1^0, \ldots, \tau_{r-1}^0, \tau_r^0 - \delta, \tau_r^0 + \delta, \tau_{r+1}^0, \ldots, \tau_{l^0}^0\}$ and let $\xi_0 = -\infty$, $\xi_{l+l^0+2} = \infty$. Then it follows from Proposition

3.1 (ii) that uniformly in A_r,

$$\begin{aligned}
&\frac{1}{n}S_n(\tau_1,\cdots,\tau_l)\\
\geq&\frac{1}{n}S_n(\xi_1,\cdots,\xi_{l+l^0+1})\\
=&\frac{1}{n}\sum_{j=1}^{l+l^0+2}S_n(\xi_{j-1},\xi_j)\\
=&\frac{1}{n}[\sum_{\{j:\ \xi_j\neq\tau_r^0+\delta\}}S_n(\xi_{j-1},\xi_j)+S_n(\tau_r^0-\delta,\tau_r^0)+S_n(\tau_r^0,\tau_r^0+\delta)]\\
&+\frac{1}{n}[S_n(\tau_r^0-\delta,\tau_r^0+\delta)-S_n(\tau_r^0-\delta,\tau_r^0)-S_n(\tau_r^0,\tau_r^0+\delta)]\\
=&\frac{1}{n}\tilde{\epsilon}_n'\tilde{\epsilon}_n+O_p(\ln^2(n)/n)+\frac{1}{n}[S_n(\tau_r^0-\delta,\tau_r^0+\delta)-S_n(\tau_r^0-\delta,\tau_r^0)-S_n(\tau_r^0,\tau_r^0+\delta)].
\end{aligned}\tag{3.7}$$

By the strong law of large numbers the first term on the RHS is $\sigma_0^2+o(1)$ a.s.. By Lemma 3.3, the third term on the RHS is $C_r+o_p(1)$ a.s.. Thus

$$\frac{1}{n}S_n(\tau_1,\cdots,\tau_l)\geq\sigma_0^2+C_r+o_p(1),$$

where C_r is defined in (3.5).

(ii) Let $\xi_1\leq\cdots\leq\xi_{l+l^0}$ be the ordered set, $\{\hat{\tau}_1,\cdots,\hat{\tau}_l,\tau_1^0,\cdots,\tau_{l^0}^0\}$, $\xi_0=\tau_0^0=-\infty$ and $\xi_{l+l^0+1}=\tau_{l^0+1}^0=\infty$. Since $l\geq l^0$, by Proposition 3.1 (ii) again,

$$\begin{aligned}
\tilde{\epsilon}_n'\tilde{\epsilon}_n\geq&S_n(\tau_1^0,\cdots,\tau_{l^0}^0)\\
\geq&S_n(\hat{\tau}_1,\cdots,\hat{\tau}_l)\\
=&n\hat{\sigma}_l^2\\
\geq&S_n(\xi_1,\cdots,\xi_{l+l^0})\\
=&\sum_{j=1}^{l+l^0+1}S_n(\xi_{j-1},\xi_j)\\
=&\tilde{\epsilon}_n'\tilde{\epsilon}_n+O_p(ln^2(n)).
\end{aligned}$$

This proves (ii). ¶

Proof of Theorem 3.1 By Lemma 3.4 (i), for $l<l^0$ and sufficiently large n, there exists

$C > 0$ such that

$$MIC(l) = \ln(\hat{\sigma}_l^2) + p^*(\ln n)^{2+\delta}/n \geq \ln(\sigma_0^2 + C/2) \geq \ln(\sigma_0^2) + \ln(1 + C/(2\sigma_0^2))$$

with probability approaching 1. By Lemma 3.4 (ii), for $l \geq l^0$,

$$MIC(l) = \ln(\hat{\sigma}_l^2) + p^*(\ln n)^{2+\delta}/n \xrightarrow{p} \ln \sigma_l^2.$$

Thus, $P\{\hat{l} \geq l^0\} \to 1$ as $n \to \infty$. By Lemma 3.4 (ii) and the strong law of large numbers, for $l^0 < l \leq L$,

$$0 \geq [\hat{\sigma}_l^2 - \frac{1}{n}\tilde{\epsilon}_n'\tilde{\epsilon}_n] - [\hat{\sigma}_{l^0}^2 - \frac{1}{n}\tilde{\epsilon}_n'\tilde{\epsilon}_n] = O_p(\ln^2 n/n),$$

and

$$[\hat{\sigma}_{l^0}^2 - \sigma_0^2] = [\hat{\sigma}_{l^0}^2 - \frac{1}{n}\tilde{\epsilon}_n'\tilde{\epsilon}_n] + [\frac{1}{n}\tilde{\epsilon}_n'\tilde{\epsilon}_n - \sigma_0^2] = O_p(\ln^2 n/n) + o_p(1) = o_p(1).$$

Hence $0 \leq (\hat{\sigma}_{l^0}^2 - \hat{\sigma}_l^2)/\hat{\sigma}_{l^0}^2 = O_p(\ln^2 n/n)$. Note that for $0 \leq x < 1/2, \ln(1-x) \geq -2x$. Therefore,

$$\begin{aligned}
MIC(l) - MIC(l^0) &= \ln(\hat{\sigma}_l^2) - \ln(\hat{\sigma}_{l^0}^2) + c_0(l - l^0)(\ln n)^{2+\delta_0}/n \\
&= \ln(1 - (\hat{\sigma}_{l^0}^2 - \hat{\sigma}_l^2)/\hat{\sigma}_{l^0}^2) + c_0(l - l^0)(\ln n)^{2+\delta_0}/n \\
&\geq - 2O_p(\ln^2(n)/n) + c_0(l - l^0)(\ln n)^{2+\delta_0}/n \\
&> 0
\end{aligned}$$

for sufficiently large n. Whence $\hat{l} \xrightarrow{p} l^0$ as $n \to \infty$. ¶

Remark: From the proof of Theorem 3.1 it can be seen that if the term $c_0 l(\ln n)^{2+\delta_0}/n$ is replaced by $l \cdot cn^{\alpha-1}$, where $\alpha \in (0,1)$ and c is a constant, the model selection procedure is still consistent. In fact, such a penalty is proposed by Yao (1989) for a one-dimensional piecewise constant model.

Remark If the assumed δ in *Assumption 3.1* is replaced by assumed sequences $\{a_j\}$, $\{b_j\}$ such that $-\infty \leq a_1 < \tau_1^0 < b_1 \leq \cdots \leq a_{l^0} < \tau_{l^0}^0 < b_{l^0} \leq \infty$, and such that both $E\{\mathbf{x}_1\mathbf{x}_1' 1_{(x_{1d} \in (a_j, \tau_j^0])}\}$

and $E\{\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_j^0,b_j])}\}$ are positive definite for $j = 1, \ldots, l^0$, then the conclusion of Lemma 3.3 still holds with δ replaced by a_j and b_j, respectively. Therefore, the conclusion of Theorem 3.1 still holds.

To prove Theorem 3.2, we need the following lemma.

Lemma 3.5 *Under the assumptions of Theorem 3.2, for any sufficiently small $\delta \in (0, \min_{1\le j\le l^0}(\tau_{j+1}^0 - \tau_j^0)/2)$, there exists a constant $C_r > 0$ such that*

$$\frac{1}{n}[S_n(\tau_r^0 - \delta, \tau_r^0 + \delta) - S_n(\tau_r^0 - \delta, \tau_r^0) - S_n(\tau_r^0, \tau_r^0 + \delta)] \xrightarrow{p} C_r, \ \textit{as } n \to \infty,$$

where $r = 1, \cdots, l^0$.

Proof It suffices to prove the result for the case when $l^0 = 1$. For any small $\delta > 0$, all the arguments in the proof of Lemma 3.3 apply, under *Assumption 3.2.* Hence the result holds. ¶

Remark: Although the proofs of Lemma 3.3 and Lemma 3.5 are essentially the same, the assumptions, and hence the conclusions of these lemmas are different. In Lemma 3.3 C_r is fixed for the existing δ. While Lemma 3.5 implies that for any sequence of $\{\delta_m\}$ such that $\delta_m > 0$ and $\delta_m \to 0$ as $m \to \infty$, there exist $\{C_r(m)\}$ such that the conclusion of Lemma 3.5 holds for all m.

Proof of Theorem 3.2 By Theorem 3.1, the problem can be restricted to $\{\hat{l} = l^0\}$. For any sufficiently small $\delta' > 0$, substituting δ' for the δ in (3.7) in the proof of Lemma 3.4 (i), we have the following inequality

$$
\begin{aligned}
&\frac{1}{n}S_n(\tau_1,\cdots,\tau_{l^0})\\
\geq&\frac{1}{n}\tilde{\epsilon}_n'\tilde{\epsilon}_n+O_p(ln^2(n)/n)\\
&+\frac{1}{n}[S_n(\tau_r^0-\delta',\tau_r^0+\delta')-S_n(\tau_r^0-\delta',\tau_r^0)-S_n(\tau_r^0,\tau_r^0+\delta')],
\end{aligned}
$$

uniformly in $(\tau_1,\cdots,\tau_{l^0})\in A_r:=\{(\tau_1,\cdots,\tau_{l^0}):|\tau_s-\tau_r^0|>\delta',1\leq s\leq l^0\}$. By Lemma 3.5, the last term on the RHS converges to a positive C_r. For sufficiently large n, this C_r will dominate the term $O_p(\ln^2 n/n)$. Thus, uniformly in A_r, $r=1,\ldots,l^0$, and with probability tending to 1,

$$\frac{1}{n}S_n(\tau_1,\cdots,\tau_{l^0})>\frac{1}{n}\tilde{\epsilon}_n'\tilde{\epsilon}_n+\frac{C_r}{2}.$$

This implies that with probability approaching 1 no τ in A_r is qualified as a candidate for the role of $\hat{\tau}$, where $\hat{\tau}=(\hat{\tau}_1,\cdots,\hat{\tau}_{l^0})$. In other words, $P(\hat{\tau}\in A_r^c)\to 1$ as $n\to\infty$. Since this is true for all r, $P(\hat{\tau}\in\bigcap_{r=1}^{l^0}A_r^c)\to 1,\ as\ n\to\infty$. Note that for $\delta'\leq\min_{0\leq i\leq l^0}\{(\tau_{i+1}^0-\tau_i^0)/2\}$,

$$\bigcap_{r=1}^{l^0}\{|\hat{\tau}_r-\tau_r^0|\leq\delta'\}=\bigcap_{r=1}^{l^0}\{|\hat{\tau}_{i_r}-\tau_r^0|\leq\delta',for\ some\ 1\leq i_r\leq l^0\}=\{\hat{\tau}\in\bigcap_{r=1}^{l^0}A_r^c\}.$$

Thus we have,

$$P(|\hat{\tau}_r-\tau_r^0|\leq\delta'\ for\ r=1,\ldots,l^0)=P(\hat{\tau}\in\bigcap_{r=1}^{l^0}A_r^c)\to 1,\ as\ n\to\infty,$$

which completes the proof. ¶

The proof of Theorem 3.3 requires a series of preliminary results. The key step is to establish Lemma 3.6 which implies the estimation errors of the regression coefficients are controlled by the estimation errors of the thresholds.

Proposition 3.2 Let $\{x_n\}$ be a sequence of random variables. If $x_n=o_p(1)$, then there exists a positive sequence $\{a_n\}$, such that $a_n\to 0$ as $n\to\infty$ and $x_n=O_p(a_n)$.

Proof Let $\epsilon_k = \delta_k = 1/2^k$, $k = 1, 2, \cdots$. Since $x_n = o_p(1)$, for ϵ_1 and δ_1, there exists $N_1 > 0$ such that for all $n \geq N_1$

$$P(|x_n| > \delta_1) < \epsilon_1.$$

And for each pair of ϵ_k and δ_k, there exists $N_k \geq N_{k-1}$ such that for all $n \geq N_k$,

$$P(|x_n| > \delta_k) < \epsilon_k.$$

Let $a_n = 1$ if $n < N_1$ and $a_n = \delta_k$ if $N_k \leq n < N_{k+1}$, $k = 1, 2, \cdots$. Then $a_n \to 0$ as $n \to \infty$. Also, for any $\epsilon > 0$, there exists k_0 such that $0 < \epsilon_{k_0} < \epsilon$. Thus for any $n \geq N_{k_0}$, $N_k \leq n < N_{k+1}$ for some $k \geq k_0$, and

$$P(|x_n| > a_n) = P(|x_n| > \delta_k) < \epsilon_k \leq \epsilon_{k_0} < \epsilon.$$

Again by $x_n = o_p(1)$, there exists $M > 1$ such that

$$P(|x_n| > M) < \epsilon$$

for all $n < N_{k_0}$. This completes the proof. ¶

Lemma 3.6 *Let $R_j = (\tau_{j-1}^0, \tau_j^0]$, $\hat{R}_j = (\hat{\tau}_{j-1}, \hat{\tau}_j]$, $\tau_0^0 = \hat{\tau}_0 = -\infty$, $\tau_{l^0+1}^0 = \hat{\tau}_{l^0+1} = \infty$, and $\Delta_{n,j} = |\hat{\tau}_j - \tau_j^0| = O_p(a_n)$, $j = 1, \cdots, l^0+1$, where $\{a_n\}$ is a sequence of positive numbers. Suppose that $\{(z_t, x_{td})\}$ is a strictly stationary and ergodic sequence and that the marginal cdf, F_d, of x_{1d} satisfies the Lipschitz condition, $|F_d(x') - F_d(x'')| \leq C|x' - x''|$, for some constant C in a small neighborhood of $x_{1d} = \tau_j^0$ for every j. If for some $u > 1$, $E|z_1|^u < \infty$, then*

$$\frac{1}{n}\sum_{t=1}^{n} z_t(\mathbf{1}_{(x_{td}\in\hat{R}_j)} - \mathbf{1}_{(x_{td}\in R_j)}) = O_p((a_n)^{1/v})$$

where $1/v = 1 - 1/u$.

Proof It suffices to show that $\frac{1}{n}\sum_{t=1}^{n}|z_t||\mathbf{1}_{(x_{td}\in\hat{R}_j)} - \mathbf{1}_{(x_{td}\in R_j)}| = O_p((a_n)^{1/v})$. Since, for every $j = 1,\cdots,l^0$,

$$|\mathbf{1}_{(x_{td}\in\hat{R}_j)} - \mathbf{1}_{(x_{td}\in R_j)}| = \mathbf{1}_{(x_{td}\in(\hat{R}_j\cap R_j^c)\cup(\hat{R}_j^c\cap R_j))} \leq \mathbf{1}_{(|x_{td}-\tau_{j-1}^0|<\Delta_{n,j-1})} + \mathbf{1}_{(|x_{td}-\tau_j^0|<\Delta_{n,j})}$$

where for $j = 1$, the first term is defined as 0. Hence it suffices to show that for every j,

$$\frac{1}{n}\sum_{t=1}^{n}|z_t|\mathbf{1}_{(|x_{td}-\tau_j^0|<\Delta_{n,j})} = O_p(a_n^{1/v}).$$

By assumption, $\Delta_{n,j} = O_p(a_n)$. So for all $\epsilon > 0$ there exists $M > 0$ such that $P(\Delta_{n,j} > a_n M) < \epsilon$ for all n. Thus

$$\begin{aligned}
&P(\frac{1}{n}\sum_{t=1}^{n}|z_t|\mathbf{1}_{(|x_{td}-\tau_j^0|<\Delta_{n,j})} > a_n^{1/v}M)\\
=&P(\frac{1}{n}\sum_{t=1}^{n}|z_t|\mathbf{1}_{(|x_{td}-\tau_j^0|<\Delta_{n,j})} > a_n^{1/v}M, \Delta_{n,j} \leq a_n M)\\
&+ P(\frac{1}{n}\sum_{t=1}^{n}|z_t|\mathbf{1}_{(|x_{td}-\tau_j^0|<\Delta_{n,j})} > a_n^{1/v}M, \Delta_{n,j} > a_n M)\\
\leq&P(\frac{1}{n}\sum_{t=1}^{n}|z_t|\mathbf{1}_{(|x_{td}-\tau_j^0|<a_n M)} > a_n^{1/v}M) + \epsilon.
\end{aligned}$$

Hence it remains to show that $\frac{1}{a_n^{1/v}n}\sum_{t=1}^{n}|z_t|\mathbf{1}_{(|x_{td}-\tau_j^0|<a_n M)}$ is bounded in probability. However, in view of the Hölder's inequality and the assumptions, the expected value of this last quantity is bounded above by $(E|z_1|^u)^{1/u}a_n^{-1/v}(Ca_nM)^{1/v}$ for some constant C. This shows that

$$\frac{1}{a_n^{1/v}n}\sum_{t=1}^{n}|z_t|\mathbf{1}_{(|x_{td}-\tau_j|<a_n M)}$$

is bounded in L^1 and hence in probability. ¶

Proof of Theorem 3.3 Let $\tilde{\beta}_j^*$ be the "least squares estimates" of $\tilde{\beta}_j$, $j = 1,\cdots,l^0+1$, when l^0 and $(\tau_1^0,\cdots,\tau_{l^0}^0)$ are assumed known. Then by the law of large numbers, $\tilde{\beta}_j^* - \tilde{\beta}_j = o_p(1)$, $j = 1,\cdots,l^0+1$. So it suffices to show that $\hat{\tilde{\beta}}_j - \tilde{\beta}_j^* = o_p(1)$ for each j.

Set $X_j^* = I_n(\tau_{j-1}^0, \tau_j^0)X_n$ and $\hat{X}_j = I_n(\hat{\tau}_{j-1}, \hat{\tau}_j)X_n$. Then,

$$
\begin{aligned}
&\hat{\tilde{\beta}}_j - \tilde{\beta}_j^* \\
=&[(\frac{1}{n}\hat{X}_j'\hat{X}_j)^- - (\frac{1}{n}X_j^{*'}X_j^*)^-][\frac{1}{n}\hat{X}_j'Y_n] + [(\frac{1}{n}X_j^{*'}X_j^*)^-][\frac{1}{n}(\hat{X}_j - X_j^*)'Y_n] \\
=&[(\frac{1}{n}\hat{X}_j'\hat{X}_j)^- - (\frac{1}{n}X_j^{*'}X_j^*)^-]\{\frac{1}{n}(\hat{X}_j' - X_j^*)'Y_n + \frac{1}{n}X_j^*Y_n\} + [(\frac{1}{n}X_j^{*'}X_j^*)^-][\frac{1}{n}(\hat{X}_j - X_j^*)'Y_n] \\
=:&(I)\{(II) + (III)\} + (IV)(II),
\end{aligned}
$$

where $(I) = [(\frac{1}{n}\hat{X}_j'\hat{X}_j)^- - (\frac{1}{n}X_j^{*'}X_j^*)^-]$, $(II) = \frac{1}{n}(\hat{X}_j' - X_j^*)'Y_n$, $(III) = \frac{1}{n}X_j^*Y_n$ and $(IV) = [(\frac{1}{n}X_j^{*'}X_j^*)^-]$. By the strong law of large numbers, both (III) and (IV) are $O_p(1)$. By Theorem 3.2, $\hat{\tau} - \tau^0 = o_p(1)$. Proposition 3.2 implies that there exists a sequence $\{a_n\}$, $a_n \to 0$ as $n \to \infty$ such that $\hat{\tau} - \tau^0 = O_p(a_n)$. Note that $(II) = \frac{1}{n}\sum_{t=1}^n \mathbf{x}_t y_t(\mathbf{1}_{(x_{td}\in\hat{R}_j)} - \mathbf{1}_{(x_{td}\in R_j)})$ where $\hat{R}_j = (\hat{\tau}_{j-1}, \hat{\tau}_j]$, $R_j = (\tau_{j-1}^0, \tau_j^0]$. Taking $u > 1$ and $z_t = \mathbf{a}'\mathbf{x}_t y_t$ for any real vector $\mathbf{a}$, it follows from Lemma 3.6 that $(II) = o_p(1)$. If $(I) = o_p(1)$, then $\hat{\tilde{\beta}}_j - \tilde{\beta}_j^* = o_p(1)$, $j = 1, \cdots, l^0 + 1$. So, it remains only to show that $(I) = o_p(1)$.

By the strong law of large numbers, $\frac{1}{n}X_j^{*\prime}X_j^* \xrightarrow{a.s.} E\{\mathbf{x}_1\mathbf{x}_1\mathbf{1}_{(x_{1d}\in(\tau_{j-1}^0, \tau_j^0])}\} > 0$. If we can show that $\frac{1}{n}\hat{X}_j'\hat{X}_j - \frac{1}{n}X_j^{*\prime}X_j^* = o_p(1)$, then for sufficiently large n, $(\frac{1}{n}\hat{X}_j'\hat{X}_j)^{-1}$ and $(\frac{1}{n}X_j^{*\prime}X_j^*)^{-1}$ exist with probability approaching 1. And, $(\frac{1}{n}\hat{X}_j'\hat{X}_j)^- - (\frac{1}{n}X_j^{*\prime}X_j^*)^- = o_p(1)$. So, it suffices to show that $\frac{1}{n}\hat{X}_j'\hat{X}_j - \frac{1}{n}X_j^{*\prime}X_j^* = o_p(1)$. Let $\mathbf{a} \neq \mathbf{0}$ be a constant vector and $z_t = (\mathbf{a}'\mathbf{x}_t)^2$. Then $\mathbf{a}'(\frac{1}{n}\hat{X}_j'\hat{X}_j - \frac{1}{n}X_j^{*\prime}X_j^*)\mathbf{a} = \frac{1}{n}\sum_{t=1}^n \mathbf{a}'\mathbf{x}_t\mathbf{x}_t'\mathbf{a}(\mathbf{1}_{(x_{td}\in\hat{R}_j)} - \mathbf{1}_{(x_{td}\in R_j)}) = \frac{1}{n}\sum_{t=1}^n z_t(\mathbf{1}_{(x_{td}\in\hat{R}_j)} - \mathbf{1}_{(x_{td}\in R_j)})$. Taking the sequence $\{a_n\}$ in the last paragraph and $u > 1$, it follows from Lemma 3.6 that $\mathbf{a}'(\frac{1}{n}\hat{X}_j'\hat{X}_j - \frac{1}{n}X_j^{*\prime}X_j^*)\mathbf{a} = o_p(1)$ and hence $\frac{1}{n}\hat{X}_j'\hat{X}_j - \frac{1}{n}X_j^{*\prime}X_j^* = o_p(1)$.

This completes the proof. ¶

The proof of Theorem 3.4 depends on the following results.

Proposition 3.3 *(Serfling, 1980, p32)* *Let* $\{y_{nt}, 1 \le t \le K_n, n = 1, 2, \ldots\}$ *be a double array* *with independent random variables within rows. Suppose, for some $\nu > 2$,*

$$\sum_{t=1}^{K_n} E|y_{nt} - \mu_{nt}|^{\nu} = o(B_n^{\nu}), \quad n \to \infty.$$

Then

$$B_n^{-1}[\sum_{t=1}^{n} y_{nt} - A_n] \xrightarrow{d} N(0,1), \quad as\ n \to \infty,$$

where $\mu_{nt} = E(y_{nt}), A_n = \sum_{t=1}^{K_n} \mu_{nt}$ and $B_n^2 = \sum_{t=1}^{K_n} Var(y_{nt})$.

Lemma 3.7 *Let $\{k_n\}$ be a sequence of positive numbers such that $k_n \to 0$ and $nk_n \to \infty$. Assumptions 3.0 and 3.3 imply that for any $j = 1, \cdots, l^0$,*

(i)

$$\frac{1}{nk_n} X_n'(\tau_j^0 - k_n, \tau_j^0) X_n(\tau_j^0 - k_n, \tau_j^0) \xrightarrow{p} E(\mathbf{x}_1\mathbf{x}_1' | x_{1d} = \tau_j^0) f_d(\tau_j^0),$$
$$\frac{1}{nk_n} X_n'(\tau_j^0, \tau_j^0 + k_n) X_n(\tau_j^0, \tau_j^0 + k_n) \xrightarrow{p} E(\mathbf{x}_1\mathbf{x}_1' | x_{1d} = \tau_j^0) f_d(\tau_j^0),$$

(ii)

$$\frac{1}{nk_n} \epsilon_n'(\tau_j^0 - k_n, \tau_j^0) \epsilon_n(\tau_j^0 - k_n, \tau_j^0) \xrightarrow{p} \sigma_0^2 f_d(\tau_j^0),$$
$$\frac{1}{nk_n} \epsilon_n'(\tau_j^0, \tau_j^0 + k_n) \epsilon_n(\tau_j^0, \tau_j^0 + k_n) \xrightarrow{p} \sigma_0^2 f_d(\tau_j^0),$$

(iii)

$$\frac{1}{nk_n} \epsilon_n'(\tau_j^0 - k_n, \tau_j^0) X_n(\tau_j^0 - k_n, \tau_j^0) \xrightarrow{p} \mathbf{0},$$
$$\frac{1}{nk_n} \epsilon_n'(\tau_j^0, \tau_j^0 + k_n) X_n(\tau_j^0, \tau_j^0 + k_n) \xrightarrow{p} \mathbf{0}.$$

Proof It suffices to show the second equation in each of (i), (ii) and (iii), the proofs of the first deferring only in a formalistic sense.

(i) Note that $X_n'(\tau_j^0, \tau_j^0 + k_n) X_n(\tau_j^0, \tau_j^0 + k_n) = \sum_{t=1}^{n} \mathbf{x}_t \mathbf{x}_t' \mathbf{1}_{(x_{td} \in (\tau_j^0, \tau_j^0 + k_n])}$. Let $\mathbf{a} \neq \mathbf{0}$ be a constant vector, $y_{nt} = \mathbf{a}'\mathbf{x}_t\mathbf{x}_t'\mathbf{a}\mathbf{1}_{(x_{td} \in (\tau_j^0, \tau_j^0 + k_n])}$, $\mu_n = E(y_{nt})$, and $\sigma_n^2 = Var(y_{nt})$. If

$E[(a'\mathbf{x}_t)^2|\tau_j^0] > 0$, then $E[(a'\mathbf{x}_t)^4|\tau_j^0] > 0$ and

$$\begin{aligned}
\mu_n =& E[(\mathbf{a}'\mathbf{x}_t)^2 \mathbf{1}_{(x_{td}\in(\tau_j^0,\tau_j^0+k_n])}] \\
=& E\{\mathbf{1}_{(x_{td}\in(\tau_j^0,\tau_j^0+k_n])} E[(\mathbf{a}'\mathbf{x}_t)^2|x_{td}]\} \\
=& E[(\mathbf{a}'\mathbf{x}_1)^2|x_{1d} = \theta_n] f_d(\theta_n) k_n \\
=& E[(\mathbf{a}'\mathbf{x}_1)^2|x_{1d} = \tau_j^0] f_d(\tau_j^0) k_n + o(k_n),
\end{aligned}$$

where $\theta_n \in (\tau_j^0, \tau_j^0 + k_n]$ and $f_d(\cdot)$ is the marginal density function of x_{td}. Similarly,

$$\begin{aligned}
\sigma_n^2 =& E y_{nt}^2 - \mu_n^2 \\
=& E[(\mathbf{a}'\mathbf{x}_t)^4|x_{td} = \eta_n] f_d(\eta_n) k_n - \mu_n^2 \\
=& E[(\mathbf{a}'\mathbf{x}_t)^4|x_{td} = \tau_j^0] f_d(\tau_j^0) k_n + o(k_n),
\end{aligned}$$

where $\eta_n \in (\tau_j^0, \tau_j^0 + k_n]$ and for sufficiently large n, $\sigma_n^2 > 0$. By Minkowski's inequality, for $\nu > 2$,

$$\begin{aligned}
E|y_{n1} - \mu_n|^\nu \le& 2^{\nu-1}(E|y_{n1}|^\nu + \mu_n^\nu) \\
=& 2^{\nu-1}\{E[(\mathbf{a}'\mathbf{x}_1)^{2\nu}|x_{1d} = \xi_n] f_d(\xi_n) k_n + (E[(\mathbf{a}'\mathbf{x}_1)^2|x_{1d} = \theta_n] f_d(\theta_n) k_n)^\nu\} \\
=& 2^{\nu-1} E[(\mathbf{a}'\mathbf{x}_1)^{2\nu}|x_{1d} = \tau_j^0] f_d(\tau_j^0) k_n + o(k_n),
\end{aligned}$$

where $\xi_n \in (\tau_j^0, \tau_j^0 + k_n]$. So by setting $A_n = n\mu_n$ and $B_n^2 = n\sigma_n^2$, we have

$$\begin{aligned}
\sum_{t=1}^n E|y_{nt} - \mu_n|^\nu / B_n^\nu =& n^{-(\nu/2-1)} E|y_{nt} - \mu_n|^\nu / (E|y_{nt} - \mu_n|^2)^{\nu/2} \\
\le& n^{-(\nu/2-1)} \frac{2^{\nu-1} E[(\mathbf{a}'\mathbf{x}_1)^{2\nu}|x_{1d} = \tau_j^0] f_d(\tau_j^0) k_n + o(k_n)}{(E[(\mathbf{a}'\mathbf{x}_1)^4|x_{1d} = \tau_j^0] f_d(\tau_j^0) k_n + o(k_n))^{\nu/2}} \\
\to& 0,
\end{aligned}$$

as $n \to \infty$ since $\nu > 2$. Hence by *Proposition 3.3*,

$$B_n^{-1}[\sum_{t=1}^n y_{nt} - A_n] \xrightarrow{d} N(0,1), \quad as\ n \to \infty.$$

Now, since

$$B_n^2/(nk_n)^2 = O_p(ln^2 n)/ln^4 n = O_p(ln^{-2} n),$$

we obtain

$$\begin{aligned}
&\mathbf{a}'[\frac{1}{nk_n}X_n'(\tau_j^0,\tau_j^0+k_n)X_n(\tau_j^0,\tau_j^0+k_n)]\mathbf{a}\\
=&\frac{1}{nk_n}\sum_{t=1}^n y_{nt}\xrightarrow{p}\mathbf{a}'E(\mathbf{x}_t\mathbf{x}_t'|x_{td}=\tau_j^0)\mathbf{a}f_d(\tau_j^0),\quad as\ n\to\infty.
\end{aligned}$$

If $E[(\mathbf{a}'\mathbf{x}_1)^2|x_{1d}=\tau_j^0]=0$, it suffices to show that $\frac{1}{nk_n}\mathbf{a}'X_n'(\tau_j^0,\tau_j^0+k_n)X_n(\tau_j^0,\tau_j^0+k_n)\mathbf{a}$ converges to 0 in L_1.

$$\begin{aligned}
&E(\frac{1}{nk_n}\mathbf{a}'X_n'(\tau_j^0,\tau_j^0+k_n)X_n(\tau_j^0,\tau_j^0+k_n)\mathbf{a})\\
=&\frac{1}{nk_n}n\mu_n\\
=&E[(\mathbf{a}'\mathbf{x}_t)^2\mathbf{1}_{(x_{td}\in(\tau_j^0,\tau_j^0+k_n])}]/k_n\\
=&E\{\mathbf{1}_{(x_{td}\in(\tau_j^0,\tau_j^0+k_n])}E[(\mathbf{a}'\mathbf{x}_t)^2|x_{td}]\}/k_n\\
=&E[(\mathbf{a}'\mathbf{x}_1)^2|x_{1d}=\theta_n]f_d(\theta_n)\\
=&E[(\mathbf{a}'\mathbf{x}_1)^2|x_{1d}=\tau_j^0]f_d(\tau_j^0)+o(1)\\
=&o(1),
\end{aligned}$$

as $n\to\infty$, where $\theta_n\in(\tau_j^0,\tau_j^0+k_n)$. This completes the proof.

(ii) Similarly to (i), let $y_{nt}=\epsilon_t^2\mathbf{1}_{(x_{td}\in(\tau_j^0,\tau_j^0+k_n])}$, $\mu_n=E(y_{nt})$, and $\sigma_n^2=Var(y_{nt})$. Then

$$\begin{aligned}
\mu_n=&E[\epsilon_t^2\mathbf{1}_{(x_{td}\in(\tau_j^0,\tau_j^0+k_n])}]\\
=&\sigma_0^2P(x_{td}\in(\tau_j^0,\tau_j^0+k_n])\\
=&\sigma_0^2[f_d(\tau_j^0)k_n+o(k_n)],
\end{aligned}$$

$$\begin{aligned}
\sigma_n^2=&E(y_{nt}^2)-\mu_n^2\\
=&E(\epsilon_t^4)P(x_{td}\in(\tau_j^0,\tau_j^0+k_n])-\mu_n^2\\
=&E(\epsilon_t^4)f_d(\tau_j^0)k_n+o(k_n)-\mu_n^2\\
=&E(\epsilon_t^4)f_d(\tau_j^0)k_n+o(k_n).
\end{aligned}$$

By Minkowski's inequality, for $\nu > 2$,

$$
\begin{aligned}
E|y_{n1} - \mu_n|^\nu \leq & 2^{\nu-1}(E|y_{n1}|^\nu + \mu_n^\nu) \\
= & 2^{\nu-1}\{E(\epsilon_1^{2\nu})P(x_{td} \in (\tau_j^0, \tau_j^0 + k_n]) + (\sigma_0^2[f_d(\tau_j^0)k_n + o(k_n)])^\nu\} \\
= & 2^{\nu-1}E(\epsilon_1^{2\nu})f_d(\tau_j^0)k_n + o(k_n).
\end{aligned}
$$

So by setting $A_n = n\mu_n$ and $B_n^2 = n\sigma_n^2$, we have

$$
\begin{aligned}
\sum_{t=1}^n E|y_{nt} - \mu_n|^\nu / B_n^\nu = & n^{-(\nu/2-1)} E|y_{nt} - \mu_n|^\nu / (E|y_{nt} - \mu_n|^2)^{\nu/2} \\
\leq & n^{-(\nu/2-1)} \frac{2^{\nu-1}[E(\epsilon_1)^{2\nu} f_d(\tau_j^0)k_n + o(k_n)]}{(E(\epsilon_1)^4 f_d(\tau_j^0)k_n + o(k_n))^{\nu/2}} \\
\to & 0,
\end{aligned}
$$

as $n \to \infty$. Hence by *Proposition 3.3*,

$$
B_n^{-1}[\sum_{t=1}^n y_{nt} - A_n] \xrightarrow{d} N(0,1), \quad as\ n \to \infty.
$$

By the fact that

$$
B_n^2/(nk_n)^2 = O_p(ln^2 n)/ln^4 n = O_p(ln^{-2} n),
$$

we obtain

$$
\frac{1}{nk_n}\epsilon_n'(\tau_j^0, \tau_j^0 + k_n)\epsilon_n(\tau_j^0, \tau_j^0 + k_n) = \frac{1}{nk_n}\sum_{t=1}^n y_{nt} \xrightarrow{p} \sigma_0^2 f_d(\tau_j^0), \quad as\ n \to \infty.
$$

(iii) For any $\mathbf{a} \neq \mathbf{0}$,

$$
\begin{aligned}
& E(\frac{1}{nk_n}\epsilon_n'(\tau_j^0, \tau_j^0 + k_n)X_n(\tau_j^0, \tau_j^0 + k_n)\mathbf{a})^2 \\
= & E[\frac{1}{n^2k_n^2}(\sum_{t=1}^n \epsilon_t \mathbf{a}'\mathbf{x}_t \mathbf{1}_{(x_{td} \in (\tau_j^0, \tau_j^0 + k_n])})^2] \\
= & \frac{1}{n^2k_n^2}\sum_{t=1}^n E[\epsilon_t^2(\mathbf{a}'\mathbf{x}_t)^2 \mathbf{1}_{(x_{td} \in (\tau_j^0, \tau_j^0 + k_n])}] \\
= & \frac{1}{nk_n^2}\sigma_0^2 E[(\mathbf{a}'\mathbf{x}_t)^2 \mathbf{1}_{(x_{1d} \in (\tau_j^0, \tau_j^0 + k_n])}] \\
= & \frac{1}{nk_n^2}\sigma_0^2(E[(\mathbf{a}'\mathbf{x}_t)^2 | x_{1d} = \tau_j^0] f_d(\tau_j^0) + o(1))k_n \\
= & \frac{1}{nk_n}\sigma_0^2(E[(\mathbf{a}'\mathbf{x}_t)^2 | x_{1d} = \tau_j^0] f_d(\tau_j^0) + o(1)) \longrightarrow 0
\end{aligned}
$$

as $n \to \infty$. ¶

The approach of the following proof is to show that uniformly for all τ_j such that $|\tau_j - \tau_j^0| > O_p(\ln^2 n/n)$, $S_n(\tau_1, \cdots, \tau_{l^0}) > S_n(\tau_1^0, \cdots, \tau_{l^0}^0)$ for sufficiently large n. We shall achieve this by showing

$$S_n(\tau_{j-1}^0 + \delta, \tau_j) + S_n(\tau_j, \tau_{j+1}^0 - \delta) - [S_n(\tau_{j-1}^0 + \delta, \tau_j^0) + S_n(\tau_j^0, \tau_{j+1}^0 - \delta)] + O_p(\ln^2 n) > 0$$

for sufficiently large n.

Proof of Theorem 3.4 By Theorem 3.1, the problem can be restricted to $\{\hat{l} = l^0\}$. Suppose for some j, $P(\mathbf{x}_1'(\tilde{\beta}_{j+1}^0 - \tilde{\beta}_j^0) \neq 0 | x_d = \tau_j^0) > 0$. Hence $\Delta = E[(\mathbf{x}_1'(\tilde{\beta}_{j+1}^0 - \tilde{\beta}_j^0))^2 | x_d = \tau_j^0] > 0$. Let $\hat{\beta}(\alpha, \eta)$ be the minimizer of $||Y_n(\alpha, \eta) - X_n(\alpha, \eta)\tilde{\beta}||^2$. Set $k_n = K \ln^2 n/n$ for $n = 1, 2, \cdots$, where K will be chosen later. The proofs of Lemma 3.6 and Theorem 3.3 show that if $\alpha_n \xrightarrow{p} \alpha$, $\eta_n \xrightarrow{p} \eta$, then $\hat{\beta}(\alpha_n, \eta_n) \xrightarrow{p} \hat{\beta}(\alpha, \eta)$ as $n \to \infty$. Hence, for $\tau_j^0 + k_n \to \tau_j^0$ as $n \to \infty$, $\hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n) \xrightarrow{p} \hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0)$ as $n \to \infty$. By *Assumption 3.2*, for any sufficiently small $\delta \in (\tau_{j-1}^0, \tau_j^0)$, $E\{\mathbf{x}_1\mathbf{x}_1' \ \mathbf{1}_{(x_{1d} \in (\tau_{j-1}^0 + \delta, \tau_j^0])}\}$ is positive definite, hence $\hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0) \xrightarrow{a.s.} \tilde{\beta}_j^0$ as $n \to \infty$. Therefore $\hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n) \xrightarrow{p} \tilde{\beta}_j^0$. So, there exists a sufficiently small $\delta > 0$ such that for all sufficiently large n, $||\hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n) - \tilde{\beta}_j^0|| \leq ||\tilde{\beta}_j^0 - \tilde{\beta}_{j+1}^0||$ and $(\hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n) - \tilde{\beta}_{j+1}^0)' E(\mathbf{x}_1\mathbf{x}_1' | x_{1d} = \tau_j^0)\ (\hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n) - \tilde{\beta}_{j+1}^0) > \Delta/2$ with probability approaching 1. Hence by Theorem 3.2, for any $\epsilon > 0$, there exists N_1 such that for $n > N_1$, with probability larger than $1 - \epsilon$, we have

(i) $|\hat{\tau}_i - \tau_i^0| < \delta$, $i = 1, \cdots, l^0$,

(ii) $||\hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n) - \tilde{\beta}_{j+1}^0||^2 \leq 2||\tilde{\beta}_j^0 - \tilde{\beta}_{j+1}^0||^2$ and

(iii) $(\hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n) - \tilde{\beta}_{j+1}^0)' E(\mathbf{x}_1\mathbf{x}_1' | x_{1d} = \tau_j^0)(\hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n)) - \tilde{\beta}_{j+1}^0) > \Delta/2$.

Let $A_j = \{(\tau_1, \cdots, \tau_{l^0}) : |\tau_i - \tau_i^0| < \delta,\ i = 1, \cdots, l^0,\ |\tau_j - \tau_j^0| > k_n\}$ $j = 1, \cdots, l^0$. Since for the least squares estimates $\hat{\tau}_1, \cdots, \hat{\tau}_{l^0}$, $S_n(\hat{\tau}_1, \cdots, \hat{\tau}_{l^0}) \leq S_n(\tau_1^0, \cdots, \tau_{l^0}^0)$,

$$\inf_{(\tau_1,\cdots,\tau_{l^0}) \in A_j} \{S_n(\tau_1, \cdots, \tau_{l^0}) - S_n(\tau_1^0, \cdots, \tau_{l^0}^0)\} > 0$$

implies $(\hat{\tau}_1, \cdots, \hat{\tau}_{l^0}) \notin A_j$, or, $|\hat{\tau}_j - \tau_j^0| \leq k_n = K \ln^2 n/n$ when (i) holds. By (i), if we show that for each j, there exists $N > N_1$ such that for all $n > N$, with probability larger than $1 - 2\epsilon$, $\inf_{(\tau_1,\cdots,\tau_{l^0}) \in A_j} \{S_n(\tau_1, \cdots, \tau_{l^0}) - S_n(\tau_1^0, \cdots, \tau_{l^0}^0)\} > 0$, we will have proved the desired result. Furthermore, by symmetry, we can consider the case when $\tau_j > \tau_j^0$ only. Hence A_j may be replaced by $A_j' = \{(\tau_1, \cdots, \tau_{l^0}) : |\tau_i - \tau_i^0| < \delta,\ i = 1, \cdots, l^0,\ \tau_j - \tau_j^0 > k_n\}$. For any $(\tau_1, \cdots, \tau_{l^0}) \in A_j'$, let $\xi_1 \leq \cdots \leq \xi_{2l^0+1}$ be the set $\{\tau_1, \ldots, \tau_{l^0},\ \tau_1^0, \cdots, \tau_{j-1}^0, \tau_{j-1}^0 + \delta,\ \tau_{j+1}^0 - \delta, \tau_{j+1}^0, \cdots, \tau_{l^0}^0\}$ after ordering its elements and let $\xi_0 = -\infty$, $\xi_{2l^0+2} = \infty$. Using *Proposition 3.1* (ii) twice, we have

$$\begin{aligned}
&\sum_{i:\xi_i \neq \tau_j, \tau_{j+1}^0 - \delta} S_n(\xi_{i-1}, \xi_i) + S_n(\tau_{j-1}^0 + \delta, \tau_j^0) + S_n(\tau_j^0, \tau_{j+1}^0 - \delta) \\
=&\tilde{\epsilon}_n' \tilde{\epsilon}_n + O_p(\ln^2 n) \\
=&[S_n(\tau_1^0, \cdots, \tau_{l^0}^0) + O_p(\ln^2 n)] + O_p(\ln^2 n) \\
=&S_n(\tau_1^0, \cdots, \tau_{l^0}^0) + O_p(\ln^2 n).
\end{aligned}$$

Thus,

$$\begin{aligned}
S_n(\tau_1, \cdots, \tau_{l^0}) \geq & S_n(\xi_1, \cdots, \xi_{2l^0+1}) \\
= & \sum_{i=1}^{2l^0+2} S_n(\xi_{i-1}, \xi_i) \\
= & \sum_{i:\xi_i \neq \tau_j, \tau_{j+1}^0 - \delta} S_n(\xi_{i-1}, \xi_i) + S_n(\tau_{j-1}^0 + \delta, \tau_j) + S_n(\tau_j, \tau_{j+1}^0 - \delta) \\
= & \sum_{i:\xi_i \neq \tau_j, \tau_{j+1}^0 - \delta} S_n(\xi_{i-1}, \xi_i) + S_n(\tau_{j-1}^0 + \delta, \tau_j^0) + S_n(\tau_j^0, \tau_{j+1}^0 - \delta) \\
& + [S_n(\tau_{j-1}^0 + \delta, \tau_j) + S_n(\tau_j, \tau_{j+1}^0 - \delta)] - [S_n(\tau_{j-1}^0 + \delta, \tau_j^0) + S_n(\tau_j^0, \tau_{j+1}^0 - \delta)] \\
= & S_n(\tau_1^0, \cdots, \tau_{l^0}^0) + O_p(ln^2 n) \\
& + [S_n(\tau_{j-1}^0 + \delta, \tau_j) + S_n(\tau_j, \tau_{j+1}^0 - \delta)] - [S_n(\tau_{j-1}^0 + \delta, \tau_j^0) + S_n(\tau_j^0, \tau_{j+1}^0 - \delta)],
\end{aligned}$$

where $O_p(ln^2 n)$ is independent of $(\tau_1, \cdots, \tau_{l^0}) \in A'_j$. It suffices to show that for $B_n = \{\tau_j : \tau_j \in (\tau_j^0 + k_n, \tau_j^0 + \delta)\}$ and sufficiently large n,

$$\begin{aligned}&\inf_{\tau_j \in B_n} \{S_n(\tau_{j-1}^0 - \delta, \tau_j) + S_n(\tau_j, \tau_{j+1}^0 - \delta) - [S_n(\tau_{j-1}^0 + \delta, \tau_j^0) + S_n(\tau_j^0, \tau_{j+1}^0 - \delta)]\} \\ &>M' ln^2 n\end{aligned} \tag{3.8}$$

with probability larger than $1 - 2\epsilon$ for some fixed $M' > 0$. Let

$$S_n(\alpha, \eta; \tilde{\beta}) = \|Y_n(\alpha, \eta) - X_n(\alpha, \eta)\tilde{\beta}\|^2 = \sum_{t=1}^{n} (y_t - \mathbf{x}_t'\tilde{\beta})^2 1_{(x_{td} \in (\alpha, \eta))}.$$

Since $S_n(\alpha, \eta) = S_n(\alpha, \eta; \hat{\beta}(\alpha, \eta))$, we have

$$\begin{aligned}&S_n(\tau_{j-1}^0 + \delta, \tau_j) \\ \geq &S_n(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n) + S_n(\tau_j^0 + k_n, \tau_j) \\ = &S_n(\tau_{j-1}^0 + \delta, \tau_j^0; \hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n)) + S_n(\tau_j^0, \tau_j^0 + k_n; \hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n)) \\ &+ S_n(\tau_j^0 + k_n, \tau_j) \\ \geq &S_n(\tau_{j-1}^0 + \delta, \tau_j^0) + S_n(\tau_j^0, \tau_j^0 + k_n; \hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n)) + S_n(\tau_j^0 + k_n, \tau_j).\end{aligned} \tag{3.9}$$

And since $(\tau_j^0 + k_n, \tau_{j+1}^0 - \delta] \subset (\tau_j^0, \tau_{j+1}^0]$ for sufficiently large n,

$$S_n(\tau_j^0 + k_n, \tau_{j+1}^0 - \delta; \tilde{\beta}_{j+1}^0) = \tilde{\epsilon}_n'(\tau_j^0 + k_n, \tau_{j+1}^0 - \delta)\tilde{\epsilon}_n(\tau_j^0 + k_n, \tau_{j+1}^0 - \delta).$$

Applying Proposition 3.1 (i), we have

$$\begin{aligned}0 \leq &S_n(\tau_j^0 + k_n, \tau_{j+1}^0 - \delta; \tilde{\beta}_{j+1}^0) - [S_n(\tau_j^0 + k_n, \tau_j) + S_n(\tau_j, \tau_{j+1}^0 - \delta)] \\ = &T_n(\tau_j^0 + k_n, \tau_j) + T_n(\tau_j, \tau_{j+1}^0 - \delta).\end{aligned}$$

By Lemma 3.2, the RHS is $O_p(\ln^2 n)$. Thus,

$$\begin{aligned}&S_n(\tau_j^0, \tau_{j+1}^0 - \delta) \\ \leq &S_n(\tau_j^0, \tau_{j+1}^0 - \delta; \tilde{\beta}_{j+1}^0) \\ = &S_n(\tau_j^0, \tau_j^0 + k_n; \tilde{\beta}_{j+1}^0) + S_n(\tau_j^0 + k_n, \tau_{j+1}^0 - \delta; \tilde{\beta}_{j+1}^0) \\ \leq &S_n(\tau_j^0, \tau_j^0 + k_n; \tilde{\beta}_{j+1}^0) + S_n(\tau_j^0 + k_n, \tau_j) + S_n(\tau_j, \tau_{j+1}^0 - \delta) + O_p(\ln^2 n),\end{aligned}$$

where $O_p(\ln^2 n)$ is independent of τ_j. Hence

$$\begin{aligned}&S_n(\tau_j, \tau_{j+1}^0 - \delta)\\ \geq &S_n(\tau_j^0, \tau_{j+1}^0 - \delta) - S_n(\tau_j^0, \tau_j^0 + k_n; \tilde{\beta}_{j+1}^0) - S_n(\tau_j^0 + k_n, \tau_j) + O_p(\ln^2 n).\end{aligned} \tag{3.10}$$

Therefore, by (3.9) and (3.10)

$$\begin{aligned}&[S_n(\tau_{j-1}^0 + \delta, \tau_j) + S_n(\tau_j, \tau_{j+1}^0 - \delta)] - [S_n(\tau_{j-1}^0 + \delta, \tau_j^0) + S_n(\tau_j^0, \tau_{j+1}^0 - \delta)]\\ \geq &S_n(\tau_j^0, \tau_j^0 + k_n; \hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n)) - S_n(\tau_j^0, \tau_j^0 + k_n; \tilde{\beta}_{j+1}^0) + O_p(\ln^2 n).\end{aligned}$$

Let $M > 0$ such that the term $|O_p(\ln^2 n)| \leq M \ln^2 n$ with probability larger than $1 - \epsilon$ for all $n > N_1$. To show (3.8), it suffices to show that for sufficiently large n,

$$S_n(\tau_j^0, \tau_j^0 + k_n; \hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n)) - S_n(\tau_j^0, \tau_j^0 + k_n; \tilde{\beta}_{j+1}^0) - M ln^2 n > M' ln^2 n,$$

or

$$S_n(\tau_j^0, \tau_j^0 + k_n; \hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n)) - S_n(\tau_j^0, \tau_j^0 + k_n; \tilde{\beta}_{j+1}^0) > (M' + M) ln^2 n \tag{3.11}$$

with large probability. Recall $S_n(\alpha, \eta; \tilde{\beta}) = ||Y_n(\alpha, \eta) - X_n(\alpha, \eta)\tilde{\beta}||^2$ and $Y_n(\tau_j^0, \tau_j^0 + k_n) = X(\tau_j^0, \tau_j^0 + k_n)\tilde{\beta}_{j+1}^0 + \tilde{\epsilon}_n(\tau_j^0, \tau_j^0 + k_n)$. Taking K sufficiently large and applying (ii), (iii) and Lemma 3.7 (i), (iii), we can see that there exists $N \geq N_1$ such that for any $n \geq N$,

$$\begin{aligned}&\frac{1}{nk_n}[S_n(\tau_j^0, \tau_j^0 + k_n; \hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n)) - S_n(\tau_j^0, \tau_j^0 + k_n; \tilde{\beta}_{j+1}^0)]\\ =&\frac{1}{nk_n}[||Y_n(\tau_j^0, \tau_j^0 + k_n) - X_n(\tau_j^0, \tau_j^0 + k_n)\hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n)||^2\\ &- ||Y_n(\tau_j^0, \tau_j^0 + k_n) - X_n(\tau_j^0, \tau_j^0 + k_n)\tilde{\beta}_{j+1}^0||^2]\\ =&\frac{1}{nk_n}[||X_n(\tau_j^0, \tau_j^0 + k_n)(\tilde{\beta}_{j+1}^0 - \hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n)) + \tilde{\epsilon}_n(\tau_j^0, \tau_j^0 + k_n)||^2\\ &- ||\tilde{\epsilon}_n(\tau_j^0, \tau_j^0 + k_n)||^2]\\ =&\frac{1}{nk_n}||X_n(\tau_j^0, \tau_j^0 + k_n)(\tilde{\beta}_{j+1}^0 - \hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n))||^2\\ &+ \frac{2}{nk_n}\tilde{\epsilon}_n'(\tau_j^0, \tau_j^0 + k_n)X_n(\tau_j^0, \tau_j^0 + k_n)(\tilde{\beta}_{j+1}^0 - \hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n))\\ >&\Delta/4 - \Delta/8 > (M' + M)/K\end{aligned}$$

with probability larger than $1 - 2\epsilon$. Since $k_n = Kln^2 n/n$, the above implies (3.11). ¶

Proof of Theorem 3.5 By Lemma 3.4 (ii), $\hat{\sigma}^2 - \frac{1}{n}\sum_{t=1}^{n}\epsilon_t^2 = O_p(\ln^2 n/n)$. So, $\hat{\sigma}^2$ and $\frac{1}{n}\sum_{t=1}^{n}\epsilon_t^2$ share the same asymptotic distribution. Applying the central limit theorem to $\{\epsilon_t^2\}$, we conclude that the asymptotic distribution of $\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\epsilon_t^2$ is normal.

Let $(\tilde{\beta}_1^*, \cdots, \tilde{\beta}_{l^0+1}^*)$ be the "least squares estimates" of $(\tilde{\beta}_1^0, \cdots, \tilde{\beta}_{l^0+1}^0)$ when l^0 and τ_j^0, $(j = 1, \cdots, l^0)$, are assumed known. Then it is clear that $\sqrt{n}[(\tilde{\beta}_1^{*'}, \cdots, \tilde{\beta}_{l^0+1}^{*'})' - (\tilde{\beta}_1^{0'}, \cdots, \tilde{\beta}_{l^0+1}^{0}{}')']$ converges in distribution to a normal distribution. So it suffices to show that $\tilde{\beta}_j^* - \hat{\tilde{\beta}}_j = o_p(n^{-1/2})$.

Set $X_j^* = I_n(\tau_{j-1}^0, \tau_j^0)X_n$ and $\hat{X}_j = I_n(\hat{\tau}_{j-1}, \hat{\tau}_j)X_n$. Then,

$$
\begin{aligned}
&\hat{\tilde{\beta}}_j - \tilde{\beta}_j^* \\
=&[(\frac{1}{n}\hat{X}_j'\hat{X}_j)^- - (\frac{1}{n}X_j^{*'}X_j^*)^-][\frac{1}{n}\hat{X}_j'Y_n] + [(\frac{1}{n}X_j^{*'}X_j^*)^-][\frac{1}{n}(\hat{X}_j - X_j^*)'Y_n] \\
=&[(\frac{1}{n}\hat{X}_j'\hat{X}_j)^- - (\frac{1}{n}X_j^{*'}X_j^*)^-]\{\frac{1}{n}(\hat{X}_j' - X_j^*)'Y_n + \frac{1}{n}X_j^*Y_n\} + [(\frac{1}{n}X_j^{*'}X_j^*)^-][\frac{1}{n}(\hat{X}_j - X_j^*)'Y_n] \\
=:&(I)\{(II) + (III)\} + (IV)(II).
\end{aligned}
$$

where $(I) = [(\frac{1}{n}\hat{X}_j'\hat{X}_j)^- - (\frac{1}{n}X_j^{*'}X_j^*)^-]$, $(II) = \frac{1}{n}(\hat{X}_j' - X_j^*)'Y_n$, $(III) = \frac{1}{n}X_j^*Y_n$ and $(IV) = [(\frac{1}{n}X_j^{*'}X_j^*)^-]$. As in the proof of Theorem 3.3, both (III) and (IV) are $O_p(1)$. By Theorem 3.4, $\hat{\tau} - \tau^0 = O_p(ln^2 n/n)$. The order of $o_p(n^{-1/2})$ of (I) and (II) follows from Lemma 3.6 by taking $a_n = ln^2 n/n$, $z_t = (\mathbf{a}'\mathbf{x}_t)^2$ and $z_t = \mathbf{a}'\mathbf{x}_t y_t$ respectively, for any real vector $\mathbf{a}$ and $u > 2$.

This completes the proof. ¶

3.2 Consistency of the estimated segmentation variable

Since d is assumed unknown in this section, we will use the notation such as $S_n(A)$, $T_n(A)$ introduced in Section 2.2. The two theorems in this section show that the two methods of estimating d^0 given in Section 2.2 produce consistent estimates, respectively.

Theorem 3.6 *If d^0 is asymptotically identifiable w.r.t. L, then under the conditions of Theorem 3.1, $\hat{d}$ given in Method 1 satisfies $P(\hat{d} = d^0) \longrightarrow 1$ as $n \to \infty$.*

Theorem 3.7 *Assume $\{\mathbf{x}_t\}$ are iid random vectors. If $\mathbf{z}_1 = (x_{11}, \ldots, x_{1p})'$ is a continuous random vector and the support of its distribution is $(a_1, b_1) \times \ldots \times (a_p, b_p)$, where $-\infty \leq a_i < b_i \leq \infty$, $i = 1, \ldots, p$, and for any $\alpha \in \mathbf{R}^p$, $E[(\mathbf{z}_1'\mathbf{z}_1)^u] < \infty$, for some $u > 1$, then $\hat{d}$ given by Method 2 satisfies $P(\hat{d} = d^0) \longrightarrow 1$ as $n \to \infty$.*

To prove Theorem 3.6, some results similar to those presented in the last section are needed. Lemmas 3.2′-3.3′ and Proposition 3.1′ below are generalizations of Lemmas 3.2-3.3 and Proposition 3.1 respectively.

Lemma 3.2′ *Assume for the segmented linear regression model (3.1) that Assumption 3.0 is satisfied. For any $d \neq d_0$ and $j = 1, \cdots, l^0 + 1$, let $R_j^d(\alpha, \eta) = \{\mathbf{x}_1 : \alpha < x_{1d} \leq \eta\} \cap R_j^0$, $-\infty \leq \alpha < \eta \leq \infty$. Then*

$$P\{\sup_{\alpha<\eta} T_n(R_j^d(\alpha, \eta)) \geq \frac{9p_0^3}{T_0^2} \ln^2 n\} \to 0, \qquad \text{as } n \to 0,$$

where p_0 is the true order of the model and T_0 is the constant associated with the local exponential boundedness condition for the $\{\epsilon_t\}$.

Proof Conditioning on X_n, we have for any j and $d \neq d_0$ that

$$\begin{aligned}
&P\{\sup_{\alpha<\eta} T_n(R_j^d(\alpha, \eta)) \geq \frac{9p_0^3}{T_0^2} \ln^2 n \mid X_n\} \\
=&P\{\max_{x_{sd}<x_{td}} \tilde{\epsilon}_n' H_n(R_j^d(x_{sd}, x_{td}))\tilde{\epsilon}_n \geq \frac{9p_0^3}{T_0^2} \ln^2 n \mid X_n\} \\
\leq& \sum_{x_{sd}<x_{td}} P\{\tilde{\epsilon}_n' H_n(R_j^d(x_{sd}, x_{td}))\tilde{\epsilon}_n \geq \frac{9p_0^3}{T_0^2} \ln^2 n \mid X_n\}.
\end{aligned}$$

Since $H_n(R_j^d(x_{sd}, x_{td}))$ is nonnegative definite and idempotent, it can be decomposed as

$$H_n(R_j^d(x_{sd}, x_{td})) = W'\Lambda W,$$

where W is orthogonal and $\Lambda = diag(1, \cdots, 1, 0, \cdots, 0)$ with $p := rank(H_n(R_j^d(x_{sd}, x_{td}))) = rank(\Lambda) \le p_0$. Set $Q = (I_p, \mathbf{0})W$. Then Q has full row rank p. Let $Q' = (\mathbf{q}_1, \cdots, \mathbf{q}_p)$ and $U_l = \mathbf{q}_l' \tilde{\epsilon}_n$, $l = 1, \ldots, p$. Then

$$\tilde{\epsilon}_n' H_n(R_j^d(x_{sd}, x_{td}))\tilde{\epsilon}_n = \tilde{\epsilon}_n' Q' Q \tilde{\epsilon}_n = \sum_{l=1}^{p} U_l^2.$$

Since $p \le p_0$ and

$$\begin{aligned} & P\{\sum_{l=1}^{p} U_l^2 \ge \frac{9p_0^3}{T_0^2} ln^2 n | X_n\} \\ \le & P\{\sum_{l=1}^{p} U_l^2 \ge p\frac{9p_0^2}{T_0^2} ln^2 n | X_n\} \\ \le & P\{U_l^2 \ge \frac{9p_0^2}{T_0^2} ln^2 n \ for \ \ some \ l | X_n\} \\ \le & \sum_{l=1}^{p} P\{U_l^2 \ge \frac{9p_0^2}{T_0^2} ln^2 n | X_n\}, \end{aligned}$$

it suffices to show, for any l, that

$$\sum_{x_{sd} < x_{td}} P\{U_l^2 \ge \frac{9p_0^2}{T_0^2} \ln^2 n \mid X_n\} \to 0, \qquad as\ n \to 0.$$

Noting that $p = trace(H_n^d(R_j^d(x_{sd}, x_{td}))) = \sum_{l=1}^{p} \| \mathbf{q}_l \|^2$, we have $\| \mathbf{q}_l \|^2 = \mathbf{q}_l' \mathbf{q}_l \le p \le p_0$, $l = 1, \ldots, p$. By Lemma 3.1, with $t_0 = T_0/p_0$ we have

$$\begin{aligned} \sum_{x_{sd} < x_{td}} P\{|U_l| \ge 3p_0 \ln n / T_0 \mid X_n\} \le & \sum_{x_{sd} < x_{td}} 2\exp(-\frac{T_0}{p_0} \cdot \frac{3p_0}{T_0} \ln n) \exp(c_0 (T_0/p_0)^2 p_0) \\ \le & \, n(n-1)/n^3 \exp(c_0 T_0^2 / p_0) \to 0, \end{aligned}$$

as $n \to \infty$, where c_0 is the constant specified in Lemma 3.1. Finally, by appealing to the dominated convergence theorem we obtain the desired result without conditioning. ¶

Proposition 3.1$'$ *Consider the segmented regression model 3.1.*

(i) For any subset B of the domain of $\mathbf{x}_1$ and any j,

$$S_n(B \cap R_j^0) = \tilde{\epsilon}_n'(B \cap R_j^0)\tilde{\epsilon}_n(B \cap R_j^0) - T_n(B \cap R_j^0).$$

(ii) Let $\{B_i\}_{i=1}^{m+1}$ be a partition of the domain of $\mathbf{x}_1$, where m is a finite positive integer. Then,

$$\sum_{i=1}^{m+1} S_n(B_i \cap R_j^0) = \tilde{\epsilon}_n'(R_j^0)\tilde{\epsilon}_n(R_j^0) - \sum_{i=1}^{m+1} T_n(B_i \cap R_j^0)$$

for all j. Further, if $B_i = \{\mathbf{x}_1 : \tau_{i-1} < x_{1d} \leq \tau_i\}$ for $d \neq d_0$ then Assumption 3.0 implies

$$\sum_{i=1}^{m+1} S_n(B_i \cap R_j^0) = \tilde{\epsilon}_n'(R_j^0)\tilde{\epsilon}_n(R_j^0) + O_p(\ln^2 n)$$

uniformly for all $\tau_1, \cdots, \tau_m$ such that $-\infty = \tau_0 < \tau_1 \cdots < \tau_m < \tau_{m+1} = \infty$.

Proof:

(i) Denote $A = B \cap R_j^0$.

$$\begin{aligned}
S_n(A) =& Y_n'(I_n(A) - H_n(A))Y_n \\
=& (X_n(A)\tilde{\beta}_j^0 + \tilde{\epsilon}_n(A))'(I_n(A) - H_n(A))(X_n(A)\tilde{\beta}_j^0 + \tilde{\epsilon}_n(A)) \\
=& \tilde{\beta}_j^{0\prime} X_n'(A)X_n(A)\tilde{\beta}_j^0 + 2\tilde{\epsilon}_n'(A)X_n(A)\tilde{\beta}_j^0 + \tilde{\epsilon}_n'(A)\tilde{\epsilon}_n(A) \\
&- [\tilde{\beta}_j^{0\prime} X_n'(A)H_n(A)X_n(A)\tilde{\beta}_j^0 + 2\tilde{\epsilon}_n'(A)H_n(A)X_n(A)\tilde{\beta}_j^0 + \tilde{\epsilon}_n'(A)H_n(A)\tilde{\epsilon}_n(A)].
\end{aligned}$$

Since $X_n'(A)H_n(A)X_n(A) = X_n'(A)X_n(A)$ and $H_n(A)$ is idempotent, we have

$$[X_n(A) - H_n(A)X_n(A)]'[X_n(A) - H_n(A)X_n(A)] = 0$$

and hence $H_n(A)X_n(A) = X_n(A)$. Thus,

$$S_n(A) = \tilde{\epsilon}_n'(A)\tilde{\epsilon}_n(A) - \tilde{\epsilon}_n(A)H_n(A)\tilde{\epsilon}_n(A) = \tilde{\epsilon}_n'(A)\tilde{\epsilon}_n(A) - T_n(A).$$

(ii) By (i),

$$\begin{aligned}
&\sum_{i=1}^{m+1} S_n(B_i \cap R_j^0) \\
=& \sum_{i=1}^{m+1} [\tilde{\epsilon}_n'(B_i \cap R_j^0)\tilde{\epsilon}_n(B_i \cap R_j^0) - T_n(B_i \cap R_j^0)] \\
=& \tilde{\epsilon}_n'(R_j^0)\tilde{\epsilon}_n(R_j^0) - \sum_{i=1}^{m+1} T_n(B_i \cap R_j^0).
\end{aligned}$$

If $B_i = \{\mathbf{x}_1 : \tau_{i-1} < x_{1d} \leq \tau_i\}$, denote $B_i \cap R_j^0$ by $R_j^d(\tau_{i-1}, \tau_i)$ for all i. Lemma 3.2$'$ implies $\sum_{i=1}^{m+1} T_n(B_i \cap R_j^0) = \sum_{i=1}^{m+1} T_n(R_j^d(\tau_{i-1}, \tau_i)) \leq (m+1)\sup_{\alpha<\eta} T_n(R_j^d(\alpha, \eta)) = O_p(\ln^2 n)$ uniformly for all $-\infty < \tau_1 < \cdots < \tau_m \leq \infty$. ¶

Lemma 3.3$'$ *Let A be a subset of the domain of $\mathbf{x}_1$. If both $E[\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(\mathbf{x}_1 \in A \cap R_r^0)}]$ and $E[\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(\mathbf{x}_1 \in A \cap R_{r+1}^0)}]$ are positive definite. Then under Assumption 3.0,*

$$[S_n(A) - S_n(A \cap R_r^0) - S_n(A \cap R_{r+1}^0)]/n \xrightarrow{a.s.} C_r$$

for some $C_r > 0$ as $n \to \infty$, $r = 1, \cdots, l^0$.

Proof It suffices to prove the result when $l^0 = 1$. For notational simplicity, we omit the subscripts and superscripts 0 in this proof. Let $X_j^* = X_n(A \cap R_j)$, $\tilde{\epsilon}_j^* = \tilde{\epsilon}_n(A \cap R_j)$, $j = 1, 2$, $X^* = X_1^* + X_2^*$, $\tilde{\epsilon}^* = \tilde{\epsilon}_1^* + \tilde{\epsilon}_2^*$ and $\hat{\tilde{\beta}} = (X^{*\prime}X^*)^- X^{*\prime}Y_n$. As in ordinary regression, we have

$$\begin{aligned}
&S_n(A)\\
=&||X_1^*\tilde{\beta}_1 + X_2^*\tilde{\beta}_2 + \tilde{\epsilon}^* - X^*\hat{\tilde{\beta}}||^2\\
=&||X_1^*(\tilde{\beta}_1 - \hat{\tilde{\beta}}) + X_2^*(\tilde{\beta}_2 - \hat{\tilde{\beta}}) + \tilde{\epsilon}^*||^2\\
=&||X_1^*(\tilde{\beta}_1 - \hat{\tilde{\beta}})||^2 + ||X_2^*(\tilde{\beta}_2 - \hat{\tilde{\beta}})||^2 + ||\tilde{\epsilon}^*||^2 + 2\tilde{\epsilon}^{*\prime}X_1^*(\tilde{\beta}_1 - \hat{\tilde{\beta}}) + 2\tilde{\epsilon}^{*\prime}X_2^*(\tilde{\beta}_2 - \hat{\tilde{\beta}}).
\end{aligned}$$

It then follows from the strong law of large numbers for stationary ergodic stochastic processes that as $n \to \infty$,

$$\frac{1}{n}X^{*\prime}X^* = \frac{1}{n}\sum_{t=1}^{n}\mathbf{x}_t\mathbf{x}_t'\mathbf{1}_{(\mathbf{x}_t \in A)} \xrightarrow{a.s.} E\{\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(\mathbf{x}_1 \in A)}\} > 0,$$
$$\frac{1}{n}X_j^{*\prime}X_j^* \xrightarrow{a.s.} E\{\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(\mathbf{x}_1 \in A \cap R_j)}\} > 0,\ j = 1, 2,$$

and

$$\frac{1}{n}X^{*\prime}Y_n \xrightarrow{a.s.} E\{y_1\mathbf{x}_1\mathbf{1}_{(\mathbf{x}_1 \in A)}\}.$$

Therefore,

$$\hat{\tilde{\beta}} \xrightarrow{a.s.} \{E\{\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(\mathbf{x}_1 \in A)}\}\}^{-1}E\{y_1\mathbf{x}_1\mathbf{1}_{(\mathbf{x}_1 \in A)}\} =: \tilde{\beta}^*.$$

Similarly, it can be shown that

$$\frac{1}{n}||X_j^*(\tilde{\beta}_j - \hat{\tilde{\beta}})||^2 \xrightarrow{a.s.} (\tilde{\beta}_j - \tilde{\beta}^*)' E(\mathbf{x}_1\mathbf{x}_1' \mathbf{1}_{(\mathbf{x}_1 \in A\cap R_j)}) \cdot (\tilde{\beta}_j - \tilde{\beta}^*),$$

$$\frac{1}{n}\tilde{\epsilon}^{*\prime} X_j^*(\tilde{\beta}_j - \hat{\tilde{\beta}}) \xrightarrow{a.s.} 0,$$

for $j = 1, 2$, and

$$\frac{1}{n}||\tilde{\epsilon}^*||^2 \xrightarrow{a.s.} \sigma^2 P\{\mathbf{x}_1 \in A\}.$$

Thus as $n \to \infty$, $\frac{1}{n}S_n(A)$ has a finite limit, this limit being given by

$$\lim_{n\to\infty} \frac{1}{n}S_n(A)$$
$$=(\tilde{\beta}_1 - \tilde{\beta}^*)' E(\mathbf{x}_1\mathbf{x}_1' \mathbf{1}_{(\mathbf{x}_1 \in A\cap R_1)}) \cdot (\tilde{\beta}_1 - \tilde{\beta}^*) + (\tilde{\beta}_2 - \tilde{\beta}^*)' E(\mathbf{x}_1\mathbf{x}_1' \mathbf{1}_{(\mathbf{x}_1 \in A\cap R_2)}) \cdot (\tilde{\beta}_2 - \tilde{\beta}^*)$$
$$+ \sigma^2 P\{\mathbf{x}_1 \in A\}.$$

It remains to show that $\frac{1}{n}S_n(A \cap R_j)$ converges to $\sigma^2 P\{\mathbf{x}_1 \in A \cap R_j\}$, $j = 1, 2$, and at least one of $(\tilde{\beta}_1 - \tilde{\beta}^*)' E(\mathbf{x}_1\mathbf{x}_1' \mathbf{1}_{(\mathbf{x}_1 \in A\cap R_1)})(\tilde{\beta}_1 - \tilde{\beta}^*)$ and $(\tilde{\beta}_2 - \tilde{\beta}^*)' E(\mathbf{x}_1\mathbf{x}_1' \mathbf{1}_{(\mathbf{x}_1 \in A\cap R_2)})(\tilde{\beta}_2 - \tilde{\beta}^*)$ is positive. The latter is a direct consequence of the assumed conditions while the former can be shown again by the strong law of large numbers. By Proposition 3.1′ (i),

$$S_n(A \cap R_1) = \tilde{\epsilon}_n'(A \cap R_1)\tilde{\epsilon}_n(A \cap R_1) - T_n(A \cap R_1) = \tilde{\epsilon}_1^{*\prime}\tilde{\epsilon}_1^* - T_n(A \cap R_1).$$

The strong law of large numbers implies

$$\frac{1}{n}\tilde{\epsilon}_1^*\tilde{\epsilon}_1 \xrightarrow{a.s.} E[\epsilon_1^2 \mathbf{1}_{(\mathbf{x}_1 \in A\cap R_1)}] = \sigma^2 P(\mathbf{x}_1 \in A \cap R_1),$$

$$\frac{1}{n}\tilde{\epsilon}_1^{*\prime} X_1^* \xrightarrow{a.s.} E[\epsilon_1 \mathbf{x}_1 \mathbf{1}_{(\mathbf{x}_1 \in A\cap R_1)}] = 0,$$

as $n \to \infty$ and $W = \lim_{n\to\infty} \frac{1}{n}X_1^{*\prime}X_1^*$ is positive definite. Therefore,

$$\frac{1}{n}T_n(A \cap R_1) = (\frac{1}{n}\tilde{\epsilon}_1' X_1^*)(\frac{1}{n}X_1^{*\prime}X_1^*)^-(\frac{1}{n}X_1^{*\prime}\epsilon_1) \xrightarrow{a.s.} \mathbf{0}W^{-1}\mathbf{0} = 0$$

and hence $\frac{1}{n}S_n(A \cap R_1) \xrightarrow{a.s.} \sigma^2 P\{\mathbf{x}_1 \in A \cap R_1\}$. The same argument can also be used to show that $\frac{1}{n}S_n(A \cap R_2) \xrightarrow{a.s.} \sigma^2 P\{\mathbf{x}_1 \in A \cap R_2\}$. This completes the proof. ¶

Proof of Theorem 3.6 For $d = d^0$, by Lemma 3.4 (ii),

$$\frac{1}{n}\tilde{S}_n^d \xrightarrow{p} \sigma_0^2.$$

Thus, it suffices to show for $d \neq d^0$, that $\frac{1}{n}\tilde{S}_n^d > \sigma_0^2 + C$ for some constant $C > 0$ with probability approaching 1. Again, $l^0 = 1$ is assumed for simplicity. If $d \neq d^0$, by the identifiability of d^0 and Theorem 2.1, for any $\{R_j^d\}_{j=1}^{L+1}$, there exist $r, s \in \{1, \cdots, L+1\}$ such that $R_r^d \supset A_s^d$ where $A_s^d = \{\mathbf{x}_1 : x_{1d} \in [a_s, b_s]\}$ is defined in Theorem 2.1. Let $B_s = \{(\tau_1, \ldots, \tau_L) : R_r^d \supset A_s^d$ for some $r\}$. Then for any $(\tau_1, \ldots, \tau_L)$, $(\tau_1, \cdots, \tau_L) \in B_s$ for at least one $s \in \{1, \cdots, L+1\}$. Since $\hat{d}$ is chosen such that $\tilde{S}_n^{\hat{d}} \leq \tilde{S}_n^d$ for all d, it suffices to show that for $d \neq d^0$ and each s, there exists $C_s > 0$ such that

$$\inf_{(\tau_1, \ldots, \tau_L) \in B_s} \frac{1}{n} S_n^d(\tau_1, \ldots, \tau_L) \geq \sigma_0^2 + C_s \tag{3.12}$$

with probability approaching 1 as $n \to \infty$. For any $(\tau_1, \ldots, \tau_L) \in B_s$, let $R_{L+2}^d = \{\mathbf{x} : x_d \in (\tau_{r-1}, a_s)\}$, $R_{L+3}^d = \{\mathbf{x} : x_d \in (b_s, \tau_r]\}$. Then $R_r^d = A_s^d \cup R_{L+2}^d \cup R_{L+3}^d$. Note that the total sum of squared errors decreases as the partition becomes finer. By Proposition 3.1$'$ and the strong law of large numbers,

$$
\begin{aligned}
&\frac{1}{n} S_n^d(\tau_1, \cdots, \tau_L) \\
=&\frac{1}{n} \sum_{j=1}^{L+1} S_n(R_j^d) \\
\geq&\frac{1}{n}[\sum_{j=1, j\neq r}^{L+3} S_n(R_j^d) + S_n(A_s^d)] \\
\geq&\frac{1}{n}\{ \sum_{j=1, j\neq r}^{L+3} [S_n(R_j^d \cap R_1^0) + S_n(R_j^d \cap R_2^0)] + [S_n(A_s^d \cap R_1^0) + S_n(A_s^d \cap R_2^0)]\} \\
&+ \frac{1}{n}[S_n(A_s^d) - S_n(A_s^d \cap R_1^0) - S_n(A_s^d \cap R_2^0)] \\
=&\frac{1}{n}\{\tilde{\epsilon}_n'(R_1^0)\tilde{\epsilon}_n(R_1^0) + \tilde{\epsilon}_n'(R_2^0)\tilde{\epsilon}_n(R_2^0) + O_p(\ln^2 n)\} \\
&+ \frac{1}{n}[S_n(A_s^d) - S_n(A_s^d \cap R_1^0) - S_n(A_s^d \cap R_2^0)] \\
=&\frac{1}{n}\{\tilde{\epsilon}_n'\tilde{\epsilon}_n + O_p(\ln^2 n)\} + \frac{1}{n}[S_n(A_s^d) - S_n(A_s^d \cap R_1^0) - S_n(A_s^d \cap R_2^0)] \\
=&\sigma_0^2 + o_p(1) + \frac{1}{n}[S_n(A_s^d) - S_n(A_s^d \cap R_1^0) - S_n(A_s^d \cap R_2^0)].
\end{aligned}
\tag{3.13}
$$

Now it remains to show that $\frac{1}{n}[S_n(A_s^d) - S_n(A_s^d \cap R_1^0) - S_n(A_s^d \cap R_2^0)] > C_s$ for some $C_s > 0$, with probability approaching 1. By Theorem 2.1, $E[\mathbf{x}_1\mathbf{x}_1' 1_{(\mathbf{x}_1 \in A_s \cap R_i^0)}]$, $i = 1, 2$, are positive definite. Applying Lemma 3.3′ we obtain the desired result. ¶

To prove Theorem 3.7, we first define the kth percentile of a distribution function F as $\rho_k := \inf_t\{t : F(t) \geq k/100\}$. Let ρ_j^d and r_j^d be the $j * 100/(2L+2)$th percentile of F^d and F_n^d respectively, where F^d is the distribution function of x_{1d} and F_n^d is the empirical distribution function of $\{x_{td}\}$, $j = 1, \ldots, 2L+2$. If x_{1d} has positive density function over a neighborhood of ρ_j^d for each j, then by Theorem 2.3.1 of Serfling (1980, p75), r_j^d converges to ρ_j^d almost surely for any j. Now, we are ready to introduce three lemmas required by the proof of Theorem 3.7. In these three lemmas, we shall omit "d" in ρ_j^d and r_j^d for notational simplicity.

Lemma 3.8 *Suppose (z_t, x_{td}) is a strictly stationary process and the marginal cdf of x_{td} has* *bounded derivative at ρ_j for all j. If $r_j - \rho_j = o_p(1)$, $j = 1, \cdots, 2L+2$, and for some $u > 1$ $E|z_t|^u < \infty$, then*

$$\frac{1}{n}\sum_{t=1}^{n} z_t(\mathbf{1}_{(x_{td}\in(r_{j-1},r_j))} - \mathbf{1}_{(x_{td}\in(\rho_{j-1},\rho_j))}) = o_p(1).$$

Proof By the assumption, the marginal *cdf*, F_d, of x_{1d} satisfies Lipschitz condition in a small neighborhood of $x_{1d} = \rho_j$ for every j. By Proposition 3.2, $r_j - \rho_j = o_p(1)$ implies that there exists a positive sequence $\{a_n\}$ such that $a_n \to 0$ as $n \to \infty$ and $r_j - \rho_j = O_p(a_n)$. Applying Lemma 3.6 in with τ_j^0 and $\hat{\tau}_j$ replaced by ρ_j and r_j respectively, we obtain the desired result.

¶

For any $j \in \{1, \cdots, 2L+2\}$, let $R_j = \{\mathbf{x}_1 : \rho_{j-1} < x_{1d} \leq \rho_j\}$ and $\hat{R}_j = \{\mathbf{x}_1 : r_{j-1} < x_{1d} \leq r_j\}$. Also let

$$X_{i\rho}^* = X_n(R_j \cap R_i^0),$$

$$X_\rho^* = X_n(R_j),$$

$$\epsilon_\rho^* = \tilde{\epsilon}_n(R_j),$$

and

$$X_{ir}^* = X_n(\hat{R}_j \cap R_i^0),$$

$$X_r^* = X_n(\hat{R}_j),$$

$$\tilde{\epsilon}_r^* = \tilde{\epsilon}_n(\hat{R}_j),$$

where $i = 1, 2$. Under the conditions of Theorem 3.7, the support of the distribution of $\mathbf{z}_1$ is $(a_1, b_1) \times \ldots \times (a_p, b_p)$. Hence, for $d \neq d^0$, $E[\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(\mathbf{x}_1 \in R_j \cap R_i^0)}]$ is of full rank, $i = 1, 2$.

Lemma 3.9 *Under the conditions of Theorem 3.7,*

(i) $\frac{1}{n}X_{ir}^{*\prime}X_{ir}^* = \frac{1}{n}X_{i\rho}^{*\prime}X_{i\rho}^* + o_p(1)$, $i = 1, 2$;

(ii) $\frac{1}{n}[\tilde{\epsilon}_r^{*\prime}\tilde{\epsilon}_r^* - \tilde{\epsilon}_\rho^{*\prime}\tilde{\epsilon}_\rho^*] = o_p(1)$; *and*

(iii) $\frac{1}{n}X_{i\rho}^{*\prime}\tilde{\epsilon}_\rho^* = O_p(n^{-1/2})$, $\frac{1}{n}X_{ir}^{*\prime}\tilde{\epsilon}_r^* = o_p(1)$, $i = 1, 2$.

Proof: With loss of generality, we can assume $P(R_j \bigcap R_i^0) > 0,\ i = 1, 2$.

(i) For any $\mathbf{a} \neq 0$,

$$\mathbf{a}'(\frac{1}{n}X_{ir}^{*}{}'X_{ir}^{*} - \frac{1}{n}X_{i\rho}^{*}{}'X_{i\rho}^{*})\mathbf{a} = \frac{1}{n}\sum_{t=1}^{n}(\mathbf{a}'\mathbf{x}_t)^2\mathbf{1}_{(\mathbf{x}_t \in R_i^0)}[\mathbf{1}_{(x_{td} \in (r_{j-1}, r_j))} - \mathbf{1}_{(x_{td} \in (\rho_{j-1}, \rho_j))}].$$

Taking $z_t = (\mathbf{a}'\mathbf{x}_t)^2\mathbf{1}_{(\mathbf{x}_t \in R_i^0)}$ and applying Lemma 3.8, we have

$$\frac{1}{n}X_{ir}^{*}{}'X_{ir}^{*} = \frac{1}{n}X_{i\rho}^{*}{}'X_{i\rho}^{*} + o_p(1),\ i = 1, 2.$$

(ii) Take $z_t = \epsilon_t^2\mathbf{1}_{(\mathbf{x}_t \in R_i^0)}$. Lemma 3.8 implies the desired result.

(iii) Take $z_t = \mathbf{a}'\mathbf{x}_t\epsilon_t$ for any $\mathbf{a}$. Lemma 3.8 implies $\frac{1}{n}[X_{ir}^{*}{}'\epsilon_r^* - X_{i\rho}^{*}{}'\epsilon_\rho^*] = o_p(1)$. So, it suffices to show that $\frac{1}{n}X_{i\rho}^{*}{}'\epsilon_\rho^* = O_p(n^{-1/2})$. For any $\mathbf{a} \neq 0$,

$$\mathbf{a}'(\frac{1}{n}X_{i\rho}^{*}{}'\epsilon_\rho^*) = \frac{1}{n}\sum_{t=1}^{n}\mathbf{a}'\mathbf{x}_t\epsilon_t\mathbf{1}_{(\mathbf{x}_t \in R_i^0 \cap R_j)},$$

where $\{\mathbf{a}'\mathbf{x}_t\epsilon_t\mathbf{1}_{(\mathbf{x}_t \in R_i^0 \cap R_j)}\}$ is a martingale difference sequence. By the central limit theorem for a martingale difference sequence (Billingsley, 1968), $\mathbf{a}'(\frac{1}{n}X_{i\rho}^{*}{}'\epsilon_\rho^*) = O_p(n^{-1/2})$. ¶

Lemma 3.10 *Let $n(A) = \sum_{t=1}^{n}\mathbf{1}_{(\mathbf{x}_t \in A)}$ for any set A in the domain of $\mathbf{x}_1$. Then under the conditions of Theorem 3.7, for $j = 1, \cdots, 2L+2$,*

(i) $\frac{1}{n}n(\hat{R}_j) = \frac{1}{n}n(R_j) + o_p(1) = \frac{1}{2L+2} + o_p(1)$,

(ii) $\hat{\tilde{\beta}}_r = \hat{\tilde{\beta}}_\rho + o_p(1) = \tilde{\beta}_\rho + o_p(1)$, *where*

$$\hat{\tilde{\beta}}_r = (X_r^{*\prime}X_r^*)^{-}X_r^{*\prime}Y_n,$$

$$\hat{\tilde{\beta}}_\rho = (X_\rho^{*\prime}X_\rho^*)^{-}X_\rho^{*\prime}Y_n,$$

$$\tilde{\beta}_\rho = \{E[\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(\mathbf{x}_1 \in R_j)}]\}^{-1}E[y_1\mathbf{x}_1\mathbf{1}_{(\mathbf{x}_1 \in R_j)}].$$

(iii) $\frac{1}{n}[S_n(\hat{R}_j) - S_n(R_j)] = o_p(1)$ *and*

(iv) $S_n(\hat{R}_j)/n(\hat{R}_j) - S_n(R_j)/n(R_j) = o_p(1)$.

Proof With loss of generality, we can assume $P(R_j \bigcap R_i^0) > 0,\ i = 1,2$.

(i) Note that $\frac{1}{n}n(\hat{R}_j) - \frac{1}{n}n(R_j) = \frac{1}{n}\sum_{t=1}^{n}[\mathbf{1}_{(x_{td}\in(r_{j-1},r_j))} - \mathbf{1}_{(x_{td}\in(\rho_{j-1},\rho_j))}]$. By applying Lemma 3.8 with $z_t = 1$, we get $\frac{1}{n}n(\hat{R}_j) = \frac{1}{n}n(R_j) + o_p(1)$. By the strong law of large numbers for ergodic processes,

$$\frac{1}{n}n(R_j) = \frac{1}{n}\sum_{t=1}^{n}\mathbf{1}_{(\mathbf{x}_t\in R_j)} = E[\mathbf{1}_{(\mathbf{x}_t\in R_j)}] + o_p(1) = P(\mathbf{x}_t \in R_j) + o_p(1) = \frac{1}{2L+2} + o_p(1).$$

(ii) By the strong law of large numbers for ergodic sequence, $\frac{1}{n}X_\rho^{*\prime}X_\rho^* \xrightarrow{a.s.} E[\mathbf{x}_1\mathbf{x}_1'\ \mathbf{1}_{(\mathbf{x}_t\in R_j)}] > 0$ and $\frac{1}{n}X_\rho^{*\prime}Y_n \xrightarrow{a.s.} E[\mathbf{x}_1'y_1\mathbf{1}_{(\mathbf{x}_t\in R_j)}]$. Hence, $\hat{\tilde{\beta}}_\rho \xrightarrow{a.s.} \tilde{\beta}_\rho$ as $n \to \infty$.

Since

$$X_\rho^{*\prime}Y_n = X_{1\rho}^{*}{}'X_{1\rho}^{*}\tilde{\beta}_1^0 + X_{2\rho}^{*}{}'X_{2\rho}^{*}\tilde{\beta}_2^0 + X_\rho^{*\prime}\tilde{\epsilon}_\rho^*$$

and

$$X_r^{*\prime}Y_n = X_{1r}^{*}{}'X_{1r}^{*}\tilde{\beta}_1^0 + X_{2r}^{*}{}'X_{2r}^{*}\tilde{\beta}_2^0 + X_{1r}^{*}{}'\tilde{\epsilon}_1^*,$$

Lemma 3.9 (i) and (iii) imply

$$(\frac{1}{n}X_{ir}^{*}{}'X_{ir}^{*})^{-} - (\frac{1}{n}X_{i\rho}^{*}{}'X_{i\rho}^{*})^{-} = o_p(1),$$

$i = 1,2$ and

$$\begin{aligned}
&\frac{1}{n}X_r^{*\prime}Y_n - \frac{1}{n}X_\rho^{*\prime}Y_n \\
=&(\frac{1}{n}X_{1r}^{*}{}'X_{1r}^{*} - \frac{1}{n}X_{1\rho}^{*}{}'X_{1\rho}^{*})\tilde{\beta}_1^0 + (\frac{1}{n}X_{2r}^{*}{}'X_{2r}^{*} - \frac{1}{n}X_{2\rho}^{*}{}'X_{2\rho}^{*})\tilde{\beta}_2^0 + \frac{1}{n}(X_r^{*\prime}\epsilon_r^* - X_\rho^{*\prime}\epsilon_\rho^*) \\
=&o_p(1).
\end{aligned}$$

This implies $\frac{1}{n}X_r^{*\prime}Y_n = O_p(1)$ since $\frac{1}{n}X_\rho^{*\prime}Y_n = O_p(1)$. Thus,

$$\begin{aligned}
&\hat{\tilde{\beta}}_r - \hat{\tilde{\beta}}_\rho = (X_r^{*\prime}X_r^*)^{-}X_r^{*\prime}Y_n - (X_\rho^{*\prime}X_\rho^*)^{-}X_\rho^{*\prime}Y_n \\
=&[(\frac{1}{n}X_r^{*\prime}X_r^*)^{-} - (\frac{1}{n}X_\rho^{*\prime}X_\rho^*)^{-}]\frac{1}{n}X_r^{*\prime}Y_n + (\frac{1}{n}X_\rho^{*\prime}X_\rho^*)^{-}[\frac{1}{n}X_r^{*\prime}Y_n - \frac{1}{n}X_\rho^{*\prime}Y_n] \\
=&o_p(1)O_p(1) + O_p(1)o_p(1) = o_p(1).
\end{aligned}$$

(iii)

$$\frac{1}{n}S_n(\hat{R}_j) = \frac{1}{n}\|X_r^*\hat{\tilde{\beta}}_r - (X_{1r}^*\tilde{\beta}_1^0 + X_{2r}^*\tilde{\beta}_2^0 + \epsilon_r^*)\|^2$$

$$= \frac{1}{n}\|X_{1r}^*(\hat{\tilde{\beta}}_r - \tilde{\beta}_1^0) + X_{2r}^*(\hat{\tilde{\beta}}_r - \tilde{\beta}_2^0) + \epsilon_r^*\|^2$$

$$= (\hat{\tilde{\beta}}_r - \tilde{\beta}_1^0)'(\frac{1}{n}X_{1r}^{*}{}'X_{1r}^*)(\hat{\tilde{\beta}}_r - \tilde{\beta}_1^0)$$

$$+ (\hat{\tilde{\beta}}_r - \tilde{\beta}_2^0)'(\frac{1}{n}X_{2r}^{*}{}'X_{2r}^*)(\hat{\tilde{\beta}}_r - \tilde{\beta}_2^0)$$

$$+ \frac{1}{n}\epsilon_r^{*}{}'\epsilon_r^* + \frac{2}{n}\epsilon_r^{*}{}'[X_{1r}^*(\hat{\tilde{\beta}}_r - \tilde{\beta}_1^0) + X_{2r}^*(\hat{\tilde{\beta}}_r - \tilde{\beta}_2^0)].$$

By (ii) and Lemma 3.9 (iii), $\hat{\tilde{\beta}}_r = \tilde{\beta}_\rho + o_p(1)$ and $\frac{1}{n}\epsilon_r^{*}{}'X_{ir}^* = o_p(1),\ i = 1,2$. Thus,

$$\frac{1}{n}S_n(\hat{R}_j)$$

$$= (\tilde{\beta}_\rho - \tilde{\beta}_1^0)'(\frac{1}{n}X_{1r}^{*}{}'X_{1r}^*)(\tilde{\beta}_\rho - \tilde{\beta}_1^0) + (\tilde{\beta}_\rho - \tilde{\beta}_2^0)'(\frac{1}{n}X_{2r}^{*}{}'X_{2r}^*)(\tilde{\beta}_\rho - \tilde{\beta}_2^0) + \frac{1}{n}\epsilon_r^{*}{}'\epsilon_r^* + o_p(1).$$

Similarly,

$$\frac{1}{n}S_n(R_j)$$

$$= (\tilde{\beta}_\rho - \tilde{\beta}_1^0)'(\frac{1}{n}X_{1\rho}^{*}{}'X_{1\rho}^*)(\tilde{\beta}_\rho - \tilde{\beta}_1^0) + (\tilde{\beta}_\rho - \tilde{\beta}_2^0)'(\frac{1}{n}X_{2\rho}^{*}{}'X_{2\rho}^*)(\tilde{\beta}_\rho - \tilde{\beta}_2^0) + \frac{1}{n}\epsilon_\rho^{*}{}'\epsilon_\rho^* + o_p(1).$$

Hence, by Lemma 3.9 (i) and (ii),

$$\frac{1}{n}S_n(\hat{R}_j) - \frac{1}{n}S_n(R_j)$$

$$= (\tilde{\beta}_\rho - \tilde{\beta}_1^0)'[\frac{1}{n}X_{1r}^{*}{}'X_{1r}^* - \frac{1}{n}X_{1\rho}^{*}{}'X_{1\rho}^*](\tilde{\beta}_\rho - \tilde{\beta}_1^0)$$

$$+ (\tilde{\beta}_\rho - \tilde{\beta}_2^0)'[\frac{1}{n}X_{2r}^{*}{}'X_{2r}^* - \frac{1}{n}X_{2\rho}^{*}{}'X_{2\rho}^*](\tilde{\beta}_\rho - \tilde{\beta}_2^0) + [\frac{1}{n}\epsilon_r^{*}{}'\epsilon_r^* - \frac{1}{n}\epsilon_\rho^{*}{}'\epsilon_\rho^*] + o_p(1)$$

$$= o_p(1).$$

(iv) By (i) and (iii),

$$\frac{S_n(\hat{R}_j)}{n(\hat{R}_j)} - \frac{S_n(R_j)}{n(R_j)}$$

$$= \frac{S_n(\hat{R}_j)}{n} \cdot \frac{n}{n(\hat{R}_j)} - \frac{S_n(R_j)}{n} \cdot \frac{n}{n(R_j)} \xrightarrow{p} 0, \quad n \to \infty. \qquad ¶$$

Lemma 3.10 sets down the fundation for Theorem 3.7 and will be used repetedly in its proof.

Proof of Theorem 3.7 Let $d \neq d^0$. Suppose a linear model is fitted on $R_j^d = \{\mathbf{x}_1 : x_{1d} \in (\rho_{j-1}^d, \rho_j^d]\}$ with the mean squared error $\hat{\sigma}_j^2(d) = S_n(R_j^d)/n(R_j^d)$. Under the assumed conditions, Lemma 3.3′ and Lemma 3.10 (i) imply $\frac{1}{n(R_j^d)} S_n(R_j^d) - \frac{1}{n(R_j^d)}[S_n(R_j^d \cap R_1^0) + S_n(R_j^d \cap R_2^0)] \xrightarrow{a.s.} C_j$ for some $C_j > 0$. Proposition 3.1′ (i) and Lemma 3.2′ imply the second term on the LHS,

$$
\begin{aligned}
&\frac{1}{n(R_j^d)}[S_n(R_j^d \cap R_1^0) + S_n(R_j^d \cap R_2^0)] \\
=&\frac{1}{n(R_j^d)}[\sum_{i=1}^{2} \tilde{\epsilon}_n'(R_j^d \cap R_i^0)\tilde{\epsilon}_n(R_j^d \cap R_i^0) + O_p(\ln^2 n)] \\
=&\frac{1}{n(R_j^d)}\tilde{\epsilon}_n'(R_j^d)\tilde{\epsilon}_n(R_j^d) + O_p(\ln^2 n/n),
\end{aligned}
$$

which converges to σ_0^2 by the strong law of large numbers. Thus, $P(\hat{\sigma}_j^2(d) > \sigma_0^2 + C_j/2) \to 1$ as $n \to \infty$. Since this holds for every j, by Lemma 3.10 (iv)

$$
\begin{aligned}
&\frac{S_n(\hat{R}_{(j)}^d)}{n(\hat{R}_{(j)}^d)} \\
=& \sum_{k=1}^{2(L+1)} \frac{S_n(\hat{R}_k^d)}{n(\hat{R}_k^d)} \mathbf{1}_{(\hat{R}_{(j)}^d = \hat{R}_k^d)} \\
=& \sum_{k=1}^{2(L+1)} (\frac{S_n(R_k^d)}{n(R_k^d)} + o_p(1))\mathbf{1}_{(\hat{R}_{(j)}^d = \hat{R}_k^d)} \\
\geq& \sum_{k=1}^{2(L+1)} (\sigma_0^2 + C_k/2)\mathbf{1}_{(\hat{R}_{(j)}^d = \hat{R}_k^d)} + o_p(1) \\
\geq& \sigma_0^2 + C + o_p(1)
\end{aligned}
$$

for some $C > 0$. By Lemma (3.10) (i)

$$
\begin{aligned}
&\frac{n(\hat{R}_{(j)}^d)}{n} \\
=& \sum_{k=1}^{2(L+1)} \frac{n(\hat{R}_k^d)}{n} \mathbf{1}_{(\hat{R}_{(j)}^d = \hat{R}_k^d)} \\
=& \sum_{k=1}^{2(L+1)} (\frac{1}{2(L+1)} + o_p(1))\mathbf{1}_{(\hat{R}_{(j)}^d = \hat{R}_k^d)} \\
=& \frac{1}{2(L+1)} + o_p(1).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\frac{1}{n}\tilde{S}_n^d &= \frac{1}{n}\sum_{j=1}^{L+1} S_n(\hat{R}_{(j)}^d) \\
&= \sum_{j=1}^{L+1} \frac{n(\hat{R}_{(j)}^d)}{n}\frac{1}{n(\hat{R}_{(j)}^d)} S_n(\hat{R}_{(j)}^d) \\
&\geq \sum_{j=1}^{L+1} [\frac{1}{2(L+1)} + o_p(1)][\sigma_0^2 + C + o_p(1)] \\
&= \frac{1}{2}\sigma_0^2 + \frac{C}{2} + o_p(1).
\end{aligned}
$$

If $d = d^0$, there are at least $L+1$ R_j^d's, say, $R_{j_i}^d$, $i = 1, \cdots, L+1$, which are entirely embedded in one of the R_j^0's. By Proposition 3.1 and Lemma 3.2,

$$
\begin{aligned}
&\hat{\sigma}_{j_i}^2(d) \\
&= \frac{1}{n(R_{j_i}^d)} S_n(R_{j_i}^d) \\
&= \frac{1}{n(R_{j_i}^d)} [\tilde{\epsilon}_n'(R_{j_i}^d)\tilde{\epsilon}_n(R_{j_i}^d) - T_n(R_{j_i}^d)] \\
&= \frac{n}{n(R_{j_i}^d)} [\frac{1}{n}\tilde{\epsilon}_n'(R_{j_i}^d)\tilde{\epsilon}_n(R_{j_i}^d) + O_p(\ln^2 n/n)], \quad i = 1, \cdots, L+1.
\end{aligned}
$$

By Lemma 3.10 (i) and the strong law of large numbers, the RHS is $\sigma_0^2 + O_p(\ln^2 n/n)$. This and Lemma 3.10 (i), (iv) imply,

$$
\begin{aligned}
\frac{1}{n}\tilde{S}_n^{d^0} &= \frac{1}{n}\sum_{j=1}^{L+1} S_n(\hat{R}_{(j)}^{d^0}) \\
&\leq \frac{1}{n}\sum_{i=1}^{L+1} S_n(\hat{R}_{j_i}^{d^0}) \\
&= \sum_{i=1}^{L+1} \frac{n(\hat{R}_{j_i}^{d^0})}{n}\frac{1}{n(\hat{R}_{j_i}^{d^0})} S_n(\hat{R}_{j_i}^{d^0}) \\
&= \sum_{i=1}^{L+1} (\frac{1}{2(L+1)} + o_p(1))(\hat{\sigma}_{j_i}^2(d^0) + o_p(1)) \\
&= \sum_{i=1}^{L+1} (\frac{1}{2(L+1)} + o_p(1))(\sigma_0^2 + o_p(1)) \\
&= \frac{1}{2}\sigma_0^2 + o_p(1).
\end{aligned}
$$

So, with probability approaching 1, $\tilde{S}_n^{d^0} < \tilde{S}_n^d$ for $d \neq d^0$. ¶

Remark The number $2(L+1)$ in Theorem 3.7 is not necessary. Actually, all we need is a number larger than $(L+1)$. So $L+2$ will do. And with probability approaching 1, $S_n(\hat{R}^{d^0}_{(1)})$, the smallest of the $\{S_n(\hat{R}^{d^0}_j)\}$ will be one of those obtained from the data entirely contained in one regime. Hence, if we let $\tilde{S}^d_n = S_n(\hat{R}^d_{(1)})$, with probability approaching 1, $\tilde{S}^{d^0}_n < \tilde{S}^d_n$ for $d \neq d^0$. However, by changing $L+2$ and $S_n(\hat{R}^d_{(1)})$ to $2(L+1)$ and $\sum_{j=1}^{L+1} S_n(\hat{R}^d_{(j)})$ respectively, we expect that the chance of $\tilde{S}^d_n < \tilde{S}^{d^0}_n$ for any $d \neq d^0$ will be reduced for small sample size. In fact, this was shown by a simulation study we performed but have not included in this thesis for the sake of brevity. The rate of correct identification is significantly higher when $\sum_{j=1}^{L+1} S_n(\hat{R}^d_{(j)})$ is used. If the number of regimes is chosen to be too large, then the number of observations in each regime will be small and the variance of $\tilde{S}^d_n$ will increase. Hence, it will undermine our selection of d. Through our simulation, we found that $2(L+1)$ is a reasonable choice. In addition, with small sample size, one of $\hat{R}^d_{(1)} \cap R^0_i$ $(i = 1, 2)$ may have very few observations for some $d \neq d^0$. In such a case $S_n(\hat{R}^d_{(1)})$ is likely to be smaller than $S_n(\hat{R}^{d^0}_{(1)})$ by chance. Using $\sum_{j=1}^{L+1} S_n(\hat{R}^d_{(j)})$ may average out this effect.

3.3 A simulation study

In this section, simulations of model (3.1) are carried out to examine the performance of the proposed procedure under various conditions. Constrained by our computing power, we study only moderate sample sizes under the segmented regression setup with two to three dependence structures, that is, $l^0 = 1$ and 2, respectively.

Let $\{\epsilon_t\}$ be iid with mean 0 and variance σ_0^2 and $\mathbf{z}_t = (x_{t1}, \cdots, x_{tp})'$ so that $\mathbf{x}'_t = (1, \mathbf{z}'_t)$, where $\{x_{tj}\}$ are iid $N(0,4)$. Let $DE(0,\lambda)$ denote the double exponential distribution with mean 0 and variance $2\lambda^2$. For $d = 1$ and $\tau_1^0 = 1$, the following 5 sets of specifications of the model are used for reasons given below:

(a) $p = 2$, $\tilde{\beta}_1 = (0, 1, 1)'$, $\tilde{\beta}_2 = (1.5, 0, 1)'$, $\epsilon_t \sim N(0, 1)$;

(b) $p = 2$, $\tilde{\beta}_1 = (0, 1, 1)'$, $\tilde{\beta}_2 = (1.5, 0, 1)'$, $\epsilon_t \sim DE(0, 1/\sqrt{2})$;

(c) $p = 2$, $\tilde{\beta}_1 = (0, 1, 0)'$, $\tilde{\beta}_2 = (1, 1, 0.5)'$, $\epsilon_t \sim DE(0, 1/\sqrt{2})$;

(d) $p = 3$, $\tilde{\beta}_1 = (0, 1, 0, 1)'$, $\tilde{\beta}_2 = (1, 0, 0.5, 1)'$, $\epsilon_t \sim DE(0, 1/\sqrt{2})$;

(e) $p = 3$, $\tilde{\beta}_1 = (0, 1, 1, 1)'$, $\tilde{\beta}_2 = (1, 0, 1, 1)'$, $\epsilon_t \sim DE(0, 1/\sqrt{2})$.

From the theory in Section 3.1 we know that the least squares estimate, $\hat{\tau}_1$, is appropriate if the model is discontinuous at τ_1^0. To explore the behavior of $\hat{\tau}_1$ for moderate sized samples, Models (a)-(d) are chosen to be discontinuous. The noise term in Model (a) is chosen to be normal as a reference, normal noise being widely used in practice. However, our emphasis is on more general noise distributions. Because the double exponential distribution is commonly used in regression modeling and it has heavier tails than the normal distribution, it is used as the distribution of the noise in all other models. The deterministic part of Model (b) is chosen to be the same as that of Model (a) to make them comparable. Note that Models (a) and (b) have a jump of size 0.5 at $x_1 = \tau_1$ while $Var(\epsilon_1) = 1$, which is twice the jump size. Except for the parameter τ_1, our model selection method and estimation procedures work for both continuous and discontinuous models. Model (e) is chosen to be a continuous model to demonstrate the behavior of the estimates for this type of model.

In all, 100 replications are simulated with different sample sizes, 30, 50, 100 and 200. Although in some experiments, $L = 3$ was tried, the number of under- and over-estimated l^0 are the same as those obtained by setting $L = 2$. The number of cases where $\hat{l} = 3$ is only 1 or 2, out of 100 replications. This agrees with our intuition that, given a two-piece model, if a two-piece model is selected over a three-piece one, it is unlikely that a four-piece model will be selected over a two-piece one. Based on this experience, the results reported in Tables 3.1 and 3.2 are obtained by setting $L = 2$ to save some computational effort. The two constants δ_0 and c_0 in MIC are chosen as 0.1 and 0.299 respectively, as explained in Section 3.1.

The results are summarized in Tables 3.1 and 3.2. Table 3.1 contains the estimates of l^0, τ_1^0 and the standard error of the estimate of τ_1^0, $\hat{\tau}_1$, based on the MIC. A number of observations may be made about the results in the table.

(i) For sample sizes greater than 30, the MIC correctly identifies l^0 in most of the cases. Hence, for estimating l^0, the result seems satisfactory. Comparing Models (a) and (b), it seems that the distribution of the noise has a significant influence on the estimation of l^0, for sample sizes of 50 or less.

(ii) For smaller sample sizes, the bias of $\hat{\tau}_1$ is related to the shape of the underlying model. It is seen that the biases are positive for Models (a) and (b), and negative for the others. In an experiment where Models (a) and (b) are changed so that the jump size at $x_1 = \tau_1$ is -0.5, instead of 0.5, negative biases are observed for every sample size. These biases decrease as the sample size becomes larger.

(iii) The standard error of $\hat{\tau}_1$ is relatively large in all the cases considered. And, as expected, the standard error decreases as the sample size increases. This suggests that a large sample size is needed for a reliable estimate of τ_1^0. An experiment with sample size of 400 for a model similar to Model (e) is reported in Section 4.3. In that experiment the standard error of $\hat{\tau}_1$ is significantly reduced.

(iv) The choice of $\delta_0 = 0.1$ seems adequate for most of the models we experimented with since it does not generate a pattern, like always overestimating l for $n = 30$ and underestimating l for $n = 50$, or vice-versa.

By the continuity of Model (e), its identification is expected to be the most difficult of all the cases considered. The c_0 chosen above seems too big for this case, since the tendency toward underestimating l is obvious when the sample size is small. However, a more plausible explanation for this is that with the small sample size and the noise level, there is simply not enough information to reveal the underlying model. Therefore, choosing a lower dimensional model with positive probability may be appropriate by the principle of parsimony.

In summary, since the optimal selection of the penalty is model dependent for samples of moderate size, no optimal pair of (c_0, δ_0) can be recommended. On the other hand, our choice of δ_0 and c_0 shows a reasonable performance for the models we experimented with.

Table 3.2 shows the estimated values of the other parameters for the models in Table 3.1 for a sample size of 200. The results indicate that, in general, the estimated $\tilde{\beta}_j$'s and σ_0^2 are quite close to their true values even when $\hat{r}_1$ is inaccurate. So, for the purpose of estimating $\tilde{\beta}_j$'s and σ_0^2, and interpolation when the model is continuous, a moderate sized sample say of size 200 may be sufficient. When the model is discontinuous, interpolation near the threshold may not be accurate due to the inaccurate $\hat{r}_1$. A careful comparison of the estimates obtained from Models (a) and (b) shows that the estimation errors are generally smaller with normally distributed errors. The estimates of β_{20} have relatively larger standard errors. This is due to the fact that a small error in $\hat{\beta}_{21}$ would result in a relatively large error in $\hat{\beta}_{20}$.

To assess the performance of the MIC when $l^0 = 2$, and to compare it with the Schwarz Criterion (SC) as well as a criterion proposed by Yao (1989), simulations were done for a much simpler model with sample sizes up to $n = 450$. Here we adopt Yao's (1989) setup where an univariate piecewise constant model is to be estimated. Note that such a model is a special case of Model (3.1). Specifically, Yao's model is

$$y_t = \beta_j^0 + \epsilon_t \quad if\ x_t \in (\tau_{j-1}^0, \tau_j^0], \ \ j = 1, \cdots, l^0 + 1,$$

where x_t is set to be t/n for $t = 1, \cdots, n$, ϵ_t is iid with mean zero and finite $2m$th moment for some positive integer m. Yao shows that with $m \geq 3$, the minimizer of $log\hat{\sigma}_l^2 + l \cdot C_n/n$ is a consistent estimate of l^0 for $l \leq L$, the known upper bound of l^0, where $\{C_n\}$ is any sequence satisfying $C_n n^{-2/m} \to \infty$ and $C_n/n \to 0$ as $n \to \infty$. Four sets of specifications of this model are experimented with:

(f) $\tau_1^0 = 1/3,\ \tau_2^0 = 2/3,\ \beta_{10}^0 = 0,\ \beta_{20}^0 = 2,\ \beta_{30}^0 = 4,\ \epsilon_t \sim DE(0, 1/\sqrt{2})$;

(g) $\tau_1^0 = 1/3,\ \tau_2^0 = 2/3,\ \beta_{10}^0 = 0,\ \beta_{20}^0 = 2,\ \beta_{30}^0 = 4,\ \epsilon_t \sim t_7/\sqrt{1.4}$;

(h) $\tau_1^0 = 1/3,\ \tau_2^0 = 2/3,\ \beta_{10}^0 = 0,\ \beta_{20}^0 = 1,\ \beta_{30}^0 = -1,\ \epsilon_t \sim DE(0, 1/\sqrt{2})$; and

(i) $\tau_1^0 = 1/3,\ \tau_2^0 = 2/3,\ \beta_{10}^0 = 0,\ \beta_{20}^0 = 1,\ \beta_{30}^0 = -1,\ \epsilon_t \sim t_7/\sqrt{1.4}$,

where t_7 refers to the Student-t distribution with degree of freedom of 7.

In each of these cases the variances of ϵ_t are scaled to 1 so the noise levels are comparable. Note that for $\epsilon_t \sim t_7/\sqrt{1.4}$, $E(\epsilon_t^6) < \infty$ and $E|\epsilon_t^7| = \infty$. It barely satisfies Yao's (1989) condition with $m = 3$ and does not satisfy our exponential boundedness condition. In Yao's (1989) paper, $\{C_n\}$ is not specified, so we have to choose a $\{C_n\}$ satisfying the conditions. The simplest $\{C_n\}$ is $c_1 n^\alpha$. With $m = 3$, we have $n^{\alpha - 2/3} \to \infty$ implying $\alpha > 2/3$. (We shall call the criterion with such a C_n, YC, hereafter.) To reduce the potential risk of underestimating l^0, we round 2/3 up to 0.7 as our choice of α. The δ_0 and c_0 in MIC are chosen as 0.1 and 0.299 respectively, for the reasons previously mentioned. c_1 is chosen by the same method as we used to choose c_0, that is, forcing $\log n_0 = c_1 n_0^\alpha$ and solving for c_1. With $n_0 = 20$ and $\alpha = 0.7$, we get $c_1 = 0.368$.

The results for model selection are reported in Tables 3.3-3.4. Table 3.3 tabulates the empirical distributions of the estimated l^0 for different sample sizes. From the table, it is seen that for most cases, MIC and YC perform significantly better than SC. And with sample size of 450, MIC and YC correctly identify l^0 in more then 90% of the cases. For Models (f) and (g), which are more easily identified, YC makes more correct identifications than MIC. But for Models (h) and (i), which are harder to identify, MIC makes more correct identifications. From Theorem 3.1 and the remark after its proof, it is known that both MIC and YC are consistent for the models with double exponential noise. This theory seems to be confirmed by our simulation.

The effect on model selection of varying the noise distribution does not seem significant. This may be due to the scaling of the noises by their variances, since variance is more sensitive to tail probabilities compared to quantiles or mean absolute deviation. Because most people are familiar with the use of variance as an index of dispersion, we adopt it, although other measures may reveal the tail effect on model identification better for our moderate sample sizes. Table 3.4 shows the estimated thresholds and their standard deviations for Models (f), (g), (h), (i), conditional on $\hat{l} = l^0$. Overall, they are quite accurate, even when the sample size is 50. For Models (h) and (i), the accuracy of $\hat{\tau}_2$ is much better than that of $\hat{\tau}_1$, since τ_2 is much easier to identify by the model specification. In general, for models which are more difficult to identify, a larger sample size is needed to achieve the same accuracy.

Finally, the small sample performance of the two methods given in Section 2.2 for the identification of the segmentation variable is examined. The experiment is carried out for Models (b), (d) and (e). Among Models (a)-(e), Models (b) and (e) seem to be the most difficult in terms of identifying l^0, and are also expected to be difficult for identifying d. Note that for all the models considered, d is asymptotically identifiable w.r.t. any $L \geq 1$ by Corollary 2.2. For $L = 2$, 100 replications are simulated with sample sizes of 50, 100 and 200. With sample sizes of 100 and 200, both methods identify l^0 correctly in every case. With sample size of 50, the correct identification rate of Method 1 is 100% for Models (b), (d), and 96% for Model (e); for Method 2 the rates are 98, 94 and 88 for Models (b), (d) and (e), respectively. From these results, we observe that for sample sizes of 100 or more, the two methods perform very well. And for a sample size of 50, Method 1 performs better than Method 2. This suggests that if the sample size is small, Method 1 may be more reliable. Otherwise, Method 2 gives a good estimate with a high computational efficiency.

3.4 General remarks

In this chapter, we proved the consistency of the estimators given in Chapter 2. In addition, when the model is discontinuous at the thresholds, we proved that the estimated thresholds converge rapidly to their true values at the rate of $\ln^2 n/n$. Consequently, the estimated regression coefficients and the estimated variance of the noise are shown to have the same asymptotic distributions as in the case where the thresholds are known, under the specified conditions. We put emphasis on the case where the model is discontinuous for the following two reasons:

First, if the model is continuous at the thresholds, then we have for any $\mathbf{z} \in \mathbf{R}^p$ and $\mathbf{x}' = (1, \mathbf{z}')$, $\mathbf{x}'\tilde{\beta}_j^0 = \mathbf{x}'\tilde{\beta}_{j+1}^0$ if $x_d = \tau_j^0$, $j = 1, \cdots, l^0$. This implies for all j, $\sum_{i \neq d}(\beta_{(j+1)i}^0 - \beta_{ji}^0)x_i = \beta_{j0}^0 - \beta_{(j+1)0}^0 + (\beta_{jd}^0 - \beta_{(j+1)d}^0)\tau_j^0$. Since this holds for any $\mathbf{x}$ such that $x_d = \tau_j^0$, we can conclude that $\beta_{(j+1)i}^0 = \beta_{ji}^0$ for $i \neq 0, d$ and all j. By aggregating the data over x_d, we obtain an ordinary linear regression problem and, hence, β_{ji}^0, $(i \neq 0, d,\ j = 1, \cdots, l^0 + 1)$, can be estimated by least squares estimates with all the properties given by the classical theory. The residuals can then be used to fit a one-dimensional continuous piecewise linear model to estimate β_{ji}^0, $(i = 0, d,\ j = 1, \cdots, l^0 + 1)$. For this one-dimensional continuous problem, Feder (1975a) shows that the restricted (by continuity) least squares estimates of the thresholds and the regression coefficient are asymptotically normally distributed when the covariates are viewed as nonrandom. So the problem is essentially solved except for a few technical points. In the Appendix of this chapter, we shall use Feder's idea to show that for a multidimensional continuous model with random covariates, the unrestricted least squares estimates possess similar properties. That is, the $\{\hat{\hat{\beta}}_j\}$ are asymptotically normally distributed, and so are the thresholds estimates given by the $\{\hat{\hat{\beta}}_j\}$ instead of least squares.

Second, noting that continuity requires $\beta^0_{(j+1)i} = \beta^0_{ji}$ for $i \neq 0, d$ and all j, it would seem that a response surface over a multidimensional space will rarely be well approximated by such a continuous piecewise model.

Problems where the models are either continuous at all thresholds or discontinuous at all thresholds have now been solved. The next question is what if the model is continuous at some thresholds, and discontinuous at others. This problem can be treated as follows. First, decide if the model is continuous at each threshold. This can be done by comparing $\hat{\tau}_j$, the least squares estimate of τ^0_j, with $\hat{\hat{\tau}}_j$, the solution of $\hat{\beta}_{j0} - \hat{\beta}_{(j+1)0} = (\hat{\beta}_{(j+1)d} - \hat{\beta}_{jd})\tau_j$. By the established convergence of the $\hat{\hat{\beta}}_j$'s and the $\hat{\tau}_j$'s, if the model were discontinuous at τ^0_j, then $\hat{\tau}_j$ would converge to τ^0_j. Meanwhile, $\hat{\beta}_{ji}$ or $\hat{\beta}_{(j+1)i}$ would converge to different values for some $i \neq 0, d$ or $\hat{\hat{\tau}}_j$ would converge to some point different from τ^0_j, or both. Thus, a large difference between $\hat{\tau}_j$ and $\hat{\hat{\tau}}_j$ or between $\hat{\beta}_{ji}$ and $\hat{\beta}_{(j+1)i}$ for some $i \neq 0, d$ would indicate discontinuity. Then, by noting that Theorem 3.4 does not assume the model is discontinuous at all τ^0_j's, we see that $\hat{\tau}_j - \tau^0_j = O_p(\ln^2 n/n)$ for all τ^0_j's which are thresholds of model discontinuity. By the proof of Theorem 3.5, it is seen that these $\hat{\tau}_j$'s can replace the corresponding τ^0_j's without changing the asymptotic distributions of the other parameters. So, between each successive pair of thresholds at which the model is discontinuous, the asymptotic results for a continuous model can be applied. In summary, regardless of whether the model is continuous or not, we can always obtain estimates of τ_j^0's which converge to their true values no slower than $O_p(1/\sqrt{n})$, and the estimated regression coefficients always have asymptotically normal distributions.

Note that most results given in this chapter do not require that $\mathbf{x}_1$ have a joint density which is everywhere positive over its domain. Hence, one component of $\mathbf{x}_1$ could be a function of other components, as long as they are not collinear. In particular, $\mathbf{x}_1$ could be a basis of pth order polynomials.

Since our estimation procedure is computationally intensive, one may worry about its computational feasibility. However, we do not think this is a serious problem, especially with the ever growing speed of modern computers. The simulations reported in the last section are done with a Sparc 2 work station. Even with our inefficient program, which inverts an order n^3 $(p+1) \times (p+1)$ matrices, 100 runs for model (a) consumes only about 9 minutes of CPU time with a sample size of $n = 50$ and only about 35 minutes with $n = 100$. Hence, each run would consume approximately .35 minutes of CPU time if $n = 100$. A more efficient program is under development; it uses an iterative method to avoid matrix inversion. A preliminary test shows that, with the same problems mentioned above, the CPU time consumed by this program is about 15 and 40 seconds for $n = 50$ and 100, respectively. Hence, each run would only take a few seconds of CPU time. Unfortunately, further modifications are needed for the new program to counter the problem of error evolution for large sample size. Nevertheless, even with our inefficient program, we believe our procedure is computationally feasible if L is small and n is not too large (say, $L \leq 5$, $n \leq 1000$). And with a better program and a faster computer, the computation time could be substantially reduced, making much more complicated model fitting computationally feasible. Finally, as we mentioned in Section 3.1, the choice of δ_0 and c_0 in MIC needs further study.

3.5 Appendix: A discussion of the continuous model

In Section 3.1, we established the asymptotic normality of coefficient estimators for Model (3.1) when it is discontinuous at the thresholds. In this section, we shall establish the corresponding result for Model (3.1) when it is everywhere continuous. If *Assumptions 3.0-3.1* are assumed by Theorem 3.1, the attention can be restricted to $\{\hat{l} = l^0\}$. First, we shall show that the $\hat{\beta}_j$'s converge at a rate no slower than $O_p(n^{-1/2}\ln n)$ by a method similar to that of Feder (1975a). Now let

$$\theta = (\tilde{\beta}_1', \cdots, \tilde{\beta}_{l^0+1}')';$$

$$\theta^0 = (\tilde{\beta}_1^{0'}, \cdots, \tilde{\beta}_{l^0+1}^{0'})';$$

$$\xi = (\theta', \tau_1, \cdots, \tau_{l^0})';$$

$$\xi^0 = (\theta^{0'}, \tau_1^0, \cdots, \tau_{l^0}^0)';$$

$$\Xi = \{\xi : \tilde{\beta}_j \neq \tilde{\beta}_{j+1},\ j = 1, \cdots, l^0,\ -\infty < \tau_1 < \cdots < \tau_{l^0} < \infty\};$$
$$\mu(\xi; \mathbf{x}) = \mathbf{x}'[\sum_{j=1}^{l^0+1} \mathbf{1}_{(x_d \in (\tau_{j-1}, \tau_j])}\tilde{\beta}_j];$$

and

$$\mu(\xi; X_k) = (\mu(\xi; \mathbf{x}_1), \cdots, \mu(\xi; \mathbf{x}_k))',$$

where $X_k = (\mathbf{x}_1, \cdots, \mathbf{x}_k)'$. Assuming no measurement errors, Feder (1975a) seeks the values at which the response must be observed to uniquely determine the model over the domain of the covariate. To find these values, he introduces a concept of identifiability. We adapt his concept to our problem.

Definition For any $\xi^* = (\theta^{*'}, \tau_1^*, \cdots, \tau_{l^0}^*)' \in \Xi$, the parameter $\theta = (\tilde{\beta}_1', \cdots, \tilde{\beta}_{l^0+1}')'$ is identified at $\mu^* = \mu(\xi^*, X_k)$ by X_k if the equation $\mu(\xi; X_k) = \mu^*$ uniquely determines $\theta = \theta^*$.

Next we prove a lemma adapted from Feder (1975a). The proof follows that of Feder (1975a).

Lemma A3.1 *If θ is identified at $\mu^0 = \mu(\xi^0, X_k)$ by $X_k = (\mathbf{x}_1, \cdots, \mathbf{x}_k)$, then there exist neighborhoods, M, of $\mu(\xi^0, X_k)$ and T of X_k such that*

(a) for all (k-dimensional) vectors $\tilde{\mu} = (\mu_1, \cdots, \mu_k)' \in M$ and $(p+1) \times k$ matrices $X_k^* \in T$ such that $\tilde{\mu}$ can be represented as $\tilde{\mu} = \mu(\xi, X_k^*)$ for some $\xi \in \Xi$, θ is identified at $\tilde{\mu}$ by X_k^*; and

(b) the induced transformation $\theta = \theta(\tilde{\mu}; X_k^*)$ satisfies the Lipschitz condition $||\theta_1 - \theta_2|| \leq C||\tilde{\mu}_1 - \tilde{\mu}_2||$ for some constant $C > 0$, whenever $X_k^* \in T$ and $\tilde{\mu}_1 = \mu(\xi_1; X_k^*)$, $\tilde{\mu}_2 = \mu(\xi_2; X_k^*) \in M$.

Proof: Since θ is identified at μ^0 by X_k, it follows that for any possible choice of parameters $\tau_1, \cdots, \tau_{l^0}$ consistent with θ^0, for each j there must exist $p+1$ components of X_k, $\mathbf{x}_{j_1}, \cdots, \mathbf{x}_{j_{p+1}}$ such that $x_{j_i,d} \in (\tau_{j-1}, \tau_j] \cap (\tau_{j-1}^0, \tau_j^0]$, $i = 1, \cdots, p+1$, and the matrix $(\mathbf{x}_{j_1}, \cdots, \mathbf{x}_{j_{p+1}})$ is nonsingular. By continuity, the $\mathbf{x}_{j_i}$'s may be perturbed slightly without disturbing the nonsingularity of $(\mathbf{x}_{j_1}, \cdots, \mathbf{x}_{j_{p+1}})$. Assertions (a) and (b) follow directly from the properties of nonsingular linear transformations. (Recall that if $\mu = X\theta$ for a nonsingular X, then $\theta = X^{-1}\mu$ and hence $||\theta|| \leq tr(X^{-1\prime}X^{-1})||\mu||$). ¶

Remark It is clear from the proof that for a continuous model, it is necessary and sufficient to identify θ^0, that within each τ-partition, there are $p+1$ observations $(\mathbf{x}_{j_1}, \cdots, \mathbf{x}_{j_{p+1}})$ such that the matrix $X = (\mathbf{x}_{j_1}, \cdots, \mathbf{x}_{j_{p+1}})$ is of full rank. In particular, if $\mathbf{z}$ has a positive density over a neighborhood of τ_j^0 for each j, then with large n, a X_k exists such that θ is identified at $\mu(\xi^0; X_k)$ by X_k.

Another concept introduced by Feder (1975a) is called the *center of observations.* This concept is modified in the next definition to fit our multivariate setup.

Definition Let $\mathbf{z} = (x_1, \cdots, x_p)'$. $\mathbf{z}^0 = (x_1^0, \cdots, x_p^0)'$ is a center of observation if for any $\delta > 0$, both $P(\{\mathbf{z} : \|\mathbf{z} - \mathbf{z}^0\| < \delta,\ x_d \leq x_d^0\})$ and $P(\{\mathbf{z} : \|\mathbf{z} - \mathbf{z}^0\| < \delta,\ x_d \geq x_d^0\})$ are positive.

Remark For any $\alpha < \eta$, if constant vectors $\mathbf{z}_1, \cdots, \mathbf{z}_{p+1}$ are centers of observations such that $x_{td} \in (\alpha, \eta)$, $t = 1, \cdots, p+1$, and the matrix $X_{p+1} = (\mathbf{x}_1, \cdots, \mathbf{x}_{p+1})$ is of full rank where $\mathbf{x}_t = (1, \mathbf{z}_t')'$, by Lemma (A3.1) there exists a neighborhood, T, of X_{p+1}, such that $T \subset \{\mathbf{x} : \alpha < x_{td} < \eta\}$, $P(T) > 0$ and X_{p+1}^* is of full rank if $X_{p+1}^* \in T$. Hence, for any $\mathbf{a} \neq \mathbf{0}$ and random vector $\mathbf{x}$,

$$E[(\mathbf{a}'\mathbf{x})^2 \mathbf{1}_{(x_d \in (\alpha, \eta])}] \geq E[(\mathbf{a}'\mathbf{x})^2 \mathbf{1}_{(\mathbf{x} \in \mathbf{T})}] > 0$$

implying that $E[\mathbf{x}\mathbf{x}'\mathbf{1}_{(x_d \in (\alpha, \eta])}]$ is positive definite. Therefore, a sufficient condition for *Assumption 3.1* to hold is that for some $\delta \in (0, \min_{1 \leq j \leq l^0}(\tau_{j+1}^0 - \tau_j^0)/2)$, within each of $\{\mathbf{x} : x_d \in (\tau_j^0 - \delta, \tau_j^0)\}$ and $\{\mathbf{x} : x_d \in (\tau_j^0, \tau_j^0 + \delta)\}$ there are $p+1$ centers of observations forming a full rank matrix for every j. In particular, ordinal categorical covariates are allowed in this assumption.

Lemma A3.2 *(Feder, 1975a) Let $\mathcal{V}$ be an inner product space and $\mathcal{X}$, $\mathcal{Y}$ subspaces of $\mathcal{V}$. Suppose $\mathbf{x} \in \mathcal{X}$, $\mathbf{y} \in \mathcal{Y}$, and $\mathbf{x}^*$, $\mathbf{y}^*$ are the orthogonal projections of $\mathbf{x} + \mathbf{y}$ onto $\mathcal{X}$, $\mathcal{Y}$ respectively. If there exists an $\alpha < 1$ such that $\mathbf{x} \in \mathcal{X}$, $\mathbf{y} \in \mathcal{Y}$ implies $|\mathbf{x}'\mathbf{y}| \leq \alpha \|\mathbf{x}\| \|\mathbf{y}\|$, then $\|\mathbf{x} + \mathbf{y}\| \leq (\|\mathbf{x}^*\| + \|\mathbf{y}^*\|)/(1 - \alpha)$.*

Lemma A3.3 *For any real $\tau_1 < \tau_1^0$, let $\hat{\mathcal{F}}$ be the random linear space spanned by the $2(p+1)$ column vectors of $(X_n(-\infty, \tau_1),\ X_n(\tau_1, \infty))$, and let $\zeta = X_n(\tau_1, \tau_1^0)\Delta\tilde{\beta}^0$, where $\Delta\tilde{\beta}^0 = \tilde{\beta}_1^0 - \tilde{\beta}_2^0$. Then under Assumptions 3.0-3.1, there exists $\alpha < 1$ such that for sufficiently large n,*

$$|\zeta' g| \leq \alpha \|\zeta\| \|g\|$$

uniformly in $\tau_1 < \tau_1^0$ and $g \in \hat{\mathcal{F}}$ with probability approaching 1.

Proof: It suffices to show that with large probability, for all $\tau_1 < \tau_1^0$ and $g \in \hat{\mathcal{F}}$,

$$(\zeta' g)^2 \leq \alpha^2 ||\zeta||^2 ||g||^2.$$

Define $\hat{X}_1 = X_n(-\infty, \tau_1)$, $\hat{X}_2 = X_n(\tau_1, \infty)$, $X_1^* = X_n(-\infty, \tau_1^0)$, $X_2^* = X_n(\tau_1^0, \infty)$. For any $g \in \hat{\mathcal{F}}$, there exist $\tilde{\beta}_1, \tilde{\beta}_2 \in \mathbf{R}^{p+1}$ such that $g = \hat{X}_1\tilde{\beta}_1 + \hat{X}_2\tilde{\beta}_2$. Noting that $||X_n(\tau_1, \tau_1^0)\tilde{\beta}_2||^2 \leq ||\hat{X}_2\tilde{\beta}_2||^2$, we have

$$\begin{aligned}\frac{||X_n(\tau_1,\tau_1^0)\tilde{\beta}_2||^2}{||g||^2} &= \frac{||X_n(\tau_1,\tau_1^0)\tilde{\beta}_2||^2}{||\hat{X}_1\tilde{\beta}_1||^2 + ||\hat{X}_2\tilde{\beta}_2||^2} \\ &\leq \frac{||X_n(\tau_1,\tau_1^0)\tilde{\beta}_2||^2}{||\hat{X}_2\tilde{\beta}_2||^2} \\ &\leq \frac{||X_n(\tau_1,\tau_1^0)\tilde{\beta}_2||^2 + ||\hat{X}_1\tilde{\beta}_2||^2}{||\hat{X}_2\tilde{\beta}_2||^2 + ||\hat{X}_1\tilde{\beta}_2||^2} \\ &= \frac{||X_1^*\tilde{\beta}_2||^2}{||X_n\tilde{\beta}_2||^2}.\end{aligned} \tag{A3.1}$$

Suppose A, B are positive definite matrices and $\lambda(M)$ denotes the largest eigenvalue of any symmetric matrix M. Then for any $\tilde{\beta} \neq 0$,

$$\frac{\tilde{\beta}' A \tilde{\beta}}{\tilde{\beta}'(A+B)\tilde{\beta}} = \frac{(B^{1/2}\tilde{\beta})'(B^{-1/2}AB^{-1/2})(B^{1/2}\tilde{\beta})}{(B^{1/2}\tilde{\beta})'(B^{-1/2}AB^{-1/2})(B^{1/2}\tilde{\beta}) + (B^{1/2}\tilde{\beta})'(B^{1/2}\tilde{\beta})} \leq \frac{\lambda(B^{-1/2}AB^{-1/2})}{\lambda(B^{-1/2}AB^{-1/2}) + 1}.$$

This result can be applied to the RHS of (A3.1) since $X_n'X_n = X_1^{*\prime}X_1^* + X_2^{*\prime}X_2^*$ and with probability approaching 1, $X_1^{*\prime}X_1^*$, $X_2^{*\prime}X_2^*$ are positive definite. Thus,

$$\frac{||X_1^*\tilde{\beta}_2||^2}{||X_n\tilde{\beta}_2||^2} = \frac{\tilde{\beta}_2'(\frac{1}{n}X_1^{*\prime}X_1^*)\tilde{\beta}_2}{\tilde{\beta}_2'(\frac{1}{n}X_1^{*\prime}X_1^* + \frac{1}{n}X_2^{*\prime}X_2^*)\tilde{\beta}_2} \leq \frac{\lambda_1}{\lambda_1 + 1}, \tag{A3.2}$$

where

$$\lambda_1 = \lambda((\frac{1}{n}X_2^{*\prime}X_2^*)^{-1/2}(\frac{1}{n}X_1^{*\prime}X_1^*)(\frac{1}{n}X_2^{*\prime}X_2^*)^{-1/2})$$

is bounded in probability since both $\frac{1}{n}X_1^{*\prime}X_1^*$ and $\frac{1}{n}X_2^{*\prime}X_2^*$ converge to positive definite matrices. Therefore, by (A3.1) and (A3.2) there exists $0 < \alpha < 1$ such that with probability approaching 1,

$$\frac{\|X_n(\tau_1,\tau_1^0)\tilde{\beta}_2\|^2}{\|g\|^2} \le \alpha^2 < 1$$

for all $\tau_1 < \tau_1^0$ and $g \in \hat{\mathcal{F}}$. Thus, with probability approaching 1,

$$\begin{aligned}
(\zeta' g)^2 =&(\sum_{t=1}^{n}(\Delta\tilde{\beta}^{0'}\mathbf{x}_t)(\mathbf{x}_t'\tilde{\beta}_2)\mathbf{1}_{(x_{td}\in(\tau_1,\tau_1^0])})^2 \\
\le&[\sum_{t=1}^{n}(\Delta\tilde{\beta}^{0'}\mathbf{x}_t)^2\mathbf{1}_{(x_{td}\in(\tau_1,\tau_1^0])}][\sum_{t=1}^{n}(\mathbf{x}_t'\tilde{\beta}_2)^2\mathbf{1}_{(x_{td}\in(\tau_1,\tau_1^0])}] \\
=&\|\zeta\|^2\|X_n(\tau_1,\tau_1^0)\tilde{\beta}_2\|^2 \\
=&\|\zeta\|^2\|g\|^2\frac{\|X_n(\tau_1,\tau_1^0)\tilde{\beta}_2\|^2}{\|g\|^2} \\
\le&\alpha^2\|\zeta\|^2\|g\|^2
\end{aligned}$$

for all $\tau_1 < \tau_1^0$ and $g \in \hat{\mathcal{F}}$. This completes the proof. ¶

Lemma A3.4 *Suppose Assumptions 3.0-3.1 are satisfied. Let W be a subset of $\mathbf{R}^p$ such that $P(W) > 0$. Then under Assumptions 3.0-3.1,* $\min_{\mathbf{x}_t\in W}|\hat{\nu}(\mathbf{x}_t)| = O_p(\ln n/\sqrt{n})$, *where* $\hat{\nu}(\mathbf{x}_t) = \mu(\hat{\xi};\mathbf{x}_t) - \mu(\xi^0;\mathbf{x}_t)$.

Proof Without loss of generality, we can assume $l^0 = 1$.

If we can show that $\sum_{t=1}^{n}\hat{\nu}_t^2(\mathbf{x}_t) = O_p(\ln^2 n)$, then for any $W \subset \mathbf{R}^p$ such that $P(W) > 0$, $\min_{\mathbf{x}_t\in W}|\hat{\nu}(\mathbf{x}_t)| = O_p(\ln n/\sqrt{n})$.

Let $\hat{\mathcal{F}}$ be the linear space spanned by the $2(p+1)$ column vectors of $(X_n(-\infty,\hat{\tau}_1)$, $X_n(\hat{\tau}_1,\infty))$, $[\mu(\xi^0;X_n)]$ be the linear space spanned by $\mu(\xi^0;X_n)$, and $\hat{\mathcal{F}}^+ = \hat{\mathcal{F}} \oplus [\mu(\xi^0;X_n)]$ be the direct sum of the two vector spaces. Let $\hat{Q}^+,\hat{Q}$ denote the orthogonal projections onto $\hat{\mathcal{F}}^+,\hat{\mathcal{F}}$ respectively. Let $\hat{\nu}(X_n) = (\hat{\nu}(\mathbf{x}_1),\cdots,\hat{\nu}(\mathbf{x}_n))'$. Then $\|\hat{\nu}(X_n) - \tilde{\epsilon}_n\|^2 = S_n(\hat{\tau}_1) \le \|\tilde{\epsilon}_n\|^2$.

Since both $\mu(\xi^0, X_n)$ and $\mu(\hat{\xi}; X_n)$ belong to $\hat{\mathcal{F}}^+$, by orthogonality,

$$\begin{aligned}
&||\mu(\hat{\xi}; X_n) - \hat{Q}^+ Y_n||^2 + ||\hat{Q}^+ Y_n - Y_n||^2 \\
=&||\mu(\hat{\xi}; X_n) - Y_n||^2 \\
=&||\hat{\nu}(X_n) - \tilde{\epsilon}_n||^2 \\
\leq&||\tilde{\epsilon}_n||^2 \\
=&||\mu(\xi^0; X_n) - Y_n||^2 \\
=&||\mu(\xi^0; X_n) - \hat{Q}^+ Y_n||^2 + ||\hat{Q}^+ Y_n - Y_n||^2.
\end{aligned}$$

Subtracting $||\hat{Q}^+ Y_n - Y_n||^2$ from both sides, we have that

$$\begin{aligned}
&||\mu(\hat{\xi}; X_n) - \hat{Q}^+ Y_n||^2 \\
\leq&||\mu(\xi^0; X_n) - \hat{Q}^+ Y_n||^2 \\
=&||\mu(\xi^0; X_n) - \hat{Q}^+ \mu(\xi^0; X_n) - \hat{Q}^+ \tilde{\epsilon}_n||^2 \\
=&||\hat{Q}^+ \tilde{\epsilon}_n||^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
||\hat{\nu}(X_n)|| =&||\mu(\hat{\xi}; X_n) - \mu(\xi^0; X_n)|| \\
\leq&||\mu(\hat{\xi}; X_n) - \hat{Q}^+ Y_n|| + ||\hat{Q}^+ Y_n - \mu(\xi^0; X_n)|| \\
\leq&||\hat{Q}^+ \tilde{\epsilon}_n|| + ||\hat{Q}^+ \tilde{\epsilon}_n|| \\
=&2||\hat{Q}^+ \tilde{\epsilon}_n||.
\end{aligned}$$

Since $\sum_{t=1}^n \hat{\nu}_t^2(\mathbf{x}_t) = \|\hat{\nu}(X_n)\|^2$, it remains to show that $\|\hat{Q}^+ \tilde{\epsilon}_n\| = O_p(lnn)$. Without loss of generality, we can assume that $\hat{\tau}_1 < \tau_1^0$. Let $\tilde{\beta}^0 = (\tilde{\beta}_1^{0'}, \tilde{\beta}_2^{0'})'$ and $\Delta\tilde{\beta}^0 = \tilde{\beta}_1^0 - \tilde{\beta}_2^0$. Note that

$$\begin{aligned}
&\mu(\xi^0, X_n) \\
=&(X_n(-\infty, \tau_1^0), X_n(\tau_1^0, \infty))\tilde{\beta}^0 \\
=&(X_n(-\infty, \hat{\tau}_1) + X_n(\hat{\tau}_1, \tau_1^0), X_n(\hat{\tau}_1, \infty) - X_n(\hat{\tau}_1, \tau_1^0))\tilde{\beta}^0 \\
=&[(X_n(-\infty, \hat{\tau}_1), X_n(\hat{\tau}_1, \infty)) + (X_n(\hat{\tau}_1, \tau_1^0), -X_n(\hat{\tau}_1, \tau_1^0))]\tilde{\beta}^0 \\
=&(X_n(-\infty, \hat{\tau}_1), X_n(\hat{\tau}_1, \infty))\tilde{\beta}^0 + X_n(\hat{\tau}_1, \tau_1^0)\Delta\tilde{\beta}^0.
\end{aligned}$$

This implies that $\hat{\mathcal{F}}^+$ is also generated by the direct sum of $\hat{\mathcal{F}}$ and vector ζ, where $\zeta = X_n(\hat{\tau}_1, \tau_1^0)\Delta\tilde{\beta}^0$.

By Lemma A3.3, there exists $\alpha < 1$ such that for sufficiently large n, $|\zeta' g| \leq \alpha \|\zeta\| \|g\|$ for all $\hat{\tau}_1 < \tau_1^0$ and $g \in \hat{\mathcal{F}}$ with probability approaching 1. Since $\hat{Q}(\hat{Q}^+\tilde{\epsilon}_n) = \hat{Q}\tilde{\epsilon}_n$ and $\zeta'(\hat{Q}^+\tilde{\epsilon}_n)/\|\zeta\| = \zeta'\tilde{\epsilon}_n/\|\zeta\|$, it follows from Lemma A3.2 that with probability approaching 1,

$$\|\hat{Q}^+\tilde{\epsilon}_n\| \leq \frac{1}{1-\alpha}\{\|\hat{Q}\tilde{\epsilon}_n\| + |\zeta'\tilde{\epsilon}_n|/\|\zeta\|\}.$$

Therefore, if it is shown that $\|\hat{Q}\tilde{\epsilon}_n\| = O_p(\ln n)$ and $\zeta'\tilde{\epsilon}_n/\|\zeta\| = O_p(\ln n)$, the desired result obtains. Define $\hat{X} = (\hat{X}_1, \hat{X}_2)$. Then

$$\begin{aligned}
\|\hat{Q}\tilde{\epsilon}_n\|^2 =& \tilde{\epsilon}_n'\hat{X}(\hat{X}'\hat{X})^-\hat{X}'\hat{X}(\hat{X}'\hat{X})^-\hat{X}'\tilde{\epsilon}_n \\
=& \tilde{\epsilon}_n'\hat{X}(\hat{X}'\hat{X})^-\hat{X}'\tilde{\epsilon}_n \\
=& \tilde{\epsilon}_n'\hat{X}_1(\hat{X}_1'\hat{X}_1)^-\hat{X}_1'\tilde{\epsilon}_n + \tilde{\epsilon}_n'\hat{X}_2(\hat{X}_2'\hat{X}_2)^-\hat{X}_2'\tilde{\epsilon}_n \\
=& T_n(-\infty, \hat{\tau}_1) + T_n(\hat{\tau}_1, \infty).
\end{aligned}$$

Therefore by Lemma 3.2, $\|\hat{Q}\tilde{\epsilon}_n\| = O_p(\ln n)$ uniformly for all $\hat{\tau}_1$.

We next show that uniformly in $\hat{\tau}_1 < \tau_1^0$, $\zeta'\tilde{\epsilon}_n/\|\zeta\| = O_p(\ln n)$ for $\|\zeta\| \neq 0$, where $\zeta = \xi(\tilde{\beta}_2^0 - \tilde{\beta}_1^0)$ and $\xi = (X_n(-\infty, \tau_1^0) - X_n(-\infty, \hat{\tau}_1))$. Let $y_t = x_t'\Delta\beta^0$. Conditional on X_n, we have that

$$\begin{aligned}
&P(\frac{|\zeta'\tilde{\epsilon}_n|}{\|\zeta\|} \geq \frac{3\ln n}{T_0}|X_n) \\
\leq &P(\max_{x_{ud}<x_{vd}} \frac{|\sum_{t=1}^n y_t \mathbf{1}_{(x_{ud}<x_{td}\leq x_{vd})}\epsilon_t|}{(\sum y_t^2 \mathbf{1}_{(x_{ud}<x_{td}\leq x_{vd})})^{1/2}} > \frac{3\ln n}{T_0}|X_n) \\
\leq &\sum_{x_{ud}<x_{vd}} P(\frac{|\sum_{t=1}^n y_t \mathbf{1}_{(x_{ud}<x_{td}\leq x_{vd})}\epsilon_t|}{(\sum y_t^2 \mathbf{1}_{(x_{ud}<x_{td}\leq x_{vd})})^{1/2}} > \frac{3\ln n}{T_0}|X_n),
\end{aligned}$$

where T_0 is specified in Lemma 3.1. Since $|y_s \mathbf{1}_{(x_{ud}<x_{sd}\leq x_{vd})}/(\sum_{t=1}^n y_t^2 \mathbf{1}_{(x_{ud}<x_{td}\leq x_{vd})})^{1/2}| \leq 1$ and

$$\sum_{s=1}^n [y_s \mathbf{1}_{(x_{ud}<x_{sd}\leq x_{vd})}/(\sum_{t=1}^n y_t^2 \mathbf{1}_{(x_{ud}<x_{td}\leq x_{vd})})^{1/2}]^2 = 1$$

for any x_{ud}, x_{vd}, by Lemma 3.1,

$$
\begin{aligned}
&P(\frac{|\zeta'\tilde{\epsilon}_n|}{\|\zeta\|} \geq \frac{3\ln n}{T_0}|X_n) \\
\leq &\sum_{x_{ud}<x_{vd}} 2\exp(-T_0 \cdot \frac{3\ln n}{T_0})\exp(c_0T_0^2) \\
\leq &n(n-1)/n^3\exp(c_0T_0^2) \longrightarrow 0,
\end{aligned}
$$

as $n \to \infty$, where c_0 is the constant specified in Lemma 3.1. Finally, by appealing to the dominated convergence theorem we obtain the desired result without conditioning.

This completes the proof. ¶

Theorem A3.1 *Suppose Assumptions 3.0 and 3.1 are satisfied. Let $X_k^0 = (\mathbf{x}_1^0, \cdots, \mathbf{x}_k^0)$. If θ is identified at $\mu(\xi^0, X_k^0)$ by X_k^0 and $\mathbf{x}_1^0, \cdots, \mathbf{x}_k^0$ are centers of observations, then*

$$\hat{\theta} - \theta^0 = O_p(\ln n/\sqrt{n}).$$

Proof Lemma A3.4 implies that with probability approaching 1, within any small neighborhood of $\mathbf{x}_i^0$, there exists a $\mathbf{x}_{t_i}$ such that

$$\mu(\hat{\xi}, \mathbf{x}_{t_i}) - \mu(\xi^0, \mathbf{x}_{t_i}) = O_p(\ln n/\sqrt{n}),$$

$i = 1, \cdots, k$. Lemma A3.1 implies the conclusion of the theorem. ¶

Corollary A3.1 *Under the conditions of Theorem A3.1, $\hat{\tau} - \tau^0 = O_p(\ln n/\sqrt{n})$ where $\hat{\tau} = (\hat{\tau}_1, \cdots, \hat{\tau}_{l^0})'$, $\hat{\tau}_j = (\hat{\beta}_{j,0} - \hat{\beta}_{j+1,0})/(\hat{\beta}_{j+1,d} - \hat{\beta}_{j,d})$, $j = 1, \cdots, l^0$.*

Proof For any $j = 1, \cdots, l^0$, by continuity of the model at the end points $x_d = \tau_j^0$,

$$\beta_{j0}^0 + \sum_{i\neq d}\beta_{ji}^0 x_i + \beta_{jd}^0\tau_j^0 = \beta_{j+1,0}^0 + \sum_{i\neq d}\beta_{j+1,i}^0 x_i + \beta_{j+1,d}^0\tau_j^0$$

for all $\{x_i,\ i \neq d\}$. Then by choosing the $\{x_i,\ i \neq d\}$ so that they are not collinear, we deduce that $\beta^0_{j+1,i} = \beta^0_{ji}$, for all $i \neq 0, d$. By assumption, $\beta^0_{jd} \neq \beta^0_{j+1,d}$. Therefore, τ_j can be reestimated by solving

$$(\hat{\beta}_{jd} - \hat{\beta}_{j+1,d})\hat{\tau}_j = (\hat{\beta}_{j+1,0} - \hat{\beta}_{j0}),$$

and hence, $\hat{\tau}_j - \tau^0_j$ has the same order as $\hat{\beta}_{j+1,0} - \hat{\beta}_{j,0}$. ¶

Next we shall establish the asymptotic normality of $\hat{\theta}$, and $\hat{\tau}$ when the model is continuous. The idea is to form a pseudo problem by deleting all the observations in a small neighborhood of each τ^0_j so that classical techniques can be applied, and then to show that the problem of concern is "close" to the pseudo problem. The term "pseudo problem" is used because in practice the τ^0_j's are unknown and so are the observations to be deleted. This idea is due to Sylwester (1965) and is used by Feder (1975a).

Assume x_d has positive density function $f_d(x_d)$ over a neighborhood of $\tau^0_j,\ j = 1, \cdots, l^0$. Our pseudo problem is formed by deleting all the observations in $\{\mathbf{x} : \tau^0_j - d_n \leq x_d \leq \tau^0_j + d_n\}$ where $d_n = 1/\ln^3 n$. Intuitively speaking, the number of observations deleted will be $O_p(nd_n)$. This will be confirmed later in Lemma A3.6. Adopting Feder's (1975a) notation, we define n^* as the sample size in the pseudo problem, and let $n^{**} = n - n^*$, $\hat{\theta}^*$ be the least squares estimate in the pseudo problem, $\sum^*$, the summation over the n^* terms of the pseudo problem, and $\sum^{**} = \sum_{t=1}^n - \sum^*$. Generally, a single asterisk refers to the pseudo problem.

Theorem A3.1 and Corollary A3.1 carry over directly to the pseudo problem. Thus,

Theorem A3.2 *If the conditions of Theorem A3.1 is satisfied in the pseudo problem, then*

$$\hat{\theta}^* - \theta^0 = O_p(\ln n/\sqrt{n}).$$

Further, if Model (3.1) is continuous, $\hat{\tau}^* - \tau^0 = O_p(\ln n/\sqrt{n})$.

Lemma A3.5 *Suppose $\{\mathbf{x_t}\}$ is an iid sequence. Under the conditions of Theorem A3.2*

$$\sqrt{n}(\hat{\tilde{\beta}}_j^* - \tilde{\beta}_j^0) \xrightarrow{d} N(0, \sigma_0^2 G_j^{-1}),$$

where $G_j = E[\mathbf{x}\mathbf{x}'\mathbf{1}_{(x_d \in (\tau_{j-1}^0, \tau_j^0])}]$, $j = 1, \cdots, l^0 + 1$.

Proof Let $S^*(\xi) = \frac{1}{n}\sum^*(y_t - \mu(\xi, \mathbf{x}_t))^2$. Theorem A3.2 implies that $\hat{\tau}_j^* \in (\tau_j^0 - d_n, \tau_j^0 + d_n]$ with probability approaching 1. Since there are no observations within this region, it follows that $S^*(\xi)$ computed within this region does not depend on τ and is a paraboloid in θ. In particular, it is twice differentiable in θ. For the reminder of the proof, denote $S^*(\xi)$ by $S^*(\theta)$. Thus, with probability approaching 1, $\hat{\theta}^*$ may be obtained by setting the derivative of $S^*(\theta)$ to 0:

$$\begin{aligned}
0 =& \frac{\partial}{\partial \tilde{\beta}_j} S^*(\theta) = \frac{\partial}{\partial \tilde{\beta}_j}\{\frac{1}{n}\sum_{t=1}^{n}\sum_{j=1}^{l^0+1}(\mathbf{x}_t'(\tilde{\beta}_j - \tilde{\beta}_j^0) - \epsilon_t)^2 \mathbf{1}_{(x_{td} \in (\tau_{j-1}^0 + d_n, \tau_j^0 - d_n])}\} \\
=& \frac{2}{n}\sum_{t=1}^{n} \mathbf{x}_t(\mathbf{x}_t'(\tilde{\beta}_j - \tilde{\beta}_j^0) - \epsilon_t)\mathbf{1}_{(x_{td} \in (\tau_{j-1}^0 + d_n, \tau_j^0 - d_n])}.
\end{aligned}$$

Hence, $\frac{1}{n}\sum_{t=1}^{n} \mathbf{x}_t\mathbf{x}_t'\mathbf{1}_{(x_{td} \in (\tau_{j-1}^0 + d_n, \tau_j^0 - d_n])})(\hat{\tilde{\beta}}_j^* - \tilde{\beta}_j^0) = \frac{1}{n}\sum_{t=1}^{n} \mathbf{x}_t\epsilon_t\mathbf{1}_{(x_{td} \in (\tau_{j-1}^0 + d_n, \tau_j^0 - d_n])}$. By Lemma 3.6 and the strong law of large numbers,

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n} \mathbf{x}_t\mathbf{x}_t'\mathbf{1}_{(x_{td} \in (\tau_{j-1}^0 + d_n, \tau_j^0 - d_n])}) \\
=& \frac{1}{n}\sum_{t=1}^{n} \mathbf{x}_t\mathbf{x}_t'\mathbf{1}_{(x_{td} \in (\tau_{j-1}^0, \tau_j^0])} + o_p(1) \\
=& G_j + o_p(1),
\end{aligned}$$

where $G_j = E[\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{td} \in (\tau_{j-1}^0, \tau_j^0])}]$. Under the assumptions of the pseudo problem, G_j is positive definite. Thus,

$$\sqrt{n}(\hat{\tilde{\beta}}_j^* - \tilde{\beta}_j^0) = [G_j + o_p(1)]^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n} \mathbf{x}_t\epsilon_t\mathbf{1}_{(x_{td} \in (\tau_{j-1}^0 + d_n, \tau_j^0 - d_n])}.$$

The Lindeberg-Feller central limit theorem for double sequences implies the assertion of the lemma. ¶

It now remains to show that $\hat{\theta}$ in the original problem and $\hat{\theta}^*$ in the pseudo problem do not differ by too much. In fact, we shall show that $\hat{\theta} - \hat{\theta}^* = o_p(n^{-1/2})$ and hence that the two have the same asymptotic distribution.

Lemma A3.6 *Suppose Assumptions 3.0, 3.1 and 3.3 are satisfied. Then under the conditions of Theorem A3.2,* $\hat{\theta} - \hat{\theta}^* = o_p(n^{-1/2})$.

Proof The hypotheses imply that θ is identified at $\mu(\xi^0, X_k^0)$ both in original problem and in the pseudo problem, by some $X_k^0 = (\mathbf{x}_1^0, \cdots, \mathbf{x}_k^0)$, where $\mathbf{x}_1^0, \cdots, \mathbf{x}_k^0$ are centers of observations. It follows from Theorems A3.1 and A3.2 that $\hat{\theta} - \theta^0 = O_p(n^{-1/2} \ln n)$, and $\hat{\theta}^* - \theta^0 = O_p(n^{-1/2} \ln n)$.

Let $a_n = (\ln n)^{5/4}$ and $\mathcal{U}_n = \{\xi : |\theta - \theta^0| < a_n/\sqrt{n}, |\tau_j - \tau_j^0| < d_n,\ j = 1, \cdots, l^0\}$. Then $\hat{\xi}$ *and* $\hat{\xi}^*$ both lie in $\mathcal{U}_n$ with probability approaching 1. Note that function $S^*(\xi)$ depends only on θ for $\xi \in \mathcal{U}_n$, so that $S^*(\xi) = S^*(\theta)$. Recall that

$$S(\xi) = \frac{1}{n} \sum_{t=1}^{n} (\epsilon_t + \nu(\mathbf{x}_t))^2,$$

and

$$S^*(\xi) = \frac{1}{n} \sum^{*} (\epsilon_t + \nu(\mathbf{x}_t))^2.$$

Thus,

$$\begin{aligned} S(\xi) =& S^*(\xi) + \frac{1}{n} \sum^{**} (\epsilon_t + \nu(\mathbf{x}_t))^2 \\ =& S^*(\xi) + \frac{1}{n} \sum^{**} \epsilon_t^2 + \frac{2}{n} \sum^{**} \epsilon_t \nu(\mathbf{x}_t) + \frac{1}{n} \sum^{**} \nu^2(\mathbf{x}_t). \end{aligned} \tag{A3.3}$$

Without loss of generality, we can assume that $\mathbf{z}$ is bounded. It follows from the definition of $\mathcal{U}_n$ and the boundedness of $\mathbf{z}$, that

$$\sup_{\xi \in \mathcal{U}_n} \max_{x_{td} \in \bigcup_j (\tau_j^0 - d_n, \tau_j^0 + d_n]} |\nu(\xi; \mathbf{x}_t)| = O(a_n/\sqrt{n}).$$

Note that n^{**} is the (1, 1)th component of $\sum_{j=1}^{l^0} X_n'(\tau_j^0 - d_n, \tau_j^0 + d_n)X_n(\tau_j^0 - d_n, \tau_j^0 + d_n)$. By Lemma 3.7 (i), $n^{**} = O_p(nd_n)$. Thus,

$$\begin{aligned}
&\sup_{\xi \in \mathcal{U}_n} |\frac{1}{n}\sum^{**} \nu^2(\mathbf{x}_t)| \\
\le&(a_n^2/n)n^{**}/n \\
=&O_p(a_n^2 d_n/n) \\
=&o_p(\frac{1}{n}).
\end{aligned}$$

Also, for any $\delta > 0$ and $\xi \in \mathcal{U}_n$

$$\begin{aligned}
&P(|\sum^{**} \epsilon_t \nu(\xi; \mathbf{x}_t)| > \delta) \le \frac{1}{\delta^2} E(\sum^{**} \epsilon_t \nu(\xi; \mathbf{x}_t))^2 \\
=&\frac{1}{\delta^2} E[E(\sum^{**} \epsilon_s \epsilon_t \nu(\xi; \mathbf{x}_s)\nu(\xi; \mathbf{x}_t)|X_n)] \\
\le&\frac{\sigma_0^2}{\delta^2} E[\sum^{**} \nu^2(\xi; \mathbf{x}_t)] \\
\le&\frac{\sigma_0^2}{\delta^2} (\sup_{\xi \in \mathcal{U}_n} \max_{x_{td} \in \bigcup_j (\tau_j^0 - d_n, \tau_j^0 + d_n]} |\nu(\xi; \mathbf{x}_t)|)^2 E(n^{**}) \\
\le&\frac{\sigma_0^2}{\delta^2} O(\frac{a_n^2}{n}) O_p(nd_n)
\end{aligned}$$

for some $M > 0$, where $O(a_n^2/n)$ and $O_p(nd_n)$ are independent of $\xi \in \mathcal{U}_n$. Since $a_n^2 d_n \to 0$ *as* $n \to \infty$, $\frac{1}{n}\sum^{**} \epsilon_t \nu(\xi; \mathbf{x}_t) = o_p(1/n)$ uniformly for all $\xi \in \mathcal{U}_n$. Thus, by (A3.3)

$$S(\xi) = S^*(\xi) + \frac{1}{n}\sum^{**} \epsilon_t^2 + o_p(\frac{1}{n}), \tag{A3.4}$$

where $o_p(1/n)$ is uniformly small for $\xi \in \mathcal{U}_n$.

Since $\hat{\xi}$ *and* $\hat{\xi}^*$ are least squares estimates for the original and the pseudo problem respectively,

$$S(\hat{\xi}) \le S(\hat{\xi}^*),\ S^*(\hat{\xi}^*) \le S^*(\hat{\xi}). \tag{A3.5}$$

(A3.4) and (A3.5) imply

$$0 \le S(\hat{\xi}^*) - S(\hat{\xi}) = S^*(\hat{\xi}^*) - S^*(\hat{\xi}) + o_p(\frac{1}{n}) \le o_p(\frac{1}{n}). \tag{A3.6}$$

Therefore, $S^*(\hat{\xi}) - S^*(\hat{\xi}^*) = o_p(\frac{1}{n})$. Since $\partial S^*(\hat{\xi}^*)/\partial\theta = 0$ and $S^*(\xi)$ is a paraboloid in θ, Taylor's expansion yields

$$S^*(\hat{\xi}) = S^*(\hat{\xi}^*)+\frac{1}{2}(\hat{\theta} - \hat{\theta}^*)'[\frac{\partial^2 S^*(\hat{\xi}^*)}{\partial\theta\partial\theta}](\hat{\theta} - \hat{\theta}^*)'. \tag{A3.7}$$

Equations (A3.6) and (A3.7) imply $\hat{\theta} - \hat{\theta}^* = o_p(n^{-\frac{1}{2}})$. ¶

Lemma A3.6 implies that $\sqrt{n}(\hat{\theta} - \theta^0)$ and $\sqrt{n}(\hat{\theta}^* - \theta^0)$ have the same asymptotic distribution. Thus, by Lemma A3.5 we have

Theorem A3.3 *Suppose the conditions of Lemma A3.6 are satisfied. Then,*

$$\sqrt{n}(\hat{\beta}_j - \beta_j^0) \xrightarrow{d} N(\mathbf{0}, \sigma_0^2 G_j^{-1}), \qquad j = 1, \cdots, l^0 + 1$$

where G_j is defined in Lemma A3.5.

For any $j = 1, \cdots, l^0 + 1$, let

$$\Delta\beta_0^0 = \beta_{j,0}^0 - \beta_{j+1,0}^0,\ \Delta\beta_d^0 = \beta_{j,d}^0 - \beta_{j+1,d}^0$$

and

$$\Delta\hat{\beta}_0 = \hat{\beta}_{j,0} - \hat{\beta}_{j+1,0},\ \Delta\hat{\beta}_d = \hat{\beta}_{j,d} - \hat{\beta}_{j+1,d}.$$

Then $\tau_j^0 = \frac{-\Delta\beta_0^0}{\Delta\beta_d^0}$, $\hat{\tau}_j = \frac{-\Delta\hat{\beta}_0}{\Delta\hat{\beta}_d}$, hence,

$$\begin{aligned}
\hat{\tau}_j - \tau_j^0 =& \frac{-\Delta\hat{\beta}_0}{\Delta\hat{\beta}_d} - \frac{-\Delta\beta_0^0}{\Delta\beta_d^0} = \frac{-1}{\Delta\hat{\beta}_d}(\Delta\hat{\beta}_0 - \Delta\beta_0^0) - \Delta\beta_0^0(\frac{1}{\Delta\hat{\beta}_d} - \frac{1}{\Delta\beta_d^0}) \\
=& \frac{1}{-\Delta\hat{\beta}_d}(\Delta\hat{\beta}_0 - \Delta\beta_0^0) - \frac{\Delta\beta_0^0}{\Delta\hat{\beta}_d\Delta\beta_d^0}(\Delta\beta_d^0 - \Delta\hat{\beta}_d) \\
=& \frac{1}{-\Delta\hat{\beta}_d}(\Delta\hat{\beta}_0 - \Delta\beta_0^0) + \frac{\tau_j^0}{-\Delta\beta_d^0}(\Delta\hat{\beta}_d - \Delta\beta_d^0).
\end{aligned}$$

$$\sqrt{n}(\hat{\tau}_j - \tau_j^0) = \frac{1}{-\Delta\beta_d^0}\sqrt{n}(\Delta\hat{\beta}_0 - \Delta\beta_0^0) + \frac{\tau_j^0}{-\Delta\beta_d^0}\sqrt{n}(\Delta\hat{\beta}_d - \Delta\beta_d^0) + o_p(1).$$

So we have

Theorem A3.4 *Under the conditions of Theorem A3.3, if Model (3.1) is continuous, then* $(\hat{\tau}_j - \tau_j^0)$ *and* $\frac{1}{-\Delta\beta_d^0}(\Delta\hat{\beta}_0 - \Delta\beta_0^0) + \frac{\tau_j^0}{-\Delta\beta_d^0}(\Delta\hat{\beta}_d - \Delta\beta_d^0)$ *have the same asymptotic distribution,* $j = 1, \cdots, l^0 + 1$.

Chapter 4

SEGMENTED REGRESSION MODELS WITH HETEROSCEDASTIC AUTOCORRELATED NOISE

In this chapter, we consider the situation where the noise is autocorrelated and the noise levels are different in different regimes. Specifically, consider the model

$$y_t = \mathbf{x}_t'\tilde{\beta}_j + \sigma_j \epsilon_t, \;\; if \;\; x_{td} \in (\tau_{j-1}, \tau_j], \;\; j = 1, \ldots, l+1, t = 1, \ldots, n, \tag{4.1}$$

where $\epsilon_t = \sum_0^\infty \psi_i \zeta_{t-i}$, with $\sum_0^\infty |\psi_i| < \infty$. The $\{\zeta_t\}$ are iid, have mean zero, have variance σ_ζ^2, and are independent of the $\{\mathbf{x}_t\}$, $\mathbf{x}_t = (1, x_{t1}, \ldots, x_{tp})'$. And $-\infty = \tau_0 < \tau_1 < \cdots < \tau_{l+1} = \infty$, while the σ_j $(j = 1, \ldots, l+1)$ are positive parameters. We adopt the parametrization which forces $\sigma_\zeta^2 = 1/\sum_0^\infty \psi_i^2$ so that the $\{\epsilon_t\}$ have unit variances. Further, we assume that there exists a $\delta > 3/2$, $k_0 > 0$ such that $|\psi_i| \le k/(i+1)^\delta$ for all i. Note that this implies $\{\epsilon_t\}$ is a stationary ergodic process.

Estimation procedures are given in Section 4.1. In Section 4.2, it is shown that the asymptotic results obtained in Chapter 3 remain valid. Since a major part of the proofs formally resemble those in Chapter 3, all the proofs are put in Section 4.5 as an appendix. Simulation results are reported in Section 4.3. Section 4.4 contains some remarks.

4.1 Estimation procedures

With the notation introduced in Chapter 3, the model can be rewritten in the vector form,

$$Y_n = \sum_{j=1}^{l^0+1} X_n(\tau_{j-1}^0, \tau_j^0)\tilde{\beta}_j + \tilde{\epsilon}_n^{\sigma}, \tag{4.2}$$

where $\tilde{\epsilon}_n^{\sigma} := [\sum_{j=1}^{l^0+1} \sigma_j I_n(\tau_{j-1}^0, \tau_j^0)]\tilde{\epsilon}_n$.

All the parameters are estimated as in Chapter 2 except for the variances $\{\sigma_1^2, \ldots, \sigma_{l^0+1}^2\}$. These are estimated by

$$\hat{\sigma}_j^2 = S_n(\hat{\tau}_{j-1}, \hat{\tau}_j)/\hat{n}_j. \quad j = 1, \ldots, \hat{l} + 1,$$

where $\hat{n}_j$ is the number of observations falling in the jth estimated regime and $\hat{l}$ is the estimate of l^0 produced by the estimation procedure in Section 2.2. We shall see in the next section that the asymptotic results in Section 3.2 are essentially unchanged for this modification of the model.

After estimating $\tilde{\beta}_j$ and σ_j we may use the estimated residuals, $\hat{\epsilon}_t = (y_t - \mathbf{x}_t'\hat{\tilde{\beta}}_j)/\hat{\sigma}_j$, if $x_{td} \in (\hat{\tau}_{j-1}, \hat{\tau}_j]$, to estimate the parameters in the moving average model for the $\epsilon_t's$.

4.2 Asymptotic properties of the parameter estimates

To establish the asymptotic theory, we need to make some assumptions for Model (4.2). Below is a basic assumption which is assumed to hold throughout this section.

Assumption 4.0*:*

The $\{\mathbf{x}_t\}$ is a strictly stationary ergodic process with $E(\mathbf{x}_1'\mathbf{x}_1) < \infty$. The ϵ_t are given by $\epsilon_t = \sum_0^{\infty} \psi_i \zeta_{t-i}$, where $\psi_i \leq k_0/(i+1)^{\delta}$ for some $k_0 > 0$, $\delta > 3/2$ and all i, the $\{\zeta_t\}$ are iid, locally exponentially bounded random variables with mean zero, variance $\sigma_\zeta^2 = 1/\sum_{i=0}^{\infty} \psi_i^2$, and *are independent of the* $\{\mathbf{x}_t\}$. *For the number of threshold* l^0, *there exists a specified* L *such that* $l^0 \le L$. *Also, for any* $j = 1, \cdots, l^0$, $\tilde{\beta}_j^0 \ne \tilde{\beta}_{j+1}^0$.

Note that $\{\epsilon_t\}$ is a stationary ergodic process and each ϵ_t has unit variance. Additional assumptions analogous to those in Section 3.1 are also needed to establish the consistency of the estimates. For convenience, we restate *Assumptions 3.1-3.2* as *Assumptions 4.1-4.2*, respectively.

Assumption 4.1

There exists $\delta \in (0, \min_{1\le j\le l^0}(\tau_{j+1}^0 - \tau_j^0)/2)$ *such that both* $E\{\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_i^0-\delta,\tau_i^0])}\}$ *and* $E\{\mathbf{x}_1\mathbf{x}_1'$ $\mathbf{1}_{(x_{1d}\in(\tau_i^0,\tau_i^0+\delta])}\}$ *are positive definite for each of the true thresholds* $\tau_1^0, \ldots, \tau_{l^0}^0$.

Assumption 4.2

For any sufficiently small $\delta > 0$, $E\{\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_i^0-\delta,\tau_i^0])}\}$ *and* $E\{\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_i^0,\tau_i^0+\delta])}\}$ *are positive definite,* $i = 1, \cdots, l^0$. *Also,* $E(\mathbf{x}_1'\mathbf{x}_1)^u < \infty$ *for some* $u > 1$.

To establish the asymptotic normality for the $\hat{\beta}_j$'s and $\hat{\sigma}_j^2$'s, we need to establish it for the least squares estimates of the β_j's and σ_j^2's with l^0 and $\tau_1^0, \cdots, \tau_{l^0}^0$ known. To this end, we specify the probability structure of $\{\mathbf{x}_t\}$ and $\{\zeta_t\}$ explicitly.

If $(\Omega, \mathcal{F}, \mathcal{P})$ is a probability space, a measurable transformation $T : \Omega \to \Omega$ is said to be measure-preserving if $P(T^{-1}A) = P(A)$ for all $A \in \mathcal{F}$. If T is measure-preserving, a set $A \in \mathcal{F}$ is called invariant if $T^{-1}(A) = A$. The class $\mathcal{T}$ of all invariant sets is a sub-σ-field of $\mathcal{F}$, called the invariant $\sigma-$field, and T is said to be ergodic if all the sets in $\mathcal{T}$ have probability zero or one. (c.f. Hall and Heyde, 1980, P281.)

As Hall and Heyde point out (1980, P281): "Any stationary process $\{\mathbf{x}_n\}$ may be thought of as being generated by a measure-preserving transformation, in the sense that there exists a variable $\mathbf{x}$ defined on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$, and a measure-preserving map $T : \Omega \to \Omega$, such that the sequence $\{\mathbf{x}'_n\}$ defined by $\mathbf{x}'_0 = \mathbf{x}$ and $\mathbf{x}'_n(\omega) = \mathbf{x}(T^n\omega)$, $n \geq 1$, $\omega \in \Omega$ has the same distribution as $\{\mathbf{x}_n\}$." Therefore, we can assume that the stationary and ergodic sequence $\{\mathbf{x}_t, \zeta_t\}$ is generated by a measure preserving transformation T on a probability space without loss of generality.

Assumption 4.3

(A.4.3.1) Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space. Let $\{\mathbf{x}_t, \zeta_t\}_{t=-\infty}^{\infty}$ be the iid random sequence such that

(i) $\{\mathbf{x}_t\}$ and $\{\zeta_t\}$ are independent;

(ii) $(\mathbf{x}_t, \zeta_t) = (\mathbf{x}(T^t\omega), \zeta(T^t\omega))$, $\omega \in \Omega$, $t = 0, \pm 1, \cdots$, where T is an ergodic measure-preserving transformation and $(\mathbf{x}, \zeta)$ is a random variable defined on the probability space $(\Omega, \mathcal{F}, \mathcal{P})$; and

(iii) $E(\mathbf{x}'_1\mathbf{x}_1)^u < \infty$ for some $u > 2$.

(A.4.3.2) Within some small neighborhoods of the true thresholds, x_{1d} has a positive and continuous probability density function $f_d(\cdot)$ with respect to the one dimensional Lebesgue measure.

(A.4.3.3) There exists one version of $E[\mathbf{x}_1\mathbf{x}'_1 | x_{1d} = x]$ which is continuous within some neighborhoods of the true thresholds and that version has been adopted.

Consider the segmented linear regression model (4.2) of the previous section. Let $\hat{l}$ be the minimizer of $MIC(l)$.

Theorem 4.1 *For the segmented linear regression model (4.2) suppose Assumptions 4.0 and 4.1 are satisfied. Then $\hat{l}$ converges to l^0 in probability as $n \to \infty$.*

The next two theorems show that the estimates $\hat{\tau}$, $\hat{\beta}_j$ and $\hat{\sigma}_j^2$ are consistent, under *Assumptions 4.0 and 4.2.*

Theorem 4.2 *Assume for the segmented linear regression model (4.2) Assumptions 4.0 and 4.2 are satisfied. Then*

$$\hat{\tau} - \tau^0 = o_p(1),$$

where $\tau^0 = (\tau_1^0, \ldots, \tau_{l^0}^0)$ *and* $\hat{\tau} = (\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{l}})$ *is the least squares estimate of* τ^0 *based on* $l = \hat{l}$, *and* $\hat{l}$ *is a minimizer of* $MIC(l)$ *subject to* $l \leq L$.

Theorem 4.3 *If the marginal cdf* F_d *of* x_{1d} *satisfies Lipschitz Condition* $|F_d(x') - F_d(x'')| \leq C|x' - x''|$ *for some constant* C *at a small neighborhood of* $x_{1d} = \tau_j^0$ *for every* j, *then under the conditions of Theorem 4.2, the least squares estimates* $\hat{\tilde{\beta}}_j$ *and* $\hat{\sigma}_j^2$, $j = 1, \ldots, \hat{l} + 1$, *based on the estimates* $\hat{l}$ *and* $\hat{\tau}_j$*'s as defined in Section 2.2, are consistent.*

Next, we show that if Model (4.2) is discontinuous at τ_j^0 for some $j = 1, \cdots, l^0$, then the threshold estimates, $\hat{\tau}_j$, converge to the true thresholds, τ_j^0, at the rate of $O_p(\ln^2 n/n)$, and the least squares estimates of $\tilde{\beta}_j$ and σ_j^2 based on the estimated thresholds are asymptotically normally distributed.

Theorem 4.4 *Suppose for the segmented linear regression model (4.2) that Assumptions 4.0, 4.2 and 4.3 are satisfied. If* $P(\mathbf{x}_1'(\tilde{\beta}_{j+1} - \tilde{\beta}_j) \neq 0 | x_d = \tau_j^0) > 0$ *for some* $j = 1, \cdots, l^0$, *then*

$$\hat{\tau}_j - \tau_j^0 = O_p(\frac{\ln^2 n}{n}).$$

For $j = 1, \cdots, l^0 + 1$, let $\hat{\tilde{\beta}}_j$ be the least squares estimates of $\tilde{\beta}_j$ based on the estimates $\hat{l}$ and $\hat{\tau}_j$'s as defined in Section 2.2, and $\hat{\sigma}_j^2$ be as defined in Section 4.1. Define

$$G_j = E(\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_{j-1}^0, \tau_j^0])}),$$

$$\Sigma_j = \sigma_j^2[G_j^{-1} + 2\sum_{i=1}^{\infty}\gamma(i)G_j^{-1}E(\mathbf{x}_1\mathbf{1}_{(x_{1d}\in(\tau_{j-1}^0, \tau_j^0])})E(\mathbf{x}_1\mathbf{1}_{(x_{1d}\in(\tau_{j-1}^0, \tau_j^0])})'G_j^{-1}],$$

$$p_j = P(\tau_{j-1}^0 < x_{1d} \leq \tau_j^0)$$

and

$$v_j = p_j(1-p_j)E(\epsilon_1^4) + p_j^2[(\eta-3)\gamma^2(0) + 2\sum_{i=-\infty}^{\infty}\gamma^2(i)],$$

where $\gamma(i) = E(\epsilon_1\epsilon_{1+i})$, $\eta = \sigma_\zeta^4/E(\zeta_1^4)$ and $j = 1,\cdots,l^0+1$. Then, we have the following result.

Theorem 4.5 *Suppose for the segmented linear regression model (4.2) Assumptions 4.0, 4.2 and 4.3 are satisfied. If* $P(\mathbf{x}_1'(\tilde{\beta}_{j+1}-\tilde{\beta}_j) \neq 0 | x_d = \tau_j^0) > 0$ *for all* $j = 1,\cdots,l^0$, *then*

$$\sqrt{n}(\hat{\tilde{\beta}}_j - \tilde{\beta}_j) \xrightarrow{d} N(\mathbf{0}, \Sigma_j) \quad and \quad \sqrt{n}p_j(\hat{\sigma}_j^2 - \sigma_j^2) \xrightarrow{d} N(0, v_j\sigma_j^4),$$

as $n \to \infty$, $j = 1,\cdots,l^0+1$.

Note that if $\gamma(i) = 0$, $i > 0$, then $\sum_j = \sigma_0^2 G_j^{-1}$ as shown in Section 3.1. The next theorem shows that Method 1 of Section 2.2 for estimating d^0 produces a consistent estimate.

Theorem 4.6 *If* d^0 *is asymptotically identifiable w.r.t.* L, *then under the conditions of Theorem 4.1,* $\hat{d}$ *given in Method 1 of Section 2.2 satisfies* $P(\hat{d} = d^0) \longrightarrow 1$ *as* $n \longrightarrow \infty$.

Remark: Although the result of Theorem 3.7 is expected to carry over if $\sigma_j = \sigma$ for all j, it does not carry over in general. Hence, Method 2 given in Section 2.2 is not generally consistent. Below is a counterexample.

Example 4.1. Let $\mathbf{x} = (1, x_1, x_2)'$ where (x_1, x_2) is a random vector with domain $[0,6]\times[0,6]$. Divide the domain into six parts as shown in Figure 4.1. On each part, (x_1, x_2) is uniformly distributed with mass indicated in the figure. Let $d = 1$, $l^0 = 2$, $L = 2$ and $(\tau_1, \tau_2) = (0.5, 1)$. Hence, $R_1^0 = \{\mathbf{x} : 0 < x_1 \leq 0.5\}$, $R_2^0 = \{\mathbf{x} : 0.5 < x_1 \leq 1\}$ and $R_3^0 = \{\mathbf{x} : 1 < x_1 \leq 6\}$. The model is

$$y_t = 2 \times \mathbf{1}_{(\mathbf{x}_t \in R_2^0)} + \sigma_j\epsilon_t, \quad if\ \mathbf{x}_t \in R_j^0,$$

where the $\{\mathbf{x}_t\}$ are independent samples from the distribution of $\mathbf{x}$, the $\{\epsilon_t\}$ are iid $N(0,1)$ and independent of $\{\mathbf{x}_t\}$. Let $\sigma_1^2 = 1$ and $\sigma_2^2 = \sigma_3^2 = 10$. Define $R_j^i = \{\mathbf{x} : x_i \in (j-1, j]\}$, $i = 1, 2$, $j = 1, \cdots, 6$. It is easy to see that on each R_j^i, the mass is $1/6 = 1/(2L+2)$. Suppose we fit a constant on each of R_j^i. Let us calculate $AMSE(R_j^i)$, the asymptotic mean squared error on R_j^i. For $j > 1$, $AMSE(R_j^1) = \sigma_3^2 = 10$. And

$$AMSE(R_1^1) = \sigma_1^2 \times \frac{1}{2} + \sigma_2^2 \times \frac{1}{2} + B_1^2 = \frac{11}{2} + B_1^2,$$

where B_1 is the asymptotic mean bias. Observe that the marginal distribution of x_1 on $(0,1]$ is uniform and symmetric about $\tau_1 = 0.5$; hence $B_1 = 1$ and $AMSE(R_1^1) = 13/2 < 10$. Therefore, with probability approaching 1 as $n \longrightarrow \infty$, the MSE on R_1^1 will be chosen as the smallest MSE among those on R_j^1, $j = 1, \cdots, 6$.

For $i = 2$ and $j > 1$,

$$AMSE(R_j^2) = \sigma_1^2 \times \frac{1}{100} + \sigma_2^2 \times \frac{8}{100} + \sigma_3^2 \times \frac{91}{100} + B_2^2 \geq \frac{991}{100},$$

where B_2 represents the asymptotic mean bias on each of R_j^2, $j > 1$. The asymptotic mean squared error on R_1^2 should be no larger than the asymptotic mean squared error obtained by setting the model to 0:

$$AMSE(R_1^2) \leq \sigma_1^2 \times \frac{9}{20} + \sigma_2^2 \times \frac{2}{20} + \sigma_3^2 \times \frac{9}{20} + 2^2 \times \frac{2}{20} = \frac{127}{20} < \frac{991}{100}.$$

Thus, with large probability as $n \to \infty$, the MSE on R_1^2 will be chosen as the smallest MSE among those on R_j^2, $j = 1, \cdots, 6$. Since $AMSE(R_1^1) > AMSE(R_1^2)$, x_2, rather than x_1, will be chosen by Method 2 as the segmentation variable with probability approaching 1 as $n \to \infty$. ¶

4.3 A simulation study

In this section, simulation experiments involving model (4.2) are carried out to examine the small sample performance of our proposed procedures under various conditions. As in Section 3.3, segmented regression models with two to three regimes are investigated.

Let

$$\epsilon'_t = 0.7\epsilon'_{t-1} - 0.1\epsilon'_{t-2} + \zeta_t,$$

where the $\{\zeta_t\}$ are iid with a locally exponentially bounded distribution having zero means and unit variances. Note that the $\{\epsilon'_t\}$ can alternatively be defined by

$$(1 - \xi_1^{-1}B)(1 - \xi_2^{-1}B)\epsilon'_t = \zeta_t,$$

where B is the backward shift operator defined by $B^j\epsilon'_t = \epsilon'_{t-j}$, $j = 0, \pm1, \pm2, \cdots$, and $(\xi_1, \xi_2) = (2, 5)$. Since $|\xi_i| > 1$ for $i = 1, 2$, $\{\epsilon'_t\}$ is a causal AR(2) process. Hence, it can be written as $\epsilon'_t = \sum_{j=0}^{\infty} \psi_j\zeta_{t-j}$, where ψ_j is the coefficient of z^j in the polynomial, $\psi(z) = 1/[(1-\frac{1}{2}z)(1-\frac{1}{5}z)]$. Expanding $\psi(z)$, we get

$$\psi(z) = \sum_{i=0}^{\infty}(\frac{1}{2}z)^i \sum_{k=0}^{\infty}(\frac{1}{5}z)^k = \sum_{i=0}^{\infty}\sum_{k=0}^{\infty}(\frac{1}{2})^i(\frac{1}{5})^k z^{i+k}.$$

Let $j = i + k$, then

$$\psi(z) = \sum_{i=0}^{\infty}\sum_{j=i}^{\infty}(\frac{1}{2})^i(\frac{1}{5})^{j-i}z^j = \sum_{j=0}^{\infty}\sum_{i=0}^{j}(\frac{1}{2})^i(\frac{1}{5})^{j-i}z^j.$$

So

$$|\psi_j| = \sum_{i=0}^{j}(\frac{1}{2})^i(\frac{1}{5})^{j-i} \le \sum_{i=0}^{j}(\frac{1}{2})^j = (j+1)(\frac{1}{2})^j.$$

Thus for any $\delta > 3/2$, taking $k_0 > 0$ sufficiently large, we have $|\psi_j| \le k_0/(j+1)^\delta$. Let $\epsilon_t = \epsilon'_t/\sqrt{Var(\epsilon'_t)}$, so that $Var(\epsilon_t) = 1$ for all t. Then the $\{\epsilon_t\}$ satisfy the condition of Model (4.2) [In this case $\sqrt{Var(\epsilon'_t)} \doteq 1.33$ (c.f. Example 3.3.5, Brockwell and Davis, 1987)].

Let $\mathbf{z}_t = (x_{t1}, \cdots, x_{tp})'$ and $\mathbf{x}_t' = (1, \mathbf{z}_t')$, where $\{x_{tj}\}$ are iid $N(0,4)$. Let $DE(0,\lambda)$ denote the double exponential distribution with mean 0 and variance $2\lambda^2$. For $d = 1$ and $\tau_1^0 = 1$, the following 3 sets of model specifications are used:

(a′) $p = 2$, $\tilde{\beta}_1 = (0,1,1)'$, $\tilde{\beta}_2 = (1.5,0,1)'$, $\sigma_1 = 0.8$, $\sigma_2 = 1$, $\zeta_t \sim N(0,1)$,

(d′) $p = 3$, $\tilde{\beta}_1 = (0,1,0,1)'$, $\tilde{\beta}_2 = (1,0,0.5,1)'$, $\sigma_1 = 0.8$, $\sigma_2 = 1$, $\zeta_t \sim DE(0,1/\sqrt{2})$,

(e′) $p = 3$, $\tilde{\beta}_1 = (0,1,1,1)'$, $\tilde{\beta}_2 = (1,0,1,1)'$, $\sigma_1 = 0.8$, $\sigma_2 = 1$, $\zeta_t \sim DE(0,1/\sqrt{2})$.

Note that the regression coefficients in Models (a′), (d′) and (e′) are the same as those in Models (a), (d) and (e). Beyond the reasons given in Section 3.3, these models are selected so that the results in this section will be comparable to those in Section 3.3.

In all, 100 replications are simulated with different sample sizes, 50, 100 and 200. For the reason given in Section 3.3, the results reported in Tables 4.1 and 4.2 are obtained by setting $L = 2$ to save some computational effort. The two constants, δ_0 and c_0 in MIC, are chosen as 0.1 and 0.299 respectively, as explained in Section 3.1. Table 4.1 shows the estimates $\hat{l}$, $\hat{\tau}_1$ and its standard error, based on the MIC. The following observations derive from the table.

(i) For all models, in more than 90% of the cases l^0 is correctly identified. Hence, for estimating l^0 our results seem satisfactory. Comparing these results to those in Table 3.1, it seems that Models (a′), (d′) and (e′) are more difficult to identify than Models (a), (d) and (e).

(ii) As in Section 3.3, $\hat{\tau}_1$ seems biased for small sample size. This bias is related to the shape of the model. Note that the biases for Model (a′) are all positive and those for Model (d′) are all negative. These biases decrease as the sample size becomes larger.

(iii) The standard error of $\hat{\tau}_1$ is relatively large in all the cases considered. And, as expected, the standard error decreases as the sample size increases. This suggests that a large sample size is needed for reliable estimation of τ_1^0. An experiment of $n = 400$ is carried out for Model (e$'$). We again obtained correct identification in 99% of the cases. But the standard error of $\hat{\tau}_1$ reduces from 1.111 for $n = 200$ to 0.707 when $n = 400$.

(iv) A larger δ_0 may perform better in these cases, since there seems to be a tendency to over estimate l^0, especially as n becomes large. Because in practice, the model structure is unknown and one cannot choose the best (δ_0, c_0), we adopt the same values for these parameters as in Section 3.3.

Table 4.2 shows the estimated values of the other parameters for the models in Table 4.1 only for a sample size of 200. The results indicate that, except for β_{20}, the estimated $\tilde{\beta}_j$'s are quite close to their true values even when $\hat{\tau}_1$ is inaccurate. So, for the purpose of estimating the $\tilde{\beta}_j$'s, and interpolation when the model is continuous, a moderate sample size such as 200 may be sufficient. When the model is discontinuous, interpolation near the threshold may not be accurate due to the inaccurate $\hat{\tau}_1$. As we saw in Section 3.3, the estimates of β_{20} have relatively large standard errors. This is due to the fact that a small error in $\hat{\beta}_{21}$ would result in a relatively large error in $\hat{\beta}_{20}$. The relatively large error for $\hat{\sigma}_2^2$ may also be due to the inaccurate $\hat{\tau}_1$.

Simulations have also been carried out for a model with $l^0 = 2$. Specifically, the model is:

(j) $p = 2$, $\tilde{\beta}_1 = (1,1,0)'$, $\tilde{\beta}_2 = (0,0,1)$, $\tilde{\beta}_3 = (0.5,0,0.5)$, $\sigma_1 = 0.7$, $\sigma_2 = 0.8$, $\sigma_3 = 1$ $\tau_1^0 = -1$, $\tau_2^0 = 1$, $\zeta_t \sim DE(0, 1/\sqrt{2})$.

The results are reported in Tables 4.3-4.4. Table 4.3 tabulates the empirical distributions of the estimated l^0 for different sample sizes. With $n = 200$, l^0 is correctly identified 95 out 100 replications. The standard errors of $\hat{\tau}_j$ $(j = 1, 2)$ are relatively small indicating that the thresholds in this model are easier to identify. The $\hat{\tilde{\beta}}_j$'s and the $\hat{\sigma}_j^2$'s are given in Table 4.4. The results are similar to those in Table 4.2.

4.4 General remarks

In this chapter, we generalized the results in Chapter 3 to the case where the noise is heteroscedastic and autocorrelated. Although the ideas used in this generalization are the same as those of Chapter 3, it can be seen in Section 4.5 that a more technical analysis is required to prove these results. The simulation results given in the last section indicate that this model is in general more difficult to identify, compared with the model discussed in the last chapter.

There are several questions which need further investigation. First, can the residuals be used to estimate the ψ_i's in the moving average specification of the noise once the estimates of the regression coefficients are obtained? If so, what procedure should be used to reduce the impact of the bias in the estimated τ_j^0's? Once the ψ_i's are estimated, can the information obtained be used to reestimate the other parameters of the model to obtain better estimates? Second, the asymptotic distribution of the estimates given in this chapter are for discontinuous models. If the model were continuous, one could aggragate the data over the segmentation variable regions to obtain a linear regression problem. The β_{ji}'s ($i \neq 0, d$) can be estimated by least squares. The residuals can be then be used to estimate β_{j1}, β_{jd} and σ_j^2 ($j = 1, \cdots, l^0 + 1$) by least squares again in a one-dimensional segmented regression problem. A number of questions remain to be answered: Are these estimates consistent? What are their asymptotic distributions? If the parameters are estimated directly by least squares, are the estimates, unrestricted by continuity, consistent? What are their asymptotic distributions? Some of these problems will be discussed further in the next chapter as future research topics.

4.5 Appendix: Proofs

Although a major part of the proof appear to resemble those in Chapter 3, there are some extra difficulties resulted from the correlated errors. First, we have to show that the result of Lemma 3.2 still holds under dependent assumptions. This is accomplished in Lemmas 4.1 and 4.2. Second, the results of Lemma 3.7 have to be re-established by calculating the limits of sample moments. Third, we have to establish the asymptotic normality of the estimated regression coefficients and the variances of the errors for known thresholds. This is done in Lemmas 4.9 and 4.10 by using a central limit theorem for stationary processes.

The proof of Theorem 4.1 will be given after a series of related lemmas.

Lemma 4.1 (Susko, 1991) Suppose $|a_i| \leq k_0/i^\delta$ for some $k_0 > 0,\ \delta > 3/2$. Then $\sum_{i=1}^{\infty}(\sum_{l=1}^{\infty} |a_{l+i}|)^2 < \infty$.

Proof: By assumption, $|a_i| \leq k_0/i^\delta$ for some $k_0 > 0,\ \delta > 3/2$. Therefore,

$$\sum_{i=1}^{\infty}(\sum_{l=1}^{\infty}|a_{l+i}|)^2 \leq k_0^2 \sum_{i=1}^{\infty}(\sum_{l=1}^{\infty}\frac{1}{(i+l)^\delta})^2.$$

Now,

$$\begin{aligned}
\sum_{l=1}^{\infty}\frac{1}{(i+l)^\delta} &= \sum_{j=i+1}^{\infty} 1/j^\delta \\
&= \sum_{j=i+1}^{\infty} 1/j^\delta \int_{j-1}^{j} dt \\
&= \sum_{j=i+1}^{\infty} \int_{j-1}^{j} 1/j^\delta dt \\
&= \sum_{j=i+1}^{\infty} \int_{j-1}^{j} \min_{j-1\leq t\leq j} 1/t^\delta dt \\
&\leq \sum_{j=i+1}^{\infty} \int_{j-1}^{j} 1/t^\delta dt \\
&= \int_{i}^{\infty} 1/t^\delta dt \\
&= \frac{1}{(\delta-1)i^{\delta-1}}.
\end{aligned} \tag{4.3}$$

So,

$$\sum_{i=1}^{\infty}(\sum_{l=1}^{\infty}|a_{l+i}|)^2 \leq \frac{k_0^2}{(\delta-1)^2}\sum_{i=1}^{\infty} 1/i^{2(\delta-1)}.$$

By assumption, $\delta > 3/2$, so $2(\delta - 1) > 1$, and hence

$$\sum_{i=1}^{\infty}(\sum_{l=1}^{\infty}|a_{l+i}|)^2 < \infty. \quad \P$$

The next Lemma is slightly modified version of Lemma 1 of Susko (1991).

Lemma 4.2 *Let $\{\zeta_t\}$ be iid, locally exponentially bounded random variables. Let $\epsilon_t = \sum_{i=0}^{\infty} \psi_i \zeta_{t-i}$, and assume there exists $\delta > 3/2$, $k_0 > 0$ such that $|\psi_i| \leq k_0/(i+1)^{\delta}$ for all i. Let $S_k = \sum_{i=1}^{k} a_i \epsilon_i$, where the a_i's are constants. Then there exists $0 < c_1 < \infty$ and $T_1 > 0$, such that for any $x \geq 0$, $k \geq 1$ and t satisfying $0 < t||a|| \leq T_1$,*

$$P\{|S_k| \geq x\} \leq 2e^{-tx + c_1 t^2 \|a\|^2}.$$

Proof The assumption of locally exponentially boundedness means that for some $T_0 > 0$ and $0 < c_0 < \infty$, $E(e^{t\zeta_1}) \leq e^{c_0 t^2}$ for $|t| \leq T_0$. Now it follows from Markov's inequality that for sufficiently small $t > 0$,

$$P\{S_k \geq x\} = P\{e^{tS_k} \geq e^{tx}\} \leq e^{-tx}E(e^{tS_k}).$$

And

$$S_k = \sum_{1}^{k} a_i \epsilon_i = \sum_{i=1}^{k} a_i \sum_{j=0}^{\infty} \psi_j \zeta_{i-j} = A(k) + B(k),$$

where

$$A(k) = \sum_{i=0}^{k-1} \zeta_{k-i} \sum_{j=0}^{i} a_{k-j}\psi_{i-j},$$

$$B(k) = \sum_{i=0}^{\infty} \zeta_{-i} \sum_{j=1}^{k} a_j \psi_{j+i}.$$

Hence,

$$E(e^{tB(k)}) \leq e^{c_0 t^2 \sum_{i=0}^{\infty}(\sum_{l=1}^{k} a_l \psi_{l+i})^2},$$

if $|t\sum_{l=1}^{k} a_l\psi_{l+i}| \leq T_0$ for all i. Let $M_1 = \sum_{i=0}^{\infty}(\sum_{l=1}^{\infty}|\psi_{l+i}|)^2$. Note that we can assume $\sqrt{M_1} > 0$ without loss of generality (since otherwise $\epsilon_t \equiv 0$ $a.s.$). Since $|\psi_i| \leq k_0/(i+1)^{\delta}$, from the previous lemma $M_1 < \infty$. Observe that for all i,

$$
\begin{aligned}
(\sum_{l=1}^{k} a_l\psi_{l+i})^2 &\leq (\sum_{l=1}^{k} a_l^2)(\sum_{l=1}^{k} \psi_{l+i}^2) \\
&\leq ||a||^2(\sum_{l=1}^{k} |\psi_{l+i}|)^2 \leq ||a||^2(\sum_{l=1}^{\infty} |\psi_{l+i}|)^2.
\end{aligned}
$$

Hence if t is such that $|t|||a|| \leq T_0/\sqrt{M_1}$, then for all i

$$
|t\sum_{l=1}^{k} a_l\psi_{l+i}| \leq |t|||a||(\sum_{l=1}^{\infty} |\psi_{l+i}|) \leq |t|||a||\sqrt{M_1} \leq T_0.
$$

Therefore, for any t such that $|t|||a|| \leq T_0/\sqrt{M_1}$ and $c = c_0M_1$,

$$
\begin{aligned}
E(e^{tB(k)}) &\leq e^{c_0t^2\sum_{i=0}^{\infty}||a||^2(\sum_{l=1}^{\infty}|\psi_{l+i}|)^2} \\
&= e^{c_0\sum_{i=0}^{\infty}(\sum_{l=1}^{\infty}|\psi_{l+i}|)^2t^2||a||^2} \leq e^{ct^2||a||^2}.
\end{aligned}
$$

Also,

$$
E(e^{tA(k)}) \leq e^{c_0t^2\sum_{i=0}^{k-1}(\sum_{j=0}^{i} a_{k-j}\psi_{i-j})^2}
$$

if $|\sum_{j=0}^{i} a_{k-j}\psi_{i-j}| \le T_0$ for all i. Let $n = i-j,\ m = i-l$, then

$$\begin{aligned}
&\sum_{i=0}^{k-1}(\sum_{j=0}^{i} a_{k-j}\psi_{i-j})^2 \\
=&\sum_{i=0}^{k-1}[\sum_{j=0}^{i} a_{k-j}^2\psi_{i-j}^2 + 2\sum_{j=1}^{i}\sum_{l=0}^{j-1} a_{k-j}a_{k-l}\psi_{i-j}\psi_{i-l}] \\
=&\sum_{i=0}^{k-1}\sum_{n=i}^{0} a_{k-i+n}^2\psi_n^2 + 2\sum_{i=1}^{k-1}\sum_{n=0}^{i-1}\sum_{m=i}^{n+1} a_{k-(i-n)}a_{k-(i-m)}\psi_n\psi_m \\
=&\sum_{i=0}^{k-1}\sum_{n=0}^{i} a_{k-i+n}^2\psi_n^2 + 2\sum_{n=0}^{k-2}\sum_{i=n+1}^{k-1}\sum_{m=n+1}^{i} a_{k+n-i}a_{k+m-i}\psi_n\psi_m \\
\le&\sum_{n=0}^{k-1}\psi_n^2\sum_{i=n}^{k-1} a_{k-i+n}^2 + 2|\sum_{n=0}^{k-2}\sum_{m=n+1}^{k-1}\sum_{i=m}^{k-1} a_{k+n-i}a_{k+m-i}\psi_n\psi_m| \\
\le&\sum_{n=0}^{k-1}\psi_n^2||a||^2 + 2\sum_{n=0}^{k-2}\sum_{m=n+1}^{k-1}|\psi_n\psi_m||\sum_{i=m}^{k-1} a_{k+n-i}a_{k+m-i}| \\
\le&\sum_{n=0}^{k-1}\psi_n^2||a||^2 + 2\sum_{n=0}^{k-2}\sum_{m=n+1}^{k-1}|\psi_n\psi_m|||a||^2 \\
=&||a||^2(\sum_{n=1}^{k-1}|\psi_n|)^2 \\
\le&||a||^2 M_1.
\end{aligned}$$

Therefore, for any t such that $|t|||a|| \le T_0/\sqrt{M_1}$ and the $c = c_0 M_1$, we have

$$\begin{aligned}
&|t\sum_{j=0}^{i} a_{k-j}\psi_{i-j}| \le |t||\sum_{j=0}^{i} a_{k-j}\psi_{i-j}| \\
\le&|t|[\sum_{i=0}^{k-1}(\sum_{j=0}^{i} a_{k-j}\psi_{i-j})^2]^{1/2} \le |t|||a||\sqrt{M_1} \\
\le&T_0,
\end{aligned}$$

and hence

$$E(e^{tA(k)}) \le e^{c_0 t^2||a||^2 M_1} = e^{ct^2||a||^2}.$$

Since $A(k)$ and $B(k)$ are independent we get that for $T_1 = T_0/\sqrt{M_1}$ and any k,

$$P\{S_k \ge x\} \le e^{-tx}E(e^{tA(k)})E(e^{tB(k)}) \le e^{-tx}e^{2ct^2||a||^2} = e^{-tx}e^{c_1 t^2||a||^2},$$

where $c_1 = 2c$ and $|t|||a|| \le T_1$.

Finally, to conclude the proof, we note that

$$P\{S_k \le -x\} = P\{-S_k \ge x\}. \qquad \P$$

Lemma 4.3 *Assume for the segmented linear regression model (4.2) that Assumption 4.0 is satisfied. Define* $\sigma_{\max} := \max_i \sigma_i$ *and redefine* $T_n(\alpha,\eta) := \tilde{\epsilon}_n^{\sigma\prime} H_n(\alpha,\eta)\tilde{\epsilon}_n^{\sigma}$, $-\infty \le \alpha < \eta \le \infty$. *Then*

$$P\{\sup_{\alpha<\eta} T_n(\alpha,\eta) \ge \frac{9\sigma_{\max}^2 p_0^3}{T_1^2}\ln^2 n\} \to 0, \qquad \textit{as } n \to 0,$$

where p_0 *is the true order of the model and* T_1 *is the constant specified in Lemma 4.2.*

Proof Conditioning on X_n, we have

$$\begin{aligned}
&P\{\sup_{\alpha<\eta} T_n(\alpha,\eta) \ge \frac{9\sigma_{\max}^2 p_0^3}{T_1^2}\ln^2 n \mid X_n\}\\
=&P\{\max_{x_{sd}<x_{td}} \tilde{\epsilon}_n^{\sigma\prime} H_n(x_{sd},x_{td})\tilde{\epsilon}_n^{\sigma} \ge \frac{9\sigma_{\max}^2 p_0^3}{T_1^2}\ln^2 n \mid X_n\}\\
\le& \sum_{x_{sd}<x_{td}} P\{\tilde{\epsilon}_n^{\sigma\prime} H_n(x_{sd},x_{td})\tilde{\epsilon}_n^{\sigma} \ge \frac{9\sigma_{\max}^2 p_0^3}{T_1^2}\ln^2 n \mid X_n\}.
\end{aligned}$$

Since $H_n(x_{sd},x_{td})$ is nonnegative definite and idempotent, it can be decomposed as $H_n(x_{sd},x_{td}) = W'\Lambda W$, where W is orthogonal and $\Lambda = diag(1,\cdots,1,0,\cdots,0)$ with $p := rank(H_n(x_{sd},x_{td})) = rank(\Lambda) \le p_0$. Set $Q = (I_p, \mathbf{0})W$. Then Q has full row rank p. Let $Q' = (\mathbf{q}_1,\cdots,\mathbf{q}_p)$ and $U_l = \mathbf{q}_l'\tilde{\epsilon}_n^{\sigma} = \mathbf{q}_l'[\sum_{i=1}^{l^0+1}\sigma_i I_n(\tau_{i-1}^0,\tau_i^0)]\tilde{\epsilon}_n$, $l = 1,\ldots,p$. Then,

$$\tilde{\epsilon}_n^{\sigma\prime} H_n(x_{sd},x_{td})\tilde{\epsilon}_n^{\sigma} = \tilde{\epsilon}_n^{\sigma\prime} Q'Q\tilde{\epsilon}_n^{\sigma} = \sum_{l=1}^{p} U_l^2.$$

Since $p \le p_0$, as in the proof of Lemma 3.2, it suffices to show, for any l, that

$$\sum_{x_{sd}<x_{td}} P\{U_l^2 \ge \frac{9\sigma_{\max}^2 p_0^2}{T_1^2}\ln^2 n \mid X_n\} \to 0, \qquad \textit{as } n \to 0.$$

Noting that $p = trace(H_n(x_{sd},x_{td})) = \sum_{l=1}^p \| \mathbf{q}_l \|^2$, we have $\| \mathbf{q}_l \|^2 = \mathbf{q}_l'\mathbf{q}_l \le p \le p_0$ and $\| \mathbf{q}_l' \sum_{i=1}^{l^0+1}\sigma_i I_n(\tau_{i-1}^0,\tau_i^0) \|^2 \le \sigma_{max}^2 \| \mathbf{q}_l \|^2 \le \sigma_{max}^2 p_0 \le \sigma_{\max}^2 p_0^2$, where $l = 1,\ldots,p$. By Lemma

4.2, with $t_0 = T_1/\sigma_{max}p_0$ we have

$$\begin{aligned}
&\sum_{x_{sd}<x_{td}} P\{U_l^2 \geq \frac{9\sigma_{\max}^2 p_0^2}{T_1^2}\ln^2 n \mid X_n\} \\
=&\sum_{x_{sd}<x_{td}} P\{|U_l| \geq \frac{3\sigma_{max}p_0 \ln n}{T_1} \mid X_n\} \\
\leq&\sum_{x_{sd}<x_{td}} 2\exp(-\frac{T_1}{\sigma_{max}p_0}\cdot\frac{3\sigma_{\max}p_0}{T_1}\ln n)\exp(c_1(\frac{T_0}{\sigma_{max}p_0})^2\sigma_{max}^2 p_0) \\
\leq& n(n-1)/n^3\exp(c_1T_0^2/p_0) \to 0,
\end{aligned}$$

as $n \to \infty$, where c_1 is the constant specified in Lemma 4.2. Finally, by appealing to the dominated convergence theorem we obtain the desired result without conditioning. ¶

Corollary 4.1 *Consider the segmented regression model 4.1.*

(i) For any j and $(\alpha, \eta] \subset (\tau_{j-1}^0, \tau_j^0]$,

$$S_n(\alpha,\eta) = \sigma_j^2 \tilde{\epsilon}_n'(\alpha,\eta)\tilde{\epsilon}_n(\alpha,\eta) - T_n(\alpha,\eta).$$

(ii) Suppose Assumption 4.0 is satisfied. Let $m \geq 1$. Then uniformly for all $(a_1, \cdots, a_m)$ such that $-\infty < a_1 < \cdots < a_m < \infty$,

$$S_n(\xi_1,\cdots,\xi_{m+l^0}) = \sum_{i=1}^{m+l^0+1} S_n(\xi_{i-1},\xi_i) = \tilde{\epsilon}_n^{\sigma\prime}\tilde{\epsilon}_n^{\sigma} + O_p(\ln^2 n),$$

where $\xi_0 = -\infty$, $\xi_{m+l^0+1} = \infty$, and $\{\xi_1,\cdots,\xi_{m+l^0}\}$ is the set $\{\tau_1^0,\cdots,\tau_{l^0}^0, a_1,\cdots,a_m\}$ after ordering its elements.

Proof: (i) Replace $\tilde{\epsilon}_n(\alpha,\eta)$ in the proof of Proposition 3.1 (i) by $\tilde{\epsilon}_n^{\sigma}(\alpha,\eta) = I_n(\alpha,\eta)\tilde{\epsilon}_n^{\sigma}$ and note $\tilde{\epsilon}_n^{\sigma}(\alpha,\eta) = \sigma_j\tilde{\epsilon}_n(\alpha,\eta)$ when $(\alpha,\eta) \subset (\tau_{j-1}^0, \tau_j^0]$. The result obtains immediately.

(ii) By (i),

$$\begin{aligned}
& S_n(\xi_1, \cdots, \xi_{m+l^0}) \\
= & \sum_{i=1}^{m+l^0+1} S_n(\xi_{i-1}, \xi_i) \\
= & \sum_{i=1}^{m+l^0+1} [\tilde{\epsilon}_n^{\sigma\prime}(\xi_{i-1}, \xi_i)\tilde{\epsilon}_n^{\sigma}(\xi_{i-1}, \xi_i) - T_n(\xi_{i-1}, \xi_i)] \\
= & \tilde{\epsilon}_n^{\sigma\prime}\tilde{\epsilon}_n^{\sigma} - \sum_{i=1}^{m+l^0+1} T_n(\xi_{i-1}, \xi_i).
\end{aligned}$$

Note that each of $(\xi_{i-1}, \xi_i]$ is contained in one of $(\tau_{j-1}^0, \tau_j^0]$, $j = 1, \cdots, l^0 + 1$. By Lemma 4.3, $\sum_{i=1}^{m+l^0+1} T_n(\xi_{i-1}, \xi_i) \leq (m + l^0 + 1) \sup_{\alpha<\eta} T_n(\alpha < \eta) = O_p(\ln^2 n)$. ¶

Lemma 4.4 *Under the condition of Theorem 4.1, there exists $\delta \in (0, \min_{1\leq j\leq l^0}(\tau_{j+1}^0 - \tau_j^0)/2)$ such that for $r = 1, \ldots, l^0$,*

$$[S_n(\tau_r^0 - \delta, \tau_r^0 + \delta) - S_n(\tau_r^0 - \delta, \tau_r^0) - S_n(\tau_r^0, \tau_r^0 + \delta)]/n \xrightarrow{a.s.} C_r \tag{4.4}$$

for some $C_r > 0$ as $n \to \infty$, $r = 1, \ldots, l^0 + 1$.

Proof It suffices to prove the result when $l^0 = 1$. For notational simplicity, we omit the subscripts and superscripts 0 in this proof. For the δ in *Assumption 4.1*, denote $X_1^* = X_n(\tau_1 - \delta, \tau_1)$, $X_2^* = X_n(\tau_1, \tau_1 + \delta)$, $X^* = X_n(\tau_1 - \delta, \tau_1 + \delta)$, $\tilde{\epsilon}_1^* = \sigma_1 I_n(\tau_1 - \delta, \tau_1)\tilde{\epsilon}_n$, $\tilde{\epsilon}_2^* = \sigma_2 I_n(\tau_1, \tau_1 + \delta)\tilde{\epsilon}_n$, $\tilde{\epsilon}^* = \tilde{\epsilon}_1^* + \tilde{\epsilon}_2^*$, and $\hat{\tilde{\beta}} = (X^{*\prime}X^*)^- X^{*\prime}Y_n$. As in ordinary regression, we have

$$\begin{aligned}
& S_n(\tau_1 - \delta, \tau_1 + \delta) \\
= & \|X_1^*\tilde{\beta}_1 + X_2^*\tilde{\beta}_2 + \tilde{\epsilon}^* - X^*\hat{\tilde{\beta}}\|^2 \\
= & \|X_1^*(\tilde{\beta}_1 - \hat{\tilde{\beta}}) + X_2^*(\tilde{\beta}_2 - \hat{\tilde{\beta}}) + \tilde{\epsilon}^*\|^2 \\
= & \|X_1^*(\tilde{\beta}_1 - \hat{\tilde{\beta}})\|^2 + \|X_2^*(\tilde{\beta}_2 - \hat{\tilde{\beta}})\|^2 + \|\tilde{\epsilon}^*\|^2 + 2\tilde{\epsilon}^{*\prime}X_1^*(\tilde{\beta}_1 - \hat{\tilde{\beta}}) + 2\tilde{\epsilon}^{*\prime}X_2^*(\tilde{\beta}_2 - \hat{\tilde{\beta}}).
\end{aligned}$$

Note that $\{\mathbf{x}_t\}$ and $\{y_t\}$ in Model (4.2) are strictly stationary and ergodic. It then follows from the strong law of large numbers for stationary ergodic stochastic processes that as $n \to \infty$,

$$\frac{1}{n}X^{*\prime}X^{*} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\mathbf{1}_{(x_{id}\in(\tau_1-\delta,\tau_1+\delta])} \xrightarrow{a.s.} E\{\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_1-\delta,\tau_1+\delta])}\} > 0,$$

$$\frac{1}{n}X_j^{*\prime}X_j^{*} \xrightarrow{a.s.} \begin{cases} E\{\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_1-\delta,\tau_1])}\} > 0, & \text{if } \text{ j=1}, \\ E\{\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_1,\tau_1+\delta])}\} > 0, & \text{if } \text{ j=2}, \end{cases}$$

and

$$\frac{1}{n}X^{*\prime}\mathbf{Y}_n \xrightarrow{a.s.} E\{y_1\mathbf{x}_1\mathbf{1}_{(x_{1d}\in(\tau_1-\delta,\tau_1+\delta])}\},$$

where $E\{y_1\mathbf{x}_1\mathbf{1}_{(x_{1d}\in(\tau_1-\delta,\tau_1+\delta])}\} = E\{\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_1-\delta,\tau_1])}\}\tilde{\beta}_1 + E\{\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_1,\tau_1+\delta])}\}\tilde{\beta}_2$.

Therefore,

$$\hat{\tilde{\beta}} \xrightarrow{a.s.} \{E\{\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_1-\delta,\tau_1+\delta])}\}\}^{-1}E\{y_1\mathbf{x}_1\mathbf{1}_{(x_{1d}\in(\tau_1-\delta,\tau_1+\delta])}\} =: \tilde{\beta}^*.$$

Similarly, it can be shown that

$$\frac{1}{n}\|X_j^*(\tilde{\beta}_j - \hat{\tilde{\beta}})\|^2 \xrightarrow{a.s.} \begin{cases} (\tilde{\beta}_1 - \tilde{\beta}^*)'E(\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_1-\delta,\tau_1])})(\tilde{\beta}_1 - \tilde{\beta}^*), & \text{if } \text{ j=1}, \\ (\tilde{\beta}_2 - \tilde{\beta}^*)'E(\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_1,\tau_1+\delta])})(\tilde{\beta}_2 - \tilde{\beta}^*), & \text{if } \text{ j=2}, \end{cases}$$

$$\frac{1}{n}\tilde{\epsilon}^{*\prime}X_j^*(\tilde{\beta}_j - \hat{\tilde{\beta}}) \xrightarrow{a.s.} 0, \ for\ j = 1, 2,$$

and

$$\frac{1}{n}\|\tilde{\epsilon}^*\|^2 \xrightarrow{a.s.} \sigma_1^2 p_1 + \sigma_2^2 p_2,$$

where $p_1 = P\{x_{1d} \in (\tau_1 - \delta, \tau_1]\}$ and $p_2 = P\{x_{1d} \in (\tau_1, \tau_1 + \delta]\}$. Thus, as $n \to \infty$, $\frac{1}{n}S_n(\tau_1 - \delta, \tau_1 + \delta)$ has a finite limit, given by

$$\begin{aligned} &\lim_{n\to\infty}\frac{1}{n}S_n(\tau_1 - \delta, \tau_1 + \delta) \\ =&(\tilde{\beta}_1 - \tilde{\beta}^*)'E(\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_1-\delta,\tau_1])})(\tilde{\beta}_1 - \tilde{\beta}^*) + (\tilde{\beta}_2 - \tilde{\beta}^*)'E(\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_1,\tau_1+\delta])})(\tilde{\beta}_2 - \tilde{\beta}^*) \\ &+ \sigma_1^2 p_1 + \sigma_2^2 p_2. \end{aligned}$$

It remains to show that $\frac{1}{n}S_n(\tau_1 - \delta, \tau_1)$ and $\frac{1}{n}S_n(\tau_1, \tau_1 + \delta)$ converge to $\sigma_1^2 p_1$ and $\sigma_2^2 p_2$ respectively, and either $(\tilde{\beta}_1 - \tilde{\beta}^*)'E(\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_1-\delta,\tau_1])})(\tilde{\beta}_1 - \tilde{\beta}^*) > 0$ or $(\tilde{\beta}_2 - \tilde{\beta}^*)'E(\mathbf{x}_1\mathbf{x}_1'$

$\mathbf{1}_{(x_{1d}\in(\tau_1,\tau_1+\delta])})\cdot(\tilde{\beta}_2-\tilde{\beta}^*)>0$. The latter is a direct consequence of the assumed conditions while the former can be shown again by the strong law of large numbers. To this end, we first write $S_n(\tau_1-\delta,\tau_1)$ in the following form,

$$S_n(\tau_1-\delta,\tau_1)=\tilde{\epsilon}_1^{*\prime}\tilde{\epsilon}_1^*-T_n(\tau_1-\delta,\tau_1)$$

using Corollary 4.1 (i). Bearing in mind $E\epsilon_1^2=1$, by the strong law of large numbers,

$$\frac{1}{n}\tilde{\epsilon}_1^{*\prime}\tilde{\epsilon}_1^* \xrightarrow{a.s.} \sigma_1^2E[\epsilon_1^2\mathbf{1}_{(x_{1d}\in(\tau_1-\delta,\tau_1])}]=\sigma_1^2P\{x_{1d}\in(\tau_1-\delta,\tau_1]\},$$
$$\frac{1}{n}\tilde{\epsilon}_1^*X_1^* \xrightarrow{a.s.} \sigma_1E[\epsilon_1\mathbf{x}_1\mathbf{1}_{(x_{1d}\in(\tau_1-\delta,\tau_1])}]=\mathbf{0},$$

and $W=\lim_{n\to\infty}\frac{1}{n}X_1^{*\prime}X_1^*$ is positive definite under the assumption. Therefore,

$$T_n(\tau_1-\delta,\tau_1)=(\frac{1}{n}\tilde{\epsilon}_1^{*\prime}X_1^*)(\frac{1}{n}X_1^{*\prime}X_1^*)^-(\frac{1}{n}X_1^{*\prime}\tilde{\epsilon}_1^*) \xrightarrow{a.s.} \mathbf{0}W^{-1}\mathbf{0}=0.$$

Thus, $\frac{1}{n}S_n(\tau_1-\delta,\tau_1)\xrightarrow{a.s.}\sigma_1^2p_1$. The same argument can also be used to show that $\frac{1}{n}S_n(\tau_1,\tau_1+\delta)\xrightarrow{a.s.}\sigma_2^2p_2$. This completes the proof. ¶

Now define $\sigma_0^2=\sum_{j=1}^{l^0+1}p_j\sigma_j^2$, where $p_j=P\{x_{1d}\in(\tau_{j-1}^0,\tau_j^0]\}$. Applying the strong law of large numbers to $\{\epsilon_t\mathbf{1}_{(x_{td}\in(\tau_{j-1}^0,\tau_j^0])}\}$ for all j, we obtain $\frac{1}{n}\tilde{\epsilon}_n^{\sigma\prime}\tilde{\epsilon}_n^{\sigma}\xrightarrow{a.s.}\sigma_0^2$.

Lemma 4.5 *Under the condition of Theorem 4.1, we have*

(i) for every $l<l^0$, $P\{\hat{\sigma}_l^2>\sigma_0^2+C\}\to 1$, as $n\to\infty$ for some $C>0$, and

(ii) for every l such that $l^0\le l\le L$, where L is an upper bound of l^0,

$$0\le\frac{1}{n}\tilde{\epsilon}^{\sigma\prime}\tilde{\epsilon}^{\sigma}-\hat{\sigma}_l^2=O_p(ln^2(n)/n),$$

where $\hat{\sigma}_l^2=\frac{1}{n}S_n(\hat{\tau}_1,\ldots,\hat{\tau}_l)$ is the estimated σ_0^2 when the true number of thresholds is assumed to be l.

Proof (i) Since $l < l^0$, for $\delta \in (0, \min_{1\le j\le l^0}(\tau_{j+1}^0 - \tau_j^0)/2)$ in *Assumption 4.1*, there exists $1 \le r \le l^0$, such that $(\hat{\tau}_1, \ldots, \hat{\tau}_l) \in A_r := \{(\tau_1, \ldots, \tau_l) : |\tau_s - \tau_r^0| > \delta,\ s = 1, \ldots, l\}$. Hence, if we can show that for each r, $1 \le r < l^0$, with probability approaching 1,

$$\min_{(\tau_1,\cdots,\tau_l)\in A_r} S_n(\tau_1, \cdots, \tau_l)/n > \sigma_0^2 + C_r,$$

for some $C_r > 0$, then by choosing $C := \min_{1\le r\le l^0}\{C_r\}$, we prove the desired result.

For any $(\tau_1, \cdots, \tau_l) \in A_r$, let $\xi_1 \le \cdots \le \xi_{l+l^0+1}$ be the ordered set $\{\tau_1, \ldots, \tau_l, \tau_1^0, \ldots, \tau_{r-1}^0, \tau_r^0 - \delta, \tau_r^0 + \delta, \tau_{r+1}^0, \ldots, \tau_{l^0}^0\}$ and let $\xi_0 = -\infty$, $\xi_{l+l^0+2} = \infty$. Then it follows from Corollary 4.1 (ii) that uniformly in A_r,

$$\begin{aligned}
&\frac{1}{n} S_n(\tau_1, \cdots, \tau_l)\\
\ge& \frac{1}{n} S_n(\xi_1, \cdots, \xi_{l+l^0+1})\\
=& \frac{1}{n} \sum_{j=1}^{l+l^0+2} S_n(\xi_{j-1}, \xi_j)\\
=& \frac{1}{n}\Big[\sum_{\{j:\ \xi_j \ne \tau_r^0+\delta\}} S_n(\xi_{j-1}, \xi_j) + S_n(\tau_r^0 - \delta, \tau_r^0) + S_n(\tau_r^0, \tau_r^0 + \delta)\Big]\\
&+ \frac{1}{n}[S_n(\tau_r^0 - \delta, \tau_r^0 + \delta) - S_n(\tau_r^0 - \delta, \tau_r^0) - S_n(\tau_r^0, \tau_r^0 + \delta)]\\
=& \frac{1}{n}\tilde{\epsilon}_n^{\sigma\prime}\tilde{\epsilon}_n^{\sigma} + O_p(\ln^2(n)/n) + \frac{1}{n}[S_n(\tau_r^0 - \delta, \tau_r^0 + \delta) - S_n(\tau_r^0 - \delta, \tau_r^0) - S_n(\tau_r^0, \tau_r^0 + \delta)].
\end{aligned} \tag{4.5}$$

By the strong law of large numbers the first term on the RHS is $\sigma_0^2 + o(1)$ a.s.. By Lemma 4.4, the third term on the RHS is $C_r + o(1)$ a.s.. Thus

$$\frac{1}{n} S_n(\tau_1, \cdots, \tau_l) \ge \sigma_0^2 + C_r + o_p(1),$$

where C_r is defined in (4.4).

(ii) Let $\xi_1 \le \cdots \le \xi_{l+l^0}$ be the ordered set, $\{\hat{\tau}_1, \cdots, \hat{\tau}_l, \tau_1^0, \cdots, \tau_{l^0}^0\}$, $\xi_0 = \tau_0^0 = -\infty$ and

$\xi_{l+l^0+1} = \tau^0_{l^0+1} = \infty$. Since $l \geq l^0$, by Corollary 4.1 (ii) again,

$$\begin{aligned}
\tilde{\epsilon}^{\sigma\prime}_n \tilde{\epsilon}^{\sigma}_n \geq & S_n(\tau^0_1, \cdots, \tau^0_{l^0}) \\
\geq & S_n(\hat{\tau}_1, \cdots, \hat{\tau}_l) \\
= & n\hat{\sigma}^2_l \\
\geq & S_n(\xi_1, \cdots, \xi_{l+l^0}) \\
= & \sum_{j=1}^{l+l^0+1} S_n(\xi_{j-1}, \xi_j) \\
= & \tilde{\epsilon}^{\sigma\prime}_n \tilde{\epsilon}^{\sigma}_n + O_p(ln^2(n)).
\end{aligned}$$

This proves (ii). ¶

Proof of Theorem 4.1 By Lemma 4.5 (i), for $l < l^0$ and sufficiently large n, there exists $C > 0$ such that

$$MIC(l) = \ln(\hat{\sigma}^2_l) + p^*(\ln n)^{2+\delta}/n \geq \ln(\sigma^2_0 + C/2) \geq \ln(\sigma^2_0) + \ln(1 + C/(2\sigma^2_0))$$

with probability approaching 1. By Lemma 4.5 (ii), for $l \geq l^0$,

$$MIC(l) = \ln(\hat{\sigma}^2_l) + p^*(\ln n)^{2+\delta}/n \xrightarrow{p} \ln \sigma^2_0.$$

Thus, $P\{\hat{l} \geq l^0\} \to 1$ as $n \to \infty$. By Lemma 4.5 (ii) and the strong law of large numbers, for $l^0 < l \leq L$,

$$0 \geq \hat{\sigma}^2_l - \hat{\sigma}^2_{l^0} = [\hat{\sigma}^2_l - \frac{1}{n}\tilde{\epsilon}^{\sigma\prime}_n \tilde{\epsilon}^{\sigma}_n] - [\hat{\sigma}^2_{l^0} - \frac{1}{n}\tilde{\epsilon}^{\sigma\prime}_n \tilde{\epsilon}^{\sigma}_n] = O_p(\ln^2 n/n),$$

and

$$[\hat{\sigma}^2_{l^0} - \sigma^2_0] = [\hat{\sigma}^2_{l^0} - \frac{1}{n}\tilde{\epsilon}^{\sigma\prime}_n \tilde{\epsilon}^{\sigma}_n] + [\frac{1}{n}\tilde{\epsilon}^{\sigma\prime}_n \tilde{\epsilon}^{\sigma}_n - \sigma^2_0] = O_p(\ln^2 n/n) + o_p(1) = o_p(1).$$

Hence $0 \leq (\hat{\sigma}^2_{l^0} - \hat{\sigma}^2_l)/\hat{\sigma}^2_{l^0} = O_p(ln^2(n)/n)$. Note that for $0 \leq x < 1/2$, $\ln(1-x) \geq -2x$.

Therefore,

$$\begin{aligned} MIC(l) - MIC(l^0) &= \ln(\hat{\sigma}_l^2) - \ln(\hat{\sigma}_{l^0}^2) + c_0(l - l^0)(\ln n)^{2+\delta_0}/n \\ &= \ln(1 - (\hat{\sigma}_{l^0}^2 - \hat{\sigma}_l^2)/\hat{\sigma}_{l^0}^2) + c_0(l - l^0)(\ln(n))^{2+\delta_0}/n \\ &\geq -2O_p(\ln^2(n)/n) + c_0(l - l^0)(\ln(n))^{2+\delta_0}/n \\ &> 0 \end{aligned}$$

for sufficiently large n. Whence $\hat{l} \xrightarrow{p} l^0$ as $n \to \infty$. ¶

To prove Theorem 4.2, we need the following lemma.

Lemma 4.6 *Under the assumptions of Theorem 4.2, for any sufficiently small* $\delta \in (0, \min_{1\leq j \leq l^0}(\tau_{j+1}^0 - \tau_j^0)/2)$, *there exists a constant* $C_r > 0$ *such that*

$$\frac{1}{n}[S_n(\tau_r^0 - \delta, \tau_r^0 + \delta) - S_n(\tau_r^0 - \delta, \tau_r^0) - S_n(\tau_r^0, \tau_r^0 + \delta)] \xrightarrow{a.s.} C_r, \ \text{as } n \to \infty,$$

where $r = 1, \cdots, l^0$.

Proof It suffices to prove the result for the case when $l^0 = 1$. For any small $\delta > 0$, all the arguments in the proof of Lemma 4.4 apply, under *Assumption 4.2*. Hence, the result holds. ¶

Proof of Theorem 4.2 By Theorem 4.1, the problem can be restricted to $\{\hat{l} = l^0\}$. For any sufficiently small $\delta' > 0$, substituting δ' for the δ in (4.5) in the proof of Lemma 4.5 (i), we have the following inequality:

$$\begin{aligned} &\frac{1}{n}S_n(\tau_1, \cdots, \tau_{l^0}) \\ \geq &\frac{1}{n}\tilde{\epsilon}_n^{\sigma'}\tilde{\epsilon}_n^{\sigma} + O_p(ln^2(n)/n) \\ &+ \frac{1}{n}[S_n(\tau_r^0 - \delta', \tau_r^0 + \delta') - S_n(\tau_r^0 - \delta', \tau_r^0) - S_n(\tau_r^0, \tau_r^0 + \delta')], \end{aligned}$$

uniformly in $(\tau_1, \cdots, \tau_{l^0}) \in A_r := \{(\tau_1, \cdots, \tau_{l^0}) : |\tau_s - \tau_r^0| > \delta', 1 \leq s \leq l^0\}$. By Lemma 4.6, the last term on the RHS converges to a positive C_r for every r. And for sufficiently large n, the $O_p(ln^2(n)/n) < \min_{1\leq r\leq l^0}(C_r)$. Thus, uniformly in A_r, $r = 1, \ldots, l^0$, and with probability tending to 1,

$$\frac{1}{n}S_n(\tau_1, \cdots, \tau_{l^0}) > \frac{1}{n}\tilde{\epsilon}_n^{\sigma\prime}\tilde{\epsilon}_n^{\sigma} + \frac{C_r}{2}.$$

This implies that with probability approaching 1 no τ in A_r is qualified as a candidate of $\hat{\tau}$, where $\hat{\tau} = (\hat{\tau}_1, \cdots, \hat{\tau}_{l^0})$. In other words, $P(\hat{\tau} \in A_r^c) \to 1$ as $n \to \infty$. Since this is true for all r, $P(\hat{\tau} \in \bigcap_{r=1}^{l^0} A_r^c) \to 1,$ *as* $n \to \infty$. Note that for $\delta' \leq \min_{0\leq i\leq l^0}\{(\tau_{i+1}^0 - \tau_i^0)/2\}$,

$$\bigcap_{r=1}^{l^0}\{|\hat{\tau}_r - \tau_r^0| \leq \delta'\} = \bigcap_{r=1}^{l^0}\{|\hat{\tau}_{i_r} - \tau_r^0| \leq \delta', for\ some\ 1 \leq i_r \leq l^0\} = \{\hat{\tau} \in \bigcap_{r=1}^{l^0} A_r^c\}.$$

Thus we have,

$$P(|\hat{\tau}_r - \tau_r^0| \leq \delta'\ for\ r = 1, \ldots, l^0) = P(\hat{\tau} \in \bigcap_{r=1}^{l^0} A_r^c) \to 1,\ as\ n \to \infty,$$

which completes the proof. ¶

Proof of Theorem 4.3 Let σ_j^{2*} and $\tilde{\beta}_j^*$ be the "least squares estimates" of σ_j^2 and $\tilde{\beta}_j$, $j = 1, \cdots, l^0 + 1$, when l^0 and $(\tau_1^0, \cdots, \tau_{l^0}^0)$ are assumed known. First, we shall show that the $\hat{\tilde{\beta}}_j$'s are consistent. By the strong law of large numbers for ergodic sequence, $\tilde{\beta}_j^* - \tilde{\beta}_j = o_p(1)$, $j = 1, \cdots, l^0 + 1$. So it suffices to show that $\hat{\tilde{\beta}}_j - \tilde{\beta}_j^* = o_p(1)$ for each j.

Set $X_j^* = I_n(\tau_{j-1}^0, \tau_j^0)X_n$ and $\hat{X}_j = I_n(\hat{\tau}_{j-1}, \hat{\tau}_j)X_n$. Then,

$$\begin{aligned}
&\hat{\tilde{\beta}}_j - \tilde{\beta}_j^* \\
=&[(\frac{1}{n}\hat{X}_j'\hat{X}_j)^- - (\frac{1}{n}X_j^{*\prime}X_j^*)^-][\frac{1}{n}\hat{X}_j'Y_n] + [(\frac{1}{n}X_j^{*\prime}X_j^*)^-][\frac{1}{n}(\hat{X}_j - X_j^*)'Y_n] \\
=&[(\frac{1}{n}\hat{X}_j'\hat{X}_j)^- - (\frac{1}{n}X_j^{*\prime}X_j^*)^-]\{\frac{1}{n}(\hat{X}_j' - X_j^*)'Y_n + \frac{1}{n}X_j^*Y_n\} + [(\frac{1}{n}X_j^{*\prime}X_j^*)^-][\frac{1}{n}(\hat{X}_j - X_j^*)'Y_n] \\
=:&(I)\{(II) + (III)\} + (IV)(II).
\end{aligned}$$

where $(I) = [(\frac{1}{n}\hat{X}_j'\hat{X}_j)^- - (\frac{1}{n}X_j^{*\prime}X_j^*)^-]$, $(II) = \frac{1}{n}(\hat{X}_j' - X_j^*)'Y_n$, $(III) = \frac{1}{n}X_j^*Y_n$ and $(IV) = [(\frac{1}{n}X_j^{*\prime}X_j^*)^-]$. By the strong law of large numbers, both (III) and (IV) are $O_p(1)$. By Theorem

4.2, $\hat{\tau} - \tau^0 = o_p(1)$. Proposition 3.2 implies that there exists a sequence $\{a_n\}$, $a_n \to 0$ as $n \to \infty$ such that $\hat{\tau} - \tau^0 = O_p(a_n)$. Note that $(II) = \frac{1}{n}\sum_{t=1}^{n} \mathbf{x}_t y_t(\mathbf{1}_{(x_{td}\in\hat{R}_j)} - \mathbf{1}_{(x_{td}\in R_j)})$ where $\hat{R}_j = (\hat{\tau}_{j-1}, \hat{\tau}_j]$, $R_j = (\tau_{j-1}^0, \tau_j^0]$. Taking $u > 1$ and $z_t = \mathbf{a}'\mathbf{x}_t y_t$ for any real vector $\mathbf{a}$, it follows from Lemma 3.6 that $(II) = o_p(1)$. It is shown in the proof of Theorem 3.3 that $(I) = o_p(1)$. Thus, $\tilde{\hat{\beta}}_j - \tilde{\beta}_j^* = o_p(1)$, $j = 1, \cdots, l^0 + 1$.

Next, we shall show that the $\hat{\sigma}_j^2$'s are consistent. When l^0 and $(\tau_1^0, \cdots, \tau_{l^0}^0)$ are known, the least squares estimates $\hat{\sigma}_j^{2*}$'s are obtained from each regime separately. Hence within each regime, applying Corollary 4.1 (i) and Lemma 4.3, we obtain that

$$n_j \hat{\sigma}_j^{*2} = \sigma_j^2 \sum_{t=1}^{n} \epsilon_t^2 \mathbf{1}_{(x_{td}\in R_j^0)} + O_p(ln^2 n), \tag{4.6}$$

where $n_j = \sum_{t=1}^{n} \mathbf{1}_{(x_{td}\in R_j^0)}$ is the number of observations in the jth regime. By the strong law of large numbers and Lemma 4.3 $n_j/n \xrightarrow{a.s.} p_j$ as $n \to \infty$, and

$$\hat{\sigma}_j^{*2} = \frac{1}{n_j}\sigma_j^2 \sum_{t=1}^{n} \epsilon_t^2 \mathbf{1}_{(x_{td}\in R_j^0)} + O_p(\frac{ln^2 n}{n}) = \sigma_j^2 + o_p(1).$$

Therefore, it remains to show that $\hat{\sigma}_j^2 - \hat{\sigma}_j^{*2} = o_p(1)$. Recall $\hat{n}_j = \sum_{t=1}^{n} \mathbf{1}_{(x_{td}\in\hat{R}_j)}$. Applying Lemma 3.6 to $z_t = 1$ we obtain $\frac{1}{n}\hat{n}_j = \frac{1}{n}n_j + o_p(1) = p_j + o_p(1)$. Thus, it suffices to show $S_n(\hat{\tau}_{j-1}, \hat{\tau}_j) - S_n(\tau_{j-1}^0, \tau_j^0) = o_p(1)$.

Since

$$S_n(\hat{\tau}_{j-1}, \hat{\tau}_j) = Y_n'(I_n(\hat{\tau}_{j-1}, \hat{\tau}_j) - H_n(\hat{\tau}_{j-1}, \hat{\tau}_j))Y_n,$$

and

$$S_n(\tau_{j-1}^0, \tau_j^0) = Y_n'(I_n(\tau_{j-1}^0, \tau_j^0) - H_n(\tau_{j-1}^0, \tau_j^0))Y_n,$$

we have that

$$
\begin{aligned}
&S_n(\hat{\tau}_{j-1},\hat{\tau}_j) - S_n(\tau_{j-1}^0,\tau_j^0) \\
=&Y_n'(I_n(\hat{\tau}_{j-1},\hat{\tau}_j) - I_n(\tau_{j-1}^0,\tau_j^0))Y_n - Y_n'(H_n(\hat{\tau}_{j-1},\hat{\tau}_j) - H_n(\tau_{j-1}^0,\tau_j^0))Y_n \\
=&\sum_{t=1}^{n} y_t^2(\mathbf{1}_{(x_{td}\in\hat{R}_j)} - \mathbf{1}_{(x_{td}\in R_j^0)}) - \{Y_n'\hat{X}_j(\hat{X}_j'\hat{X}_j)^-\hat{X}_j'Y_n - Y_n'X_j^*(X_j'X_j^*)^-X_j^{*\prime}Y_n\} \\
=&\sum_{t=1}^{n} y_t^2(\mathbf{1}_{(x_{td}\in\hat{R}_j)} - \mathbf{1}_{(x_{td}\in R_j^0)}) - \{Y_n'\hat{X}_j(\hat{X}_j'\hat{X}_j)^-\hat{X}_j'Y_n - Y_n\hat{X}_j(\hat{X}_j'\hat{X}_j)^-X_j^{*\prime}Y_n \\
&+ Y_n'\hat{X}_j(\hat{X}_j'\hat{X}_j)^-X_j^{*\prime}Y_n - Y_n\hat{X}_j(X_j^{*\prime}X_j^*)^-X_j^{*\prime}Y_n \\
&+ Y_n'\hat{X}_j(X_j^{*\prime}X_j^*)^-X_j^{*\prime}Y_n - Y_n'\hat{X}_j^*(X_j^{*\prime}X_j^*)^-X_j^{*\prime}Y_n\} \\
=&\sum_{t=1}^{n} y_t^2(\mathbf{1}_{(x_{td}\in\hat{R}_j)} - \mathbf{1}_{(x_{td}\in R_j^0)}) - \{Y_n'\hat{X}_j(\hat{X}_j'\hat{X}_j)^-(\hat{X}_j' - X_j^{*\prime})Y_n \qquad (4.7) \\
&+ Y_n'\hat{X}_j'[(\hat{X}_j'\hat{X}_j)^- - (X_j^{*\prime}X_j^*)^-]X_j^{*\prime}Y_n + Y_n'(\hat{X}_j - X_j^*)(X_j^{*\prime}X_j^*)^-X_j^*Y_n\} \\
=&\sum_{t=1}^{n} y_t^2(\mathbf{1}_{(x_{td}\in\hat{R}_j)} - \mathbf{1}_{(x_{td}\in R_j^0)}) \\
&- \{Y_n'\hat{X}_j[(\hat{X}_j'\hat{X}_j)^- - (X_j^{*\prime}X_j^*)^- + (X_j^{*\prime}X_j^*)^-](\hat{X}_j' - X_j^{*\prime})Y_n \\
&+ Y_n'\hat{X}_j[(\hat{X}_j'\hat{X}_j)^- - (X_j^{*\prime}X_j^*)^-]X_j^{*\prime}Y_n + Y_n'(\hat{X}_j - X_j^*)(X_j^{*\prime}X_j^*)^-X_j^*Y_n\} \\
=&\sum_{t=1}^{n} y_t^2(\mathbf{1}_{(x_{td}\in\hat{R}_j)} - \mathbf{1}_{(x_{td}\in R_j^0)}) \\
&- \{((II)+(III))'[(I)+(IV)](II) + ((II)+(III))'[(I)](III) + (II)'(IV)(III)\}.
\end{aligned}
$$

Taking $u > 1$ and $z_t = y_t^2$, it follows from Theorem 4.2 and Lemma 3.6 that $\frac{1}{n}\sum_{t=1}^{n} y_t^2(\mathbf{1}_{(x_{td}\in\hat{R}_j)}$ $-\mathbf{1}_{(x_{td}\in R_j^0)}) = o_p(1)$. As we have previously shown, $(I) = o_p(1)$, $(II) = o_p(1)$, $(III) = O_p(1)$ *and* $(IV) = O_p(1)$. Hence

$$
\begin{aligned}
&S_n(\hat{\tau}_{j-1},\hat{\tau}_j) - S_n(\tau_{j-1}^0,\tau_j^0) \\
=&o_p(1) \\
&- \{(o_p(1)+O_p(1))[o_p(1)+O_p(1)]o_p(1) + (o_p(1)+O_p(1))[o_p(1)]O_p(1) + o_p(1)O_p(1)O_p(1)\} \\
=&o_p(1) \qquad \P
\end{aligned}
$$

Proposition 4.1 (Brockwell and Davis, 1987, p219-220) *Let*

$$\epsilon_t = \sum_{j=-\infty}^{\infty} \psi_j \zeta_{t-j},$$

where $\{\zeta_t\}$ *is iid with mean zero and variance* σ_ζ^2, $E\zeta_1^4 = \eta\sigma_\zeta^4$ *and* $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$. *Then,*

$$E(\epsilon_1^4) = 3\gamma^2(0) + (\eta - 3)\sigma_\zeta^4 \sum_i \psi_i^4, \tag{4.8}$$

and

$$\lim_{n\to\infty} nVar(\frac{1}{n}\sum_{t=1}^{n} \epsilon_t^2) = (\eta - 3)\gamma^2(0) + 2\sum_{j=-\infty}^{\infty} \gamma^2(j), \tag{4.9}$$

where $\gamma(\cdot)$ *is the autocovariance function of* $\{\epsilon_t\}$.

We would remark that under *Assumption 4.0*, $\gamma(j) = \sigma_\zeta^2 \sum_{i=0}^{\infty} \psi_i\psi_{i+j}$. In particular, $r(0) = E(\epsilon_1^2) = 1$. Now, we restate Lemma 3.7 with appropriately modified hypotheses.

Lemma 4.7 *Let* $\{k_n\}$ *be a sequence of positive numbers such* $k_n \to 0$ *and* $nk_n \to \infty$. *Suppose Assumptions 4.0 and 4.3 are satisfied. Then for any* $j = 1, \cdots, l^0$,

(i)

$$\frac{1}{nk_n} X_n'(\tau_j^0 - k_n, \tau_j^0) X_n(\tau_j^0 - k_n, \tau_j^0) \xrightarrow{p} E(\mathbf{x}_1\mathbf{x}_1' | x_{1d} = \tau_j^0) f_d(\tau_j^0),$$

$$\frac{1}{nk_n} X_n'(\tau_j^0, \tau_j^0 + k_n) X_n(\tau_j^0, \tau_j^0 + k_n) \xrightarrow{p} E(\mathbf{x}_1\mathbf{x}_1' | x_{1d} = \tau_j^0) f_d(\tau_j^0),$$

(ii)

$$\frac{1}{nk_n} \tilde{\epsilon}_n^{\sigma\prime}(\tau_j^0 - k_n, \tau_j^0) \tilde{\epsilon}_n^{\sigma}(\tau_j^0 - k_n, \tau_j^0) \xrightarrow{p} \sigma_j^2 f_d(\tau_j^0),$$

$$\frac{1}{nk_n} \tilde{\epsilon}_n^{\sigma\prime}(\tau_j^0, \tau_j^0 + k_n) \tilde{\epsilon}_n^{\sigma}(\tau_j^0, \tau_j^0 + k_n) \xrightarrow{p} \sigma_{j+1}^2 f_d(\tau_j^0),$$

(iii)

$$\frac{1}{nk_n} \tilde{\epsilon}_n^{\sigma\prime}(\tau_j^0 - k_n, \tau_j^0) X_n(\tau_j^0 - k_n, \tau_j^0) \xrightarrow{p} \mathbf{0},$$

$$\frac{1}{nk_n} \tilde{\epsilon}_n^{\sigma\prime}(\tau_j^0, \tau_j^0 + k_n) X_n(\tau_j^0, \tau_j^0 + k_n) \xrightarrow{p} \mathbf{0},$$

Proof (i) is the same as in Lemma 3.7, hence, it suffices to show the second equation in each of (ii) and (iii).

(ii) Noting for sufficiently large n that $\tilde{\epsilon}_n^{\sigma}(\tau_j^0, \tau_j^0+k_n) = \sigma_j \tilde{\epsilon}_n(\tau_j^0, \tau_j^0+k_n)$, it suffices to show that $\frac{1}{nk_n}\tilde{\epsilon}_n'(\tau_j^0, \tau_j^0+k_n)\tilde{\epsilon}_n(\tau_j^0, \tau_j^0+k_n) \xrightarrow{p} f_d(\tau_j^0)$ as $n \to \infty$. Let $y_{nt} = \mathbf{1}_{(x_{td}\in(\tau_j^0, \tau_j^0+k_n])}$, $\mu_n = E(y_{nt})$ and $\sigma_n^2 = Var(y_{nt})$. Then,

$$\begin{aligned}
\mu_n =& P(x_{td} \in (\tau_j^0, \tau_j^0 + k_n]) \\
=& (f_d(\tau_j^0) + o(1))k_n, \\
\sigma_n^2 =& E(\mathbf{1}^2_{(x_{td}\in(\tau_j^0, \tau_j^0+k_n])}) - [E(\mathbf{1}_{(x_{td}\in(\tau_j^0, \tau_j^0+k_n])})]^2 \\
=& \mu_n - \mu_n^2 \\
=& (f_d(\tau_j^0) + o(1))k_n.
\end{aligned}$$

In particular, $\mu_n/k_n \to f_d(\tau_j^0)$ as $n \to \infty$. It therefore suffices to show that

$$\frac{1}{nk_n}\sum_{t=1}^{n} \epsilon_t^2 \mathbf{1}_{(x_{td}\in(\tau_j^0, \tau_j^0+k_n])} - \mu_n/k_n \xrightarrow{p} 0, \quad n \to \infty,$$

or

$$\frac{1}{nk_n}\sum_{t=1}^{n}(\epsilon_t^2 y_{nt} - \mu_n) \xrightarrow{p} 0, \quad n \to \infty.$$

Since $E(\epsilon_t^2) = 1$ and hence $E(\epsilon_t^2 y_{nt}) = E(\epsilon_t^2)E(y_{nt}) = \mu_n$, this last result would be implied by

$$E[\frac{1}{nk_n}\sum_{t=1}^{n}(\epsilon_t^2 y_{nt} - \mu_n)]^2 = \frac{1}{n^2k_n^2}Var(\sum_{t=1}^{n}\epsilon_t^2 y_{nt}) = o(1).$$

Note that

$$\begin{aligned}
&\frac{1}{n^2k_n^2}Var(\sum_{t=1}^{n}\epsilon_t^2 y_{nt}) \\
=&\frac{1}{n^2k_n^2}\{Var[E(\sum_{t=1}^{n}\epsilon_t^2\mathbf{1}_{(x_{td}\in(\tau_j^0, \tau_j^0+k_n])}|\tilde{\epsilon}_n)] + E[Var(\sum_{t=1}^{n}\epsilon_t^2\mathbf{1}_{(x_{td}\in(\tau_j^0, \tau_j^0+k_n])}|\tilde{\epsilon}_n)]\} \\
=&\frac{1}{n^2k_n^2}\{Var[\sum_{t=1}^{n}\epsilon_t^2\mu_n] + E[\sum_{t=1}^{n}\epsilon_t^4\sigma_n^2]\} \\
=&O(1)\cdot Var(\frac{1}{n}\sum_{t=1}^{n}\epsilon_t^2) + O(1)\frac{1}{nk_n}E(\epsilon_1^4) \\
=&O(1)Var(\frac{1}{n}\sum_{t=1}^{n}\epsilon_t^2) + o(1)E(\epsilon_1^4).
\end{aligned}$$

It remains to show that $Var(\frac{1}{n}\sum_{t=1}^{n}\epsilon_t^2) = o(1)$ and $E(\epsilon_1^4) = O(1)$. To this end observe that $\sum_{j=0}^{\infty}\psi_j^4 < \infty$, and hence by equation (4.8), that $E(\epsilon_1^4) \sim O(1)$. Now,

$$\sum_{j=0}^{\infty}\gamma^2(j) = \sum_{j=0}^{\infty}(\sigma_\zeta^2\sum_{i=0}^{\infty}\psi_i\psi_{i+j})^2 \le \sigma_\zeta^4\sum_{j=0}^{\infty}(\sum_{i=0}^{\infty}|\psi_i\psi_{i+j}|)^2$$
$$\le \sigma_\zeta^4\sum_{j=0}^{\infty}(\sum_{i=0}^{\infty}\frac{k_0}{(i+1)^\delta}|\psi_{i+j}|)^2 \le k_0^2\sigma_\zeta^4\sum_{j=0}^{\infty}(\sum_{i=0}^{\infty}|\psi_{i+j}|)^2 < \infty.$$

Consequently, $\sum_{-\infty}^{\infty}\gamma^2(j) = 2\sum_{j=0}^{\infty}\gamma^2(j) - \gamma^2(0) < \infty$, and hence, by equation (4.9),

$$Var(\frac{1}{n}\sum_{t=1}^{n}\epsilon_t^2) = O(\frac{1}{n}).$$

(iii) Since $\tilde{\epsilon}_n^\sigma(\tau_j^0, \tau_j^0 + k_n) = \sigma_j\tilde{\epsilon}_n(\tau_j^0, \tau_j^0 + k_n)$, it suffices to show that

$$\frac{1}{nk_n}\tilde{\epsilon}_n(\tau_j^0, \tau_j^0 + k_n)X_n(\tau_j^0, \tau_j^0 + k_n) \xrightarrow{p} 0, \quad n \to \infty,$$

or, for any $\mathbf{a} \ne \mathbf{0}$,

$$E[\frac{1}{nk_n}\tilde{\epsilon}_n(\tau_j^0, \tau_j^0 + k_n)X_n(\tau_j^0, \tau_j^0 + k_n)\mathbf{a}]^2 = o(1).$$

But

$$E[\mathbf{a}'\mathbf{x}_1\mathbf{1}_{(x_{1d}\in(\tau_j^0,\tau_j^0+k_n])}] = (E[\mathbf{a}'\mathbf{x}_1|x_{1d} = \tau_j^0]f_d(\tau_j^0) + o(1))k_n$$

and

$$E[(\mathbf{a}'\mathbf{x}_1)^2\mathbf{1}_{(x_{1d}\in(\tau_j^0,\tau_j^0+k_n])}] = (E[(\mathbf{a}'\mathbf{x}_1)^2|x_{1d} = \tau_j^0]f_d(\tau_j^0) + o(1))k_n.$$

Consequently,

$$
\begin{aligned}
&E(\frac{1}{nk_n}\tilde{\epsilon}_n'(\tau_j^0,\tau_j^0+k_n)X_n(\tau_j^0,\tau_j^0+k_n)\mathbf{a})^2\\
=&E[\frac{1}{n^2k_n^2}(\sum_{t=1}^n \epsilon_t \mathbf{a}'\mathbf{x}_t \mathbf{1}_{(x_{td}\in(\tau_j^0,\tau_j^0+k_n])})^2]\\
=&\frac{1}{(nk_n)^2}E[\sum_{t=1}^n \epsilon_t^2(\mathbf{a}'\mathbf{x}_t)^2\mathbf{1}_{(x_{td}\in(\tau_j^0,\tau_j^0+k_n])}\\
&+2\sum_{t>s}(\mathbf{a}'\mathbf{x}_t)(\mathbf{a}'\mathbf{x}_s)\mathbf{1}_{(x_{td}\in(\tau_j^0,\tau_j^0+k_n])}\mathbf{1}_{(x_{sd}\in(\tau_j^0,\tau_j^0+k_n])}\epsilon_t\epsilon_s]\\
=&\frac{1}{(nk_n)^2}\{nE[(\mathbf{a}'\mathbf{x}_1)^2\mathbf{1}_{(x_{1d}\in(\tau_j^0,\tau_j^0+k_n])}]\\
&+2\sum_{t>s}[E(\mathbf{a}'\mathbf{x}_1\mathbf{1}_{(x_{1d}\in(\tau_j^0,\tau_j^0+k_n])})]^2E[\sum_{i=0}^\infty\psi_i\zeta_{t-i}\sum_{j=0}^\infty\psi_j\zeta_{s-j}]\}\\
\le&\frac{1}{nk_n^2}(E[(\mathbf{a}'\mathbf{x}_1)^2|x_{1d}=\tau_j^0]f_d(\tau_j^0)+o(1))k_n\\
&+\frac{2}{(nk_n)^2}(E[\mathbf{a}'\mathbf{x}_1|x_{1d}=\tau_j^0]f_d(\tau_j^0)+o(1))^2k_n^2\sum_{t>s}\sum_{i,j:i-j=t-s}|\psi_i\psi_j|\sigma_\zeta^2\\
=&o(1)+O(1)\frac{1}{n^2}\sum_{t>s}\sum_{i,j:i-j=t-s}|\psi_i\psi_j|\\
=&o(1)+O(1)\frac{1}{n^2}\sum_{k=1}^{n-1}\sum_{s=1}^{n-k}\sum_{i,j:i-j=k}|\psi_i\psi_j|\\
=&o(1)+O(1)\frac{1}{n^2}\sum_{k=1}^{n-1}(n-k)\sum_{j=0}^\infty|\psi_{j+k}\psi_j|\\
\le&o(1)+O(1)\frac{1}{n}\sum_{j=0}^\infty\sum_{k=1}^{n-1}|\psi_{j+k}\psi_j|\\
\le&o(1)+O(\frac{1}{n})\sum_{j=0}^\infty\sum_{k=1}^\infty|\psi_{j+k}\psi_j|\\
\le&o(1)+O(\frac{1}{n})\sum_{j=0}^\infty(\sum_{k=0}^\infty|\psi_{j+k}|)^2\\
=&o(1).
\end{aligned}
$$

This completes the proof. ¶

With Lemmas 3.6, 4.3, 4.7 and Theorems 4.2, 4.3, the proof of Theorem 4.4 is analogous of that of Theorem 3.4.

Proof of Theorem 4.4 By Theorem 4.1, the problem can be restricted to $\{\hat{l} = l^0\}$. Suppose for some j, $P(\mathbf{x}_1'(\tilde{\beta}_{j+1} - \tilde{\beta}_j) \neq 0 | x_d = \tau_j^0) > 0$. Hence $\Delta = E[(\mathbf{x}_1'(\tilde{\beta}_{j+1} - \tilde{\beta}_j))^2 | x_d = \tau_j^0] > 0$. Let $\hat{\beta}(\alpha, \eta)$ be the minimizer of $\|Y_n(\alpha,\eta) - X_n(\alpha,\eta)\tilde{\beta}\|^2$. Set $k_n = K \ln^2 n/n$ for $n = 1, 2, \cdots$, where K will be chosen later. The proofs of Lemma 3.6 and Theorem 4.3 show that if $\alpha_n \xrightarrow{p} \alpha$, $\eta_n \xrightarrow{p} \eta$, then $\hat{\beta}(\alpha_n, \eta_n) \xrightarrow{p} \hat{\beta}(\alpha, \eta)$ as $n \to \infty$. Hence, for $\tau_j^0 + k_n \to \tau_j^0$ as $n \to \infty$, $\hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n) \xrightarrow{p} \hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0)$ as $n \to \infty$. By *Assumption 4.2*, for any sufficiently small $\delta \in (\tau_{j-1}^0, \tau_j^0)$, $E\{\mathbf{x}_1\mathbf{x}_1' \, \mathbf{1}_{(x_{1d} \in (\tau_{j-1}^0 + \delta, \tau_j^0])}\}$ is positive definite, hence, by the strong law of large numbers, $\hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0) \xrightarrow{a.s.} \tilde{\beta}_j$ as $n \to \infty$. Therefore $\hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n) \xrightarrow{p} \tilde{\beta}_j$. So, there exists a sufficiently small $\delta > 0$ such that for all sufficiently large n, $\|\hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n) - \tilde{\beta}_j\| \leq \|\tilde{\beta}_j - \tilde{\beta}_{j+1}\|$ and $(\hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n) - \tilde{\beta}_{j+1})' E(\mathbf{x}_1\mathbf{x}_1' | x_{1d} = \tau_j^0)\,(\hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n) - \tilde{\beta}_{j+1}) > \Delta/2$ with probability approaching 1. Hence by Theorem 4.2, for any $\epsilon > 0$, there exists N_1 such that for $n > N_1$, with probability larger than $1 - \epsilon$, we have

(i) $|\hat{\tau}_i - \tau_i^0| < \delta, \; i = 1, \cdots, l^0$,

(ii) $\|\hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n) - \tilde{\beta}_{j+1}\|^2 \leq 2\|\tilde{\beta}_j - \tilde{\beta}_{j+1}\|^2$ and

(iii) $(\hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n) - \tilde{\beta}_{j+1})' E(\mathbf{x}_1\mathbf{x}_1' | x_{1d} = \tau_j^0)(\hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n)) - \tilde{\beta}_{j+1}) > \Delta/2$.

Let $A_j = \{(\tau_1, \cdots, \tau_{l^0}) : |\tau_i - \tau_i^0| < \delta, \; i = 1, \cdots, l^0, \; |\tau_j - \tau_j^0| > k_n\}$ $j = 1, \cdots, l^0$. Since for the least squares estimates $\hat{\tau}_1, \cdots, \hat{\tau}_{l^0}$, $S_n(\hat{\tau}_1, \cdots, \hat{\tau}_{l^0}) \leq S_n(\tau_1^0, \cdots, \tau_{l^0}^0)$,

$$\inf_{(\tau_1, \cdots, \tau_{l^0}) \in A_j} \{S_n(\tau_1, \cdots, \tau_{l^0}) - S_n(\tau_1^0, \cdots, \tau_{l^0}^0)\} > 0$$

implies $(\hat{\tau}_1, \cdots, \hat{\tau}_{l^0}) \notin A_j$, or, $|\hat{\tau}_j - \tau_j^0| \leq k_n = K \ln^2 n/n$ when (i) holds. By (i), if we show that for each j, there exists $N > N_1$ such that for all $n > N$, with probability larger than $1 - 2\epsilon$, $\inf_{(\tau_1, \cdots, \tau_{l^0}) \in A_j} \{S_n(\tau_1, \cdots, \tau_{l^0}) - S_n(\tau_1^0, \cdots, \tau_{l^0}^0)\} > 0$, we will have proved the desired result. Furthermore, by symmetry, we can consider the case when $\tau_j > \tau_j^0$ only. Hence A_j may be replaced by $A_j' = \{(\tau_1, \cdots, \tau_{l^0}) : |\tau_i - \tau_i^0| < \delta, \; i = 1, \cdots, l^0, \; \tau_j - \tau_j^0 > k_n\}$. For any $(\tau_1, \cdots, \tau_{l^0}) \in$

A'_j, let $\xi_1 \le \cdots \le \xi_{2l^0+1}$ be the set $\{\tau_1, \ldots, \tau_{l^0},\ \tau_1^0, \cdots, \tau_{j-1}^0, \tau_{j-1}^0 + \delta,\ \tau_{j+1}^0 - \delta, \tau_{j+1}^0, \cdots, \tau_{l^0}^0\}$ after ordering its elements and let $\xi_0 = -\infty$, $\xi_{2l^0+2} = \infty$. Using Corollary 4.1 (ii) twice, we have

$$\begin{aligned}
&\sum_{i:\xi_i \ne \tau_j, \tau_{j+1}^0 - \delta} S_n(\xi_{i-1}, \xi_i) + S_n(\tau_{j-1}^0 + \delta, \tau_j^0) + S_n(\tau_j^0, \tau_{j+1}^0 - \delta) \\
=&\tilde{\epsilon}_n^{\sigma\prime} \tilde{\epsilon}_n^{\sigma} + O_p(\ln^2 n) \\
=&[S_n(\tau_1^0, \cdots, \tau_{l^0}^0) + O_p(\ln^2 n)] + O_p(\ln^2 n) \\
=&S_n(\tau_1^0, \cdots, \tau_{l^0}^0) + O_p(\ln^2 n).
\end{aligned}$$

Thus,

$$\begin{aligned}
S_n(\tau_1, \cdots, \tau_{l^0}) \ge& S_n(\xi_1, \cdots, \xi_{2l^0+1}) \\
=& \sum_{i=1}^{2l^0+2} S_n(\xi_{i-1}, \xi_i) \\
=& \sum_{i:\xi_i \ne \tau_j, \tau_{j+1}^0 - \delta} S_n(\xi_{i-1}, \xi_i) + S_n(\tau_{j-1}^0 + \delta, \tau_j) + S_n(\tau_j, \tau_{j+1}^0 - \delta) \\
=& \sum_{i:\xi_i \ne \tau_j, \tau_{j+1}^0 - \delta} S_n(\xi_{i-1}, \xi_i) + S_n(\tau_{j-1}^0 + \delta, \tau_j^0) + S_n(\tau_j^0, \tau_{j+1}^0 - \delta) \\
&+[S_n(\tau_{j-1}^0 + \delta, \tau_j) + S_n(\tau_j, \tau_{j+1}^0 - \delta)] - [S_n(\tau_{j-1}^0 + \delta, \tau_j^0) + S_n(\tau_j^0, \tau_{j+1}^0 - \delta)] \\
=&S_n(\tau_1^0, \cdots, \tau_{l^0}^0) + O_p(ln^2 n) \\
&+[S_n(\tau_{j-1}^0 + \delta, \tau_j) + S_n(\tau_j, \tau_{j+1}^0 - \delta)] - [S_n(\tau_{j-1}^0 + \delta, \tau_j^0) + S_n(\tau_j^0, \tau_{j+1}^0 - \delta)],
\end{aligned}$$

where $O_p(ln^2 n)$ is independent of $(\tau_1, \cdots, \tau_{l^0}) \in A'_j$. It suffices to show that for $B_n = \{\tau_j : \tau_j \in (\tau_j^0 + k_n, \tau_j^0 + \delta)\}$ and sufficiently large n,

$$\begin{aligned}
&\inf_{\tau_j \in B_n} \{S_n(\tau_{j-1}^0 - \delta, \tau_j) + S_n(\tau_j, \tau_{j+1}^0 - \delta) - [S_n(\tau_{j-1}^0 + \delta, \tau_j^0) + S_n(\tau_j^0, \tau_{j+1}^0 - \delta)]\} \\
&> M' ln^2 n
\end{aligned} \tag{4.10}$$

with probability larger than $1 - 2\epsilon$ for some fixed $M' > 0$. Let

$$S_n(\alpha, \eta; \tilde{\beta}) = \|Y_n(\alpha, \eta) - X_n(\alpha, \eta)\tilde{\beta}\|^2 = \sum_{t=1}^{n} (y_t - \mathbf{x}_t'\tilde{\beta})^2 \mathbf{1}_{(x_{td} \in (\alpha, \eta))}.$$

Since $S_n(\alpha,\eta) = S_n(\alpha,\eta;\hat{\beta}(\alpha,\eta))$, we have

$$\begin{aligned}
&S_n(\tau_{j-1}^0+\delta,\tau_j) \\
\geq &S_n(\tau_{j-1}^0+\delta,\tau_j^0+k_n) + S_n(\tau_j^0+k_n,\tau_j) \\
= &S_n(\tau_{j-1}^0+\delta,\tau_j^0;\hat{\beta}(\tau_{j-1}^0+\delta,\tau_j^0+k_n)) + S_n(\tau_j^0,\tau_j^0+k_n;\hat{\beta}(\tau_{j-1}^0+\delta,\tau_j^0+k_n)) \\
&+ S_n(\tau_j^0+k_n,\tau_j) \\
\geq &S_n(\tau_{j-1}^0+\delta,\tau_j^0) + S_n(\tau_j^0,\tau_j^0+k_n;\hat{\beta}(\tau_{j-1}^0+\delta,\tau_j^0+k_n)) + S_n(\tau_j^0+k_n,\tau_j).
\end{aligned} \tag{4.11}$$

And since $(\tau_j^0+k_n,\tau_{j+1}^0-\delta] \subset (\tau_j^0,\tau_{j+1}^0]$ for sufficiently large n,

$$S_n(\tau_j^0+k_n,\tau_{j+1}^0-\delta;\tilde{\beta}_{j+1}) = \sigma_{j+1}^2\vec{\epsilon}_n'(\tau_j^0+k_n,\tau_{j+1}^0-\delta)\vec{\epsilon}_n(\tau_j^0+k_n,\tau_{j+1}^0-\delta).$$

Applying Corollary 4.1 (i), we have

$$\begin{aligned}
0 \leq &S_n(\tau_j^0+k_n,\tau_{j+1}^0-\delta;\tilde{\beta}_{j+1}) - [S_n(\tau_j^0+k_n,\tau_j) + S_n(\tau_j,\tau_{j+1}^0-\delta)] \\
= &T_n(\tau_j^0+k_n,\tau_j) + T_n(\tau_j,\tau_{j+1}^0-\delta).
\end{aligned}$$

By Lemma 4.3, the RHS is $O_p(\ln^2 n)$. Thus,

$$\begin{aligned}
&S_n(\tau_j^0,\tau_{j+1}^0-\delta) \\
\leq &S_n(\tau_j^0,\tau_{j+1}^0-\delta;\tilde{\beta}_{j+1}) \\
= &S_n(\tau_j^0,\tau_j^0+k_n;\tilde{\beta}_{j+1}) + S_n(\tau_j^0+k_n,\tau_{j+1}^0-\delta;\tilde{\beta}_{j+1}) \\
\leq &S_n(\tau_j^0,\tau_j^0+k_n;\tilde{\beta}_{j+1}) + S_n(\tau_j^0+k_n,\tau_j) + S_n(\tau_j,\tau_{j+1}^0-\delta) + O_p(\ln^2 n),
\end{aligned}$$

where $O_p(\ln^2 n)$ is independent of τ_j. Hence

$$\begin{aligned}
&S_n(\tau_j,\tau_{j+1}^0-\delta) \\
\geq &S_n(\tau_j^0,\tau_{j+1}^0-\delta) - S_n(\tau_j^0,\tau_j^0+k_n;\tilde{\beta}_{j+1}) - S_n(\tau_j^0+k_n,\tau_j) + O_p(\ln^2 n).
\end{aligned} \tag{4.12}$$

Therefore, by (4.11) and (4.12)

$$\begin{aligned}
&[S_n(\tau_{j-1}^0+\delta,\tau_j) + S_n(\tau_j,\tau_{j+1}^0-\delta)] - [S_n(\tau_{j-1}^0+\delta,\tau_j^0) + S_n(\tau_j^0,\tau_{j+1}^0-\delta)] \\
\geq &S_n(\tau_j^0,\tau_j^0+k_n;\hat{\beta}(\tau_{j-1}^0+\delta,\tau_j^0+k_n)) - S_n(\tau_j^0,\tau_j^0+k_n;\tilde{\beta}_{j+1}) + O_p(\ln^2 n).
\end{aligned}$$

Let $M > 0$ such that the term $|O_p(\ln^2 n)| \le M \ln^2 n$ with probability larger than $1 - \epsilon$ for all $n > N_1$. To show (4.10), it suffices to show that for sufficiently large n,

$$S_n(\tau_j^0, \tau_j^0 + k_n; \hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n)) - S_n(\tau_j^0, \tau_j^0 + k_n; \tilde{\beta}_{j+1}) - Mln^2 n > M' ln^2 n,$$

or

$$S_n(\tau_j^0, \tau_j^0 + k_n; \hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n)) - S_n(\tau_j^0, \tau_j^0 + k_n; \tilde{\beta}_{j+1}) > (M' + M) ln^2 n \tag{4.13}$$

with large probability. Recall $S_n(\alpha, \eta; \tilde{\beta}) = \|Y_n(\alpha, \eta) - X_n(\alpha, \eta)\tilde{\beta}\|^2$ and $Y_n(\tau_j^0, \tau_j^0 + k_n) = X(\tau_j^0, \tau_j^0 + k_n)\tilde{\beta}_{j+1} + \tilde{\epsilon}_n(\tau_j^0, \tau_j^0 + k_n)$. Taking K sufficiently large and applying (ii), (iii) and Lemma 4.7 (i), (iii), we can see that there exists $N \ge N_1$ such that for any $n \ge N$,

$$\begin{aligned}
&\frac{1}{nk_n}[S_n(\tau_j^0, \tau_j^0 + k_n; \hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n)) - S_n(\tau_j^0, \tau_j^0 + k_n; \tilde{\beta}_{j+1})] \\
=&\frac{1}{nk_n}[\|Y_n(\tau_j^0, \tau_j^0 + k_n) - X_n(\tau_j^0, \tau_j^0 + k_n)\hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n)\|^2 \\
&- \|Y_n(\tau_j^0, \tau_j^0 + k_n) - X_n(\tau_j^0, \tau_j^0 + k_n)\tilde{\beta}_{j+1}\|^2] \\
=&\frac{1}{nk_n}[\|X_n(\tau_j^0, \tau_j^0 + k_n)(\tilde{\beta}_{j+1} - \hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n)) + \sigma_{j+1}\tilde{\epsilon}_n(\tau_j^0, \tau_j^0 + k_n)\|^2 \\
&- \|\sigma_{j+1}\tilde{\epsilon}_n(\tau_j^0, \tau_j^0 + k_n)\|^2] \\
=&\frac{1}{nk_n}\|X_n(\tau_j^0, \tau_j^0 + k_n)(\tilde{\beta}_{j+1} - \hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n))\|^2 \\
&+ \frac{2\sigma_{j+1}}{nk_n}\tilde{\epsilon}_n'(\tau_j^0, \tau_j^0 + k_n)X_n(\tau_j^0, \tau_j^0 + k_n)(\tilde{\beta}_{j+1} - \hat{\beta}(\tau_{j-1}^0 + \delta, \tau_j^0 + k_n)) \\
>&\Delta/4 - \Delta/8 > (M' + M)/K
\end{aligned}$$

with probability larger than $1 - 2\epsilon$. Since $k_n = Kln^2 n/n$, the above implies (4.13). ¶

The following Lemma (c.f. Hall and Heyde, 1980, Liu 1991) plays an important role in establishing the central limit theorem for the sample moments involving the $\{\epsilon_t\}$. Before we state the lemma, we need to introduce some notation.

Let T be an ergodic one-to-one measure-preserving transformation on the probability space $(\Omega,\ \mathcal{F},\ P)$. Suppose $\mathcal{U}_0$ is a sub-σ-field of $\mathcal{F}$ satisfying $\mathcal{U}_0 \subseteq T^{-1}(\mathcal{U}_0)$. Also suppose that Z_0 is a square integrable r.v. defined on $(\Omega,\ \mathcal{F},\ P)$ with $E(Z_0) = 0$, and that $\{Z_t\}$ is a sequence of r.v.'s defined by $Z_t = Z_0(T^t\omega)$, $\omega \in \Omega$. Let $\mathcal{U}_k = T^{-k}(\mathcal{U}_0)$, $k = 0, \pm 1, \cdots$.

Lemma 4.8 *Suppose that $\mathcal{U}_0 \subseteq T^{-1}(\mathcal{U}_0)$ and put $\mathcal{U}_k = T^{-k}(\mathcal{U}_0)$. Let $E(Z_0^2) < \infty$ and $E(Z_0) = 0$. If*

$$\sum_{m=1}^{\infty}\{(E[E(Z_0|\mathcal{U}_{-m})]^2)^{1/2} + (E[Z_0 - E(Z_0|\mathcal{U}_m)]^2)^{1/2}\} < \infty,$$

then $\sigma^{*2} := \lim_{n\to\infty} \frac{E(S_n^2)}{n}$ exists, where $S_n := \sum_{t=1}^{n} Z_t$. Further,

$$\frac{S_n}{\sqrt{n}} \xrightarrow{d} N(\mathbf{0}, \sigma^{*2}).$$

Proof The proof is obtained from Hall and Heyde (1980, Theorem 5.5 and Corollary 5.4) or Liu (1991, Theorem 4.1). ¶

Proposition 4.2 (Brockwell and Davis, 1987, Remark 2, p212)

Let

$$\epsilon_t = \sum_{j=-\infty}^{\infty} \psi_j \zeta_{t-j},$$

where the $\{\zeta_t\}$ is an iid sequence of random variables each with mean zero and variance σ_ζ^2. If $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$, then, $\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty$ and

$$\lim_{n\to\infty} nVar(\frac{1}{n}\sum_{t=1}^{n}\epsilon_t) = \sum_{h=-\infty}^{\infty}\gamma(h) = \sigma_\zeta^2(\sum_{j=-\infty}^{\infty}\psi_j)^2.$$

To facilitate the statement of the next result let

$$G_j = E(\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(x_{1d}\in(\tau_{j-1}^0,\tau_j^0])}),$$

$$\Gamma_j = G_j + 2\sum_{i=1}^{\infty} \gamma(i) E(\mathbf{x}_1 \mathbf{1}_{(x_{1d} \in (\tau_{j-1}^0, \tau_j^0])}) E(\mathbf{x}_1 \mathbf{1}_{(x_{1d} \in (\tau_{j-1}^0, \tau_j^0])})',$$

and

$$\Sigma_j = \sigma_j^2 G_j^{-1} \Gamma_j G_j^{-1},$$

where $\gamma(i) = E(\epsilon_1 \epsilon_{1+i})$ and $j = 1, \cdots, l^0 + 1$. Also recall that for each $j = 1, \cdots, l^0 + 1$,

$$\tilde{\beta}_j^* = (X_n'(\tau_{j-1}^0, \tau_j^0) X_n(\tau_{j-1}^0, \tau_j^0))^- X_n(\tau_{j-1}^0, \tau_j^0) Y_n$$

is the least squares estimate of $\tilde{\beta}_j$ given τ_j^0's.

Lemma 4.9 *Under the Assumptions 4.0, 4.1 and 4.3,*

$$\sqrt{n}(\tilde{\beta}_j^* - \tilde{\beta}_j) \xrightarrow{d} N(\mathbf{0}, \Sigma_j),$$

$j = 1, \cdots, l^0 + 1.$

Proof: First, we shall show that

$$\frac{1}{\sqrt{n}} X_n'(\tau_{j-1}^0, \tau_j^0) \tilde{\epsilon}_n^\sigma \xrightarrow{d} N(\mathbf{0}, \sigma_j^2 \Gamma_j).$$

It suffices to show that for any constant vector α,

$$\frac{1}{\sqrt{n}} \alpha' X_n'(\tau_{j-1}^0, \tau_j^0) \tilde{\epsilon}_n \xrightarrow{d} N(\mathbf{0}, \sigma_\alpha^2),$$

where $\sigma_\alpha^2 = \alpha' \Gamma_j \alpha$.

By *Assumption 4.3*, $\{\mathbf{x}_t\}_{t=-\infty}^{\infty}$ is an iid sequence of random variables. Let $\mathcal{F}_t = \sigma(\zeta_s, \mathbf{x}_s,\ s \le t)$ denote the σ-field generated by $\{\zeta_s, \mathbf{x}_s,\ s \le t\}$, and $Z_t = \alpha' \mathbf{x}_t \epsilon_t \mathbf{1}_{(x_{td} \in (\tau_{j-1}^0, \tau_j^0])}$ for a given constant vector α. To show that $\frac{1}{\sqrt{n}} \sum_{t=1}^n Z_t$ has an asypmtotic normal distribution, one needs to verify the conditions of Lemma 4.8. Thus, it suffices to show that $EZ_0 = 0$, $EZ_0^2 < \infty$, $\sum_{m=1}^{\infty} (E[E(Z_0|\mathcal{F}_{-m})^2])^{1/2} < \infty$, and

$$\sum_{m=1}^{\infty} (E[Z_0 - E(Z_0|\mathcal{F}_m)]^2)^{1/2} < \infty. \tag{4.14}$$

Observe that $EZ_0 = \alpha' E(\mathbf{x}_0 \mathbf{1}_{(x_{0d} \in (\tau^o_{j-1}, \tau^o_j])}) E\epsilon_0 = 0$ and $EZ_0^2 = \alpha' E(\mathbf{x}_0 \mathbf{x}_0' \mathbf{1}_{(x_{0d} \in (\tau^o_{j-1}, \tau^o_j])})\alpha < \infty$. Also, for $m \geq 1$, $Z_0 = \alpha' \mathbf{x}_0 \epsilon_0 \mathbf{1}_{(x_{0d} \in (\tau^o_{j-1}, \tau^o_j])}$ is $\mathcal{F}_m$-measurable, hence $Z_0 - E(Z_0|\mathcal{F}_m) = Z_0 - Z_0 = 0$. So (4.14) is trivial. It remains to show that $\sum_{m=1}^{\infty}(E[E(Z_0|\mathcal{F}_{-m})^2])^{1/2} < \infty$. Now, note that

$$\begin{aligned}
&E[E(Z_0|\mathcal{F}_{-m})]^2 \\
=&E[E(\alpha' \mathbf{x}_0 \mathbf{1}_{(x_{0d} \in (\tau^o_{j-1}, \tau^o_j])} \sum_{i=0}^{\infty} \psi_i \zeta_{-i} | \mathcal{F}_{-m})]^2 \\
=&E[E(\alpha' \mathbf{x}_0 \mathbf{1}_{(x_{0d} \in (\tau^o_{j-1}, \tau^o_j])}) \sum_{i=m}^{\infty} \psi_i \zeta_{-i}]^2 \\
=&[E(\alpha' \mathbf{x}_0 \mathbf{1}_{(x_{0d} \in (\tau^o_{j-1}, \tau^o_j])})]^2 E[\sum_{i=m}^{\infty} \psi_i \zeta_{-i}]^2 \\
=&[E(\alpha' \mathbf{x}_0 \mathbf{1}_{(x_{0d} \in (\tau^o_{j-1}, \tau^o_j])})]^2 \sum_{i=m}^{\infty} \psi_i^2 \sigma_\zeta^2 \\
=&c_j \sum_{i=m}^{\infty} \psi_i^2,
\end{aligned}$$

where $c_j = [E(\alpha' \mathbf{x}_0 \mathbf{1}_{(x_{0d} \in (\tau^o_{j-1}, \tau^o_j])})]^2 \sigma_\zeta^2$. Thus

$$\begin{aligned}
&\sum_{m=1}^{\infty} \{E[E(Z_0|\mathcal{F}_{-m})]^2\}^{1/2} \\
=&\sum_{m=1}^{\infty} \{c_j \sum_{i=m}^{\infty} \psi_i^2\}^{1/2} \\
=&\sqrt{c_j} \sum_{m=1}^{\infty} (\sum_{i=m}^{\infty} \psi_i^2)^{1/2} \\
\leq&\sqrt{c_j} k_0 \sum_{m=1}^{\infty} [\sum_{i=m}^{\infty} \frac{1}{(i+1)^{2\delta}}]^{1/2},
\end{aligned}$$

under our assumption that $|\psi_i| \leq k_0/(i+1)^\delta$ for all i. Replacing the δ in equation (4.3) with 2δ, we obtain that

$$\sum_{i=m}^{\infty} \frac{1}{(i+1)^{2\delta}} = \sum_{l=1}^{\infty} \frac{1}{(m+l)^{2\delta}} \leq \frac{1}{(2\delta - 1)m^{2\delta - 1}}. \tag{4.15}$$

Since $2\delta - 1 > 1$,

$$\sum_{m=1}^{\infty} \{E[E(Z_0|\mathcal{F}_{-m})]^2\}^{1/2}$$
$$\leq \sqrt{c_j} k_0 \{\sum_{m=1}^{\infty} (\frac{1}{(2\delta-1)m^{2\delta-1}})^{1/2}\}$$
$$\leq \sqrt{c_j} k_0 \{(\frac{1}{2\delta-1})^{1/2} (\sum_{m=1}^{\infty} \frac{1}{m^{2\delta-1}})^{1/2}\} < \infty.$$

This shows that $\frac{1}{\sqrt{n}} \sum_{t=1}^{n} Z_t$ has an asymptotic normal distribution. We next calculate the asymptotic variance of $n^{-1/2} \sum_{t=1}^{n} Z_t$. By Lemma 4.8, it is

$$\lim_{n\to\infty} \frac{ES_n^2}{n}$$
$$= \lim_{n\to\infty} \frac{1}{n} E(\sum_{t=1}^{n} \alpha' \mathbf{x}_t \epsilon_t \mathbf{1}_{(x_{td}\in(\tau_{j-1}^0, \tau_j^0])})^2$$
$$= \lim_{n\to\infty} \frac{1}{n} E[\sum_{s,t} \alpha' \mathbf{x}_s \alpha' \mathbf{x}_t \epsilon_s \epsilon_t \mathbf{1}_{(x_{sd}\in(\tau_{j-1}^0, \tau_j^0])} \mathbf{1}_{(x_{td}\in(\tau_{j-1}^0, \tau_j^0])}]$$
$$= \lim_{n\to\infty} \frac{1}{n} \{\sum_{t=1}^{n} E[(\alpha' \mathbf{x}_t)^2 \mathbf{1}_{(x_{td}\in(\tau_{j-1}^0, \tau_j^0])}] E\epsilon_t^2$$
$$+ \sum_{s\neq t} E(\alpha' \mathbf{x}_s \alpha' \mathbf{x}_t \mathbf{1}_{(x_{sd}\in(\tau_{j-1}^0, \tau_j^0])} \mathbf{1}_{(x_{td}\in(\tau_{j-1}^0, \tau_j^0])}) E(\epsilon_s \epsilon_t)\}$$
$$= E[(\alpha' \mathbf{x}_1)^2 \mathbf{1}_{(x_{1d}\in(\tau_{j-1}^0, \tau_j^0])}] + [E(\alpha' \mathbf{x}_1 \mathbf{1}_{(x_{1d}\in(\tau_{j-1}^0, \tau_j^0])})]^2 \lim_{n\to\infty} \frac{1}{n} \sum_{s\neq t} E\epsilon_s \epsilon_t$$
$$= \alpha' G_j \alpha + [E(\alpha' \mathbf{x}_1 \mathbf{1}_{(x_{1d}\in(\tau_{j-1}^0, \tau_j^0])})]^2 \lim_{n\to\infty} \frac{1}{n} E[\sum_{s,t} \epsilon_s \epsilon_t - \sum_{t=1}^{n} \epsilon_t^2]$$
$$= \alpha' G_j \alpha + \alpha' [E(\mathbf{x}_1 \mathbf{1}_{(x_{1d}\in(\tau_{j-1}^0, \tau_j^0])}) E(\mathbf{x}_1 \mathbf{1}_{(x_{1d}\in(\tau_{j-1}^0, \tau_j^0])})'] \alpha [\lim_{n\to\infty} nVar(\frac{1}{n} \sum_{t=1}^{n} \epsilon_t)$$
$$- \lim_{n\to\infty} \frac{1}{n} E(\sum_{t=1}^{n} \epsilon_t^2)],$$

where $\lim_{n\to\infty} \frac{1}{n} E(\sum_{t=1}^{n} \epsilon_t^2) = E\epsilon_1^2 = 1$ by our assumption. By Proposition 4.2,

$$\lim_{n\to\infty} nVar(\frac{1}{n} \sum_{t=1}^{n} \epsilon_t) = \sum_{i=-\infty}^{\infty} \gamma(i).$$

Hence, $\lim_{n\to\infty} nVar(\frac{1}{n} \sum_{t=1}^{n} \epsilon_t) - 1 = \sum_{i=-\infty}^{\infty} \gamma(i) - \gamma(0) = 2 \sum_{i=1}^{\infty} \gamma(i)$, and

$$\lim_{n\to\infty} \frac{ES_n^2}{n} = \alpha' \Gamma_j \alpha,$$

which is σ_α^2.

By the strong law of large numbers for ergodic sequences,

$$\frac{1}{n}X_n'(\tau_{j-1}^0,\tau_j^0)X_n(\tau_{j-1}^0,\tau_j^0) = \frac{1}{n}\sum_{t=1}^{n}\mathbf{x}_t\mathbf{x}_t'\mathbf{1}_{(x_{td}\in(\tau_{j-1}^0,\tau_j^0])} \xrightarrow{a.s.} G_j$$

as $n\to\infty$. With sufficiently large n, $(X_n'(\tau_{j-1}^0,\tau_j^0)X_n(\tau_{j-1}^0,\tau_j^0))^{-1}$ exists *a.s.*, and

$$(\frac{1}{n}X_n'(\tau_{j-1}^0,\tau_j^0)X_n(\tau_{j-1}^0,\tau_j^0))^{-} \xrightarrow{a.s.} G_j^{-1}$$

as $n\to\infty$. Hence,

$$\begin{aligned}
\tilde{\beta}_j^* =&(X_n'(\tau_{j-1}^0,\tau_j^0)X_n(\tau_{j-1}^0,\tau_j^0))^{-}X_n'(\tau_{j-1}^0,\tau_j^0)Y_n\\
=&(X_n'(\tau_{j-1}^0,\tau_j^0)X_n(\tau_{j-1}^0,\tau_j^0))^{-1}(X_n'(\tau_{j-1}^0,\tau_j^0)X_n(\tau_{j-1}^0,\tau_j^0)\tilde{\beta}_j + X_n'(\tau_{j-1}^0,\tau_j^0)\tilde{\epsilon}_n^\sigma)\\
=&\tilde{\beta}_j + \sigma_j(X_n'(\tau_{j-1}^0,\tau_j^0)X_n(\tau_{j-1}^0,\tau_j^0))^{-1}X_n'(\tau_{j-1}^0,\tau_j^0)\epsilon_n.
\end{aligned}$$

Since $\sigma_j^2 G_j^{-1\prime}[G_j + 2\sum_{i=1}^{\infty}\gamma(i)E(\mathbf{x}_1\mathbf{1}_{(x_{1d}\in(\tau_{j-1}^0,\tau_j^0])})E(\mathbf{x}_1\mathbf{1}_{(x_{1d}\in(\tau_{j-1}^0,\tau_j^0])})']G_j^{-1} = \Sigma_j$,

$$\sqrt{n}(\tilde{\beta}_j^* - \tilde{\beta}_j) \xrightarrow{d} N(\mathbf{0},\Sigma_j).$$

This completes the proof. ¶

Lemma 4.10 *Under the condition of Lemma 4.9,*

$$\frac{1}{\sqrt{n}}(\sum_{t=1}^{n}\epsilon_t^2\mathbf{1}_{(\mathbf{x}_{td}\in(\tau_{j-1}^0,\tau_j^0])} - np_j) \xrightarrow{d} N(0,v_j),$$

as $n\to\infty$, *where* $v_j = p_j(1-p_j)E(\epsilon_1^4) + p_j^2[(\eta-3)\gamma^2(0) + 2\sum_{i=-\infty}^{\infty}\gamma^2(i)]$ *and* $p_j = P(\tau_{j-1}^0 < x_{1d} \le \tau_j^0)$.

Proof It suffices to show that

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\epsilon_t^2\mathbf{1}_{(\mathbf{x}_{td}\in(\tau_{j-1}^0,\tau_j^0])} - p_j) \xrightarrow{d} N(0,v_j).$$

Let $\mathcal{F}_t = \sigma(\zeta_s, \mathbf{x}_s,\ s \leq t)$ be the σ-field generated by $\{\zeta_s, \mathbf{x}_s,\ s \leq t\}$ and

$$Z_t = \epsilon_t^2 \mathbf{1}_{(\mathbf{x}_{td} \in (\tau_{j-1}^0, \tau_j^0])} - p_j.$$

To show that $\frac{1}{\sqrt{n}} \sum_{t=1}^n Z_t$ has the asymptotic normal distribution, one needs to verify that the conditions of Lemma 4.8 obtain. That is, it must be shown that $EZ_0 = 0$, $EZ_0^2 < \infty$,

$$\sum_{m=1}^{\infty} (E[E(Z_0|\mathcal{F}_{-m})^2])^{1/2} < \infty,$$

and

$$\sum_{m=1}^{\infty} (E[Z_0 - E(Z_0|\mathcal{F}_m)]^2)^{1/2} < \infty,$$

the latter having the appearance of (4.14). We obtain $EZ_0 = E\epsilon_0^2 E\mathbf{1}_{(\mathbf{x}_{0d} \in (\tau_{j-1}^0, \tau_j^0])} - p_j = 1 \cdot p_j - p_j = 0$, and

$$\begin{aligned} EZ_0^2 =& E(\epsilon_0^2 \mathbf{1}_{(\mathbf{x}_{0d} \in (\tau_{j-1}^0, \tau_j^0])} - p_j)^2 \\ =& E(\epsilon_0^4 \mathbf{1}_{(\mathbf{x}_{0d} \in (\tau_{j-1}^0, \tau_j^0])}) + p_j^2 - 2p_j E(\epsilon_0^2 \mathbf{1}_{(\mathbf{x}_{0d} \in (\tau_{j-1}^0, \tau_j^0])}) \\ =& p_j E\epsilon_0^4 - p_j^2 \\ <& \infty. \end{aligned}$$

Also, for $m \geq 1$, Z_0 is $\mathcal{F}_m$-measurable. Hence, $Z_0 - E(Z_0|\mathcal{F}_m) = Z_0 - Z_0 = 0$. So (4.14) is trivial. It remains only to show that $\sum_{m=1}^{\infty} (E[E(Z_0|\mathcal{F}_{-m})^2])^{1/2} < \infty$. Recall that $E(\epsilon_0^2) = \sigma_\zeta^2 \sum_{i=0}^{\infty} \psi_i^2$

is assumed to be 1. Hence,

$$
\begin{aligned}
&E[E(Z_0|\mathcal{F}_{-m})]^2 \\
=&E[E(\epsilon_0^2 \mathbf{1}_{(\mathbf{x}_{0d}\in(\tau_{j-1}^0,\tau_j^0])} - p_j|\mathcal{F}_{-m})]^2 \\
=&E[p_j E(\epsilon_0^2|\mathcal{F}_{-m}) - p_j]^2 \\
=&p_j^2 E[E((\sum_{i=0}^{\infty}\psi_i\zeta_{-i})^2|\mathcal{F}_{-m}) - 1]^2 \\
=&p_j^2 E[\sum_{i=0}^{m-1}\psi_i^2\sigma_\zeta^2 + (\sum_{i=m}^{\infty}\psi_i\zeta_{-i})^2 - 1]^2 \\
=&p_j^2 E[(\sum_{i=m}^{\infty}\psi_i\zeta_{-i})^2 - \sum_{i=m}^{\infty}\psi_i^2\sigma_\zeta^2]^2 \\
=&p_j^2[E(\sum_{i=m}^{\infty}\psi_i\zeta_{-i})^4 - (\sum_{i=m}^{\infty}\psi_i\sigma_\zeta^2)^2].
\end{aligned}
$$

Using equation (4.8) by setting $\psi_i = 0$ for $i < m$, we have

$$
\begin{aligned}
&E[(\sum_{i=m}^{\infty}\psi_i\zeta_{-i})^4] - (\sigma_\zeta^2\sum_{i=m}^{\infty}\psi_i^2)^2 \\
=&3\sigma_\zeta^4(\sum_{i=m}^{\infty}\psi_i^2)^2 + (\eta-3)\sigma_\zeta^4\sum_{i=m}^{\infty}\psi_i^4 - \sigma_\zeta^4(\sum_{i=m}^{\infty}\psi_i^2)^2 \\
=&2\sigma_\zeta^4(\sum_{i=m}^{\infty}\psi_i^2)^2 + (\eta-3)\sigma_\zeta^4\sum_{i=m}^{\infty}\psi_i^4 \\
\le&2\sigma_\zeta^4(\sum_{i=m}^{\infty}\psi_i^2)^2 + (\eta-3)\sigma_\zeta^4(\sum_{i=m}^{\infty}\psi_i^2)^2 \\
=&(\eta-1)\sigma_\zeta^4(\sum_{i=m}^{\infty}\psi_i^2)^2 \\
\le&(\eta-1)\sigma_\zeta^4 k_0^4(\sum_{i=m}^{\infty}\frac{1}{(i+1)^{2\delta}})^2.
\end{aligned}
$$

By (4.15), $\sum_{i=m}^{\infty} 1/(i+1)^{2\delta} \le 1/(2\delta-1)m^{2\delta-1}$. Thus,

$$
\begin{aligned}
&\sum_{m=1}^{\infty}\{E[E(Z_0|\mathcal{F}_{-m})]^2\}^{1/2} \\
\le&\sum_{m=1}^{\infty} p_j\sqrt{\eta-1}\sigma_\zeta^2 k_0^2(\sum_{i=m}^{\infty}\frac{1}{(i+1)^{2\delta}}) \\
\le&\sqrt{\eta-1}p_j\sigma_\zeta^2 k_0^2\sum_{m=1}^{\infty}\frac{1}{(2\delta-1)m^{2\delta-1}} \\
<&\infty.
\end{aligned}
$$

Finally,

$$
\begin{aligned}
v_j &= \lim_{n\to\infty} \frac{ES_n^2}{n} \\
&= \lim_{n\to\infty} \frac{1}{n} E(\sum_{t=1}^{n} (\epsilon_t^2 \mathbf{1}_{(\mathbf{x}_{td}\in(\tau_{j-1}^0,\tau_j^0])} - p_j))^2 \\
&= \lim_{n\to\infty} \frac{1}{n} E[\sum_{s,t} (\epsilon_s^2 \mathbf{1}_{(\mathbf{x}_{sd}\in(\tau_{j-1}^0,\tau_j^0])} - p_j)(\epsilon_t^2 \mathbf{1}_{(\mathbf{x}_{td}\in(\tau_{j-1}^0,\tau_j^0])} - p_j)] \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{s,t} E[\epsilon_s^2\epsilon_t^2 \mathbf{1}_{(\mathbf{x}_{sd}\in(\tau_{j-1}^0,\tau_j^0])} \mathbf{1}_{(\mathbf{x}_{td}\in(\tau_{j-1}^0,\tau_j^0])} + p_j^2 \\
&\quad - p_j(\epsilon_t^2 \mathbf{1}_{(\mathbf{x}_{td}\in(\tau_{j-1}^0,\tau_j^0])} + \epsilon_s^2 \mathbf{1}_{(\mathbf{x}_{sd}\in(\tau_{j-1}^0,\tau_j^0])})] \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} E(\epsilon_t^4) p_j + \lim_{n\to\infty} \frac{1}{n} \sum_{s,t} [E(\epsilon_s^2\epsilon_t^2)p_j^2 + p_j^2 - p_j^2 E(\epsilon_s^2) - p_j^2 E(\epsilon_t^2)] \\
&\quad - \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} E(\epsilon_t^4) p_j^2 \\
&= p_j E(\epsilon_1^4) + \lim_{n\to\infty} \frac{1}{n} p_j^2 \sum_{s,t} E[(\epsilon_s^2 - 1)(\epsilon_t^2 - 1)] - p_j^2 E(\epsilon_1^4) \\
&= p_j(1-p_j)E(\epsilon_1^4) + p_j^2 \lim_{n\to\infty} nVar(\frac{1}{n}\sum_{t=1}^{\infty} \epsilon_t^2).
\end{aligned}
$$

By equation (4.9), $\lim_{n\to\infty} nVar(\frac{1}{n}\sum_{t=1}^{\infty}\epsilon_t^2) = (\eta-3)\gamma^2(0) + 2\sum_{i=-\infty}^{\infty}\gamma^2(i)$. This completes the proof. ¶

Proof of Theorem 4.5 We shall show the conclusion for the $\hat{\beta}_j$'s first.

Let $\tilde{\beta}_j^*$ denote the least squares estimate of $\tilde{\beta}_j$ when $(\tau_1^0, \cdots, \tau_{l^0}^0)$ is known, $j = 1, \cdots, l^0+1$. By Lemma 4.9, it suffices to show that $\hat{\beta}_j$ and $\tilde{\beta}_j^*$ share the same asymptotic distribution, for all j. In turn, it suffices to show that $\hat{\tilde{\beta}}_j - \tilde{\beta}_j^* = o_p(n^{-1/2})$.

Set $X_j^* = I_n(\tau_{j-1}^0, \tau_j^0)X_n$ and $\hat{X}_j = I_n(\hat{\tau}_{j-1}, \hat{\tau}_j)X_n$. Then,

$$
\begin{aligned}
&\hat{\tilde{\beta}}_j - \tilde{\beta}_j^* \\
=&[(\frac{1}{n}\hat{X}_j'\hat{X}_j)^- - (\frac{1}{n}X_j^{*\prime}X_j^*)^-][\frac{1}{n}\hat{X}_j'Y_n] + [(\frac{1}{n}X_j^{*\prime}X_j^*)^-][\frac{1}{n}(\hat{X}_j - X_j^*)'Y_n] \\
=&[(\frac{1}{n}\hat{X}_j'\hat{X}_j)^- - (\frac{1}{n}X_j^{*\prime}X_j^*)^-]\{\frac{1}{n}(\hat{X}_j' - X_j^*)'Y_n + \frac{1}{n}X_j^*Y_n\} + [(\frac{1}{n}X_j^{*\prime}X_j^*)^-][\frac{1}{n}(\hat{X}_j - X_j^*)'Y_n] \\
=:&(I)\{(II)+(III)\} + (IV)(II).
\end{aligned}
$$

where $(I) = [(\frac{1}{n}\hat{X}_j'\hat{X}_j)^- - (\frac{1}{n}X_j^{*\prime}X_j^*)^-]$, $(II) = \frac{1}{n}(\hat{X}_j' - X_j^*)'Y_n$, $(III) = \frac{1}{n}X_j^*Y_n$ and $(IV) = [(\frac{1}{n}X_j^{*\prime}X_j^*)^-]$. As in the proof of Theorem 4.3, both (III) and (IV) are $O_p(1)$. And the order of $o_p(n^{-1/2})$ of (I) and (II) follows from Lemma 3.6 by taking $a_n = ln^2n/n$, $z_t = (\mathbf{a}'\mathbf{x}_t)^2$ and $z_t = \mathbf{a}'\mathbf{x}_t y_t$ respectively, for any real vector $\mathbf{a}$ and $u > 2$. Thus, $\hat{\tilde{\beta}}_j - \tilde{\beta}_j^* = o_p(n^{-1/2})$.

Next, we proof the conclusion for the $\hat{\sigma}_j^2$'s.

Let $\hat{\sigma}_j^{2*}$ denote the least squares estimate of σ_j^2 when $(\tau_1^0, \cdots, \tau_{l^0}^0)$ is known, $j = 1, \cdots, l^0+1$. By Lemma 4.3, $T_n(\tau_{j-1}^0, \tau_j^0) = O_p(ln^2n)$. Hence,

$$\begin{aligned}&\frac{1}{n}S_n(\tau_{j-1}^0, \tau_j^0)\\ =&\frac{1}{n}\sigma_j^2\sum_{t-1}^{n}\epsilon_t^2\mathbf{1}_{(x_{td}\in(\tau_{j-1}^0,\tau_j^0])} - \frac{1}{n}T_n(\tau_{j-1}^0, \tau_j^0)\\ =&\frac{1}{n}\sigma_j^2\sum_{t-1}^{n}\epsilon_t^2\mathbf{1}_{(x_{td}\in(\tau_{j-1}^0,\tau_j^0])} + O_p(ln^2n/n).\end{aligned}$$

By Lemma 4.10,

$$\frac{1}{\sqrt{n}}(\sum_{t=1}^{n}\epsilon_t^2\mathbf{1}_{(\mathbf{x}_{td}\in(\tau_{j-1}^0,\tau_j^0])} - np_j) \xrightarrow{d} N(0, v_j).$$

Therefore

$$\frac{1}{\sqrt{n}}(S_n(\tau_{j-1}^0, \tau_j^0) - np_j\sigma_j^2) \xrightarrow{d} N(0, v_j\sigma_j^4),$$

and hence

$$\sqrt{n}p_j(\hat{\sigma}_j^{2*} - \sigma_j^2) \xrightarrow{d} N(0, v_j\sigma_j^4).$$

It remains to show that $\hat{\sigma}_j^2 - \hat{\sigma}_j^{2*} = o_p(n^{-1/2})$. As in the proof of Theorem 4.3, it suffices to show that $S_n(\hat{\tau}_{j-1}, \hat{\tau}_j) - S_n(\tau_{j-1}^0, \tau_j^0) = o_p(n^{-1/2})$. By equation (4.7),

$$\begin{aligned}&S_n(\hat{\tau}_{j-1}, \hat{\tau}_j) - S_n(\tau_{j-1}^0, \tau_j^0)\\ =&\sum_{t=1}^{n}y_t^2(\mathbf{1}_{(x_{td}\in\hat{R}_j)} - \mathbf{1}_{(x_{td}\in R_j^0)})\\ &-\{((II)+(III))'[(I)+(IV)](II) + ((II)+(III))'[(I)](III) + (II)'(IV)(III)\}.\end{aligned}$$

Taking $a_n = ln^2 n/n$, $u > 2$ and $z_t = y_t^2$, it follows from Lemma 3.6 that $n^{-1}\sum_{t=1}^n y_t^2$ $(\mathbf{1}_{(x_{td}\in \hat{R}_j)} - \mathbf{1}_{(x_{td}\in R_j^0)}) = o_p(n^{-1/2})$. Also, it is shown in the proof of Theorem 4.3 that both (III) and (IV) are $O_p(1)$. The order of $o_p(n^{-1/2})$ of (I) and (II) follows from Lemma 3.6 by taking $a_n = ln^2 n/n$, $z_t = (\mathbf{a}'\mathbf{x}_t)^2$ and $z_t = \mathbf{a}'\mathbf{x}_t y_t$ respectively, for any real vector a and $u > 2$. This shows that $\hat{\sigma}_j^2 - \hat{\sigma}_j^{2*} = o(n^{-1/2})$. ¶

Proof of Theorem 4.6 For $d = d^0$, by Lemma 4.5 (ii),

$$\frac{1}{n}\tilde{S}_n^{d^0} \xrightarrow{p} \sigma_0^2.$$

For $d \neq d^0$, we shall show that $\tilde{S}_n^d/n > \sigma_0^2 + C$ for some constant $C > 0$ with probability approaching 1. Again, $l^0 = 1$ is assumed for simplicity. If $d \neq d^0$, by the identifiability of d^0, for any $\{R_j^d\}_{j=1}^{L+1}$, there exist $r, s \in \{1, \cdots, L+1\}$ such that $R_r^d \supset A_s^d$ where A_s^d is defined in Theorem 2.1. Let $B_s = \{(\tau_1, \ldots, \tau_L) : R_r^d \supset A_s^d \text{ for some } r\}$. Then for any $(\tau_1, \ldots, \tau_L)$, $(\tau_1, \cdots, \tau_L) \in B_s$ for at least one $s \in \{1, \cdots, L+1\}$. Since $\hat{d}$ is chosen such that $\tilde{S}_n^{\hat{d}} \leq \tilde{S}_n^d$ for all d, it suffices to show that for $d \neq d^0$ and each s, there exists $C_s > 0$ such that

$$\inf_{(\tau_1,\ldots,\tau_L)\in B_s} \frac{1}{n} S_n^d(\tau_1, \ldots, \tau_L) \geq \sigma_0^2 + C_s \tag{4.16}$$

with probability approaching 1 as $n \to \infty$. For any $(\tau_1, \ldots, \tau_L) \in B_s$, let $R_{L+2}^d = \{\mathbf{x} : x_d \in (\tau_{r-1}, a_s)\}$, $R_{L+3}^d = \{\mathbf{x} : x_d \in (b_s, \tau_r]\}$. Then $R_r^d = A_s^d \cup R_{L+2}^d \cup R_{L+3}^d$. From Lemma 4.3 and the proof of Lemma 3.2′, we can see that the conclusion of Lemma 3.2′ still holds under current assumptions. Hence, the conclusions of Proposition 3.1′ and Lemma 3.3′ also hold. Therefore, by (3.13)

$$\frac{1}{n}S_n^d(\tau_1, \cdots, \tau_L) = \sigma_0^2 + o_p(1) + \frac{1}{n}[S_n(A_s^d) - S_n(A_s^d \cap R_1^0) - S_n(A_s^d \cap R_2^0)].$$

Now it remains to show that $\frac{1}{n}[S_n(A_s^d) - S_n(A_s^d \cap R_1^0) - S_n(A_s^d \cap R_2^0)] > C_s$ for some $C_s > 0$, with probability approaching 1. By Theorem 2.1, $E[\mathbf{x}_1\mathbf{x}_1'\mathbf{1}_{(\mathbf{x}_1\in A,\cap R_i^0)}]$, $i = 1, 2$, are positive definite. Applying Lemma 3.3$'$ we obtain the desired result. ¶

Chapter 5

SUMMARY AND FUTURE RESEARCH

5.1 A brief summary of previous chapters

In this thesis, we propose a set of procedures for estimating the parameters of a segmented regression model. The consistency of the estimators is established under fairly general conditions. For the "basic" model where the noise is an iid sequence and locally exponentially bounded, it is shown that if the model is discontinuous at a threshold, then the least squares estimate of the threshold converges at the rate of $O_p(ln^2 n/n)$. For both continuous and discontinuous models, the asymptotic normality of the estimated regression coefficients and the noise variance is established. The least squares "identifier" of the segmentation variable is shown to be consistent, if the segmentation variable is asymptotically identifiable. A more efficient method of identifying the segmentation variable is given under stronger conditions. Most of these results are generalized to the case where the noise is heteroscedastic and autocorrelated. A simulation study is carried out to demonstrate the small sample behavior of the proposed estimators. The proposed procedures perform reasonably well in identifying the models, but indicate the need for large sample sizes for estimating the thresholds.

5.2 Future research on the current model

First, further work on choosing δ_0 and c_0 in the MIC is needed. One way to reduce the risk of mis-specifying the model is to try different (δ_0, c_0) values over certain range. If several (δ_0, c_0) pairs produced the same $\hat{l}$, we would be more confident of our choice. Otherwise different models can be fitted. And the estimated regression coefficients and noise variance may then indicate what (δ_0, c_0) is more appropriate. In particular, when the noise is autocorrelated, recursive estimation procedures need to be investigated.

Second, the asymptotic normality of the estimated regression coefficients for continuous models need to be generalized to the case where the noise is heteroscedastic and autocorrelated. The techniques used in Sections 3.5 and 4.5 are useful but additional tools are needed, such as the central limit theorem for a double array of martingale sequences.

Third, the local exponential boundedness assumption made on the noise may be relaxed. Note that this assumption implies that ϵ_1 has moments of any order. If ϵ_1 is assumed to have only moments to finite order, a model selection criterion with a penalty term of the form Cn^{α} $(0 < \alpha < 1)$ may well be consistent. This has been shown by Yao (1989) for a one-dimensional step function with fixed covariates and iid noise.

5.3 Further generalizations

Further generalization of the segmented regression model will enable its broader applications. First, there may be more than one segmentation variable. For example, changes in economic policy may be triggered by the simultaneous extremes in a number of key economic indices. The results in this thesis may be generalized to the case where more than one segmentation variable is present. Further, since sometimes there is no reason to believe that segmentation has to be parallel to any of the axes, a threshold defined in terms of a linear combination of explanatory variables may be appropriate. A least squares approach or that of

Goldfeld and Quandt (1972, 1973a) can be applied. Large sample properties of the estimators given by these approaches would need to be investigated. In many economic problems, the explanatory variables exhibit certain kinds of dependence over time. The explanatory variables and the noise may also be dependent. Our results can be generalized in this direction, since the iid assumption on $\{\mathbf{x}_t\}$ is not essential. Once such generalizations are accomplished, we expect this model to be useful for many economic problems, since many economic policies and business decisions are threshold-based, at least to some extent. In fact, the segmented regression model has been applied to a foreign exchange rate problem by Liu and Susko (1992) with significantly better results than other approaches reported in the literature. And, the need for a theoretical justification for this approach is obvious.

If y_t and x_{ti} in Model 2.1 are replaced by x_t and x_{t-i} respectively $(i = l, \cdots, p)$, where $\{\mathbf{x}_t\}$ is a time series, then the model becomes a *threshold autoregressive model.* This interesting nonlinear time series models has been studied by many authors. See, for example, Tong (1987) for a review on some recent work on nonlinear time series analysis. Because this model is very similar to ours in its structure, the approaches used in this thesis may also shed some light on its model selection problem and the large sample properties of its least squares estimates. In particular, we expect a criterion similar to MIC can be used to select the number of threshold for the threshold autoregressive model.

REFERENCES

Bacon, D.W. and Watts, D.G. (1971). Estimating the transition between two intersecting straight lines. *Biometrika*, **58**, 525-543.

Bellman, R. (1969). Curve fitting by segmented straight lines. *J. Amer. Statist. Assoc.*, **64**, 1079-1084.

Billingsley, P. (1968). *Convergence of Probability Measures.* Wiley, N.Y.

Breiman, L., and Meisel, W.S. (1976). General estimates of the intrinsic variability of data in nonlinear regression models. *J. Amer. Statist. Assoc.*, **71**, 301-307.

Brockwell, P.J. and Davis, R.A. (1987). *Time series: Theory and methods.* Springer-Verlag, N.Y.

Broemeling, L.D. (1974). Bayesian inferences about a changing sequence of random variables. *Commun. Statist.*, **3**, 234-255.

Cleveland, W.S. (1979). Robust locally weighted regression: An approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.*, **74**, 829-836.

Cleveland, W.S. and Devlin, S.J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.*, **83**, 596-610.

Dunicz, B.L. (1969). Discontinuities in the surface structure of alcohol-water mixtures. *Kolloid-Zeitschr. u. Zeitschrift f. Polymere*, **230**, 346-357.

Ertel J.E. and Fowlkes E.B. (1976). Some algorithms for linear spline and piecewise multiple linear regression. *J. Amer. Statist. Assoc.*, **71**, 640-648.

Farley, J.U. and Hinich, M.J. (1970). A test for a shifting slope coefficient in a linear model. *J. Amer. Statist. Assoc.*, **65**, 1320-1329.

Feder, P.I. and Sylwester, D.L. (1968). On the asymptotic theory of least squares estimation in segmented regression: identified case (preliminary report) abstracted in *Ann. Math. Statist.*, **39**, 1362.

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

Friedman, J. H. (1988). Multivariate Adaptive Regression Splines, Report 102, Department of Statistics, Stanford University.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *Ann. Statist.* **19**, 1-141.

Feder, P.I. (1975b). The log likelihood ratio in segmented regression. *Ann. Statist.* **3**, 84-97.

Ferreira, P.E. (1975). A Bayesian analysis of switching regression model: Known number of regimes. *J. Amer. Statist. Assoc.*, **70**, 730-734.

Gallant, A.R. and Fuller, W.A. (1973). Fitting segmented polynomial regression models whose join points have to be estimated. *J. Amer. Statist. Assoc.*, **68**, 144-147.

Goldfeld, S.M. and Quandt, R.E. (1972). *Nonlinear Methods in Econometrics.* North-Holland Publishing Co.

Goldfeld, S.M. and Quandt, R.E. (1973a). The estimation of structural shifts by switching regressions. *Ann. Econ. Soc. Measurement*, **2**, 475-485.

Goldfeld, S.M. and Quandt, R.E. (1973b). A Markov model for switching regressions. *Journal of Econometrics*, **1**, 3-16.

Hall, P. and Heyde, C. (1980). *Martingale limit theory and its application*, Academic Press.

Hawkins, D.M. (1980). A note on continuous and discontinuous segmented regressions. *Technometrics*, **22**, 443-444.

Henderson, H.V. and Velleman, P.F. (1981). Building regression model interactively. *Biometrics*, **37**, 391-411.

Henderson, R. (1986). Change-point problem with correlated observations, with an application in material accountancy. *Technometrics*, **28**, 381-389.

Hinkley, D.V. (1969). Inference about the intersection in two-phase regression. *Biometrika*, **56**, 495-504.

Hinkley, D.V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, **57**, 1-17.

Holbert, D. and Broemling, L. (1977). Bayesian inferences related to shifting sequences and two-phase regression. *Commun. Statist. Theor. Meth.*, A6(3), 265-275.

Jennrich, R.J. (1969). Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.*, **40**, 633-643.

Hudson, D.J. (1966). Fitting segmented curves whose join points have to be estimated. *J. Amer. Statist. Assoc.*, **61**, 1097-1129.

Liu, J. and Liu, Z. (1991). Higher order moments and limit theory of a general bilinear time series. Unpublished manuscript.

Liu, J. and Susko, E.A. (1992). Forecasting exchange rates using segmented time series regression model – a nonlinear multi-country model. Unpublished manuscript.

MacNeill, I.B. (1978). Properties of sequences of partial sums of polynomial regression residuals with applications to test for change of regression at unknown times. *Ann. Statist.*, **6**, 422-433.

McGee, V.E., and Carleton, W.T. (1970). Piecewise regression. *J. Amer. Statist. Assoc.*, **65**, 1109-1124.

Miao, B.Q. (1988). Inference in a model with at most one slope-change point. *Journal of Multivariate Analysis*, **27**, 375-391.

Muller, H.G. and Stadtmuller, U. (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.*, **15**, 610-625.

Poirier, D.J. (1973). Piecewise regression using cubic splines. *J. Amer. Statist. Assoc.*, **68**, 515-524.

Quandt, R.E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *J. Amer. Statist. Assoc.*, **53**, 324-330.

Quandt, R.E. (1960). The estimation of the parameters of a linear regression system obeying two separate regimes. *J. Amer. Statist. Assoc.*, **53**, 873-880.

Quandt, R.E. (1972). A new approach to estimating switching regression. *J. Amer. Statist. Assoc.*, **67**, 306-310.

Quandt, R.E., and Ramsey, J.B. (1978). Estimating mixtures of normal distributions and switching regression. (With discussion). *J. Amer. Statist. Assoc.*, **73**, 730-752.

Robison, D.E. (1964). Estimates for the points of intersection of two polynomial regressions. *J. Amer. Statist. Assoc.*, **59**, 214-224.

Sacks, J. and Ylvisaker, D. (1978). Linear estimation for approximately linear models. *Ann. Statist.*, **6**, 1122-1137.

Schulze, U. (1984). A method of estimation of change points in multiphasic growth models. *Biometrical Journal*, **26**, 495-504.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 49-83.

Serfling, R.J. (1980). *Approximation theorems of mathematical statistics.* Wiley, New York.

Shaban, S.A. (1980) Change point problem and two-phase regression: an annotated bibliography. *International Statistical Review*, **48**, 83-93.

Shao, J. (1990). Asymptotic theory in heteroscedastic nonlinear models. *Statistics & Probability Letters*, **10**, 77-85.

Shumway, R.H. and Stoffer, D.S. (1991). Dynamic linear models with switching. *J. Amer. Statist. Assoc.*, **86**, 763-769.

Sylwester, D.L. (1965). On the maximum likelihood estimation for two-phase linear regression. Technical Report No. 11, Department of Statistics, Stanford Univ.

Sprent, P. (1961). Some hypotheses concerning two phase regression lines. *Biometrics*, **17**, 634-645. Univ.

Susko, E.A. (1991). Segmented regression modelling with an application to German exchange rate data. M.Sc. thesis, Department of Statistics, University of British Columbia.

Tong, H. (1987). Non-linear time series models of regularly sampled data: A review. *Proc. First World Congress of the Bernoulli Society, Tashkent, USSR*, **2**, 355-367, The Netherlands, VNU Science Press.

Weerahandi, W. and Zidek, J.V. (1988). Bayesian nonparametric smoothers for regular processes. *The Canandian journal of Statistics*, **16**, 61-73.

Worsley, K.J. (1983). Testing for a two-phase multiple regression. *Technometrics*, **25**, 35-42.

Yao, Y. (1988). Estimating the number of change-points via Schwarz' criterion. *Statistics & Probability Letters*, **6**, 181-189.

Wu, C.F.J. (1981). Asymptotic theory of nonlinear least squares estimation. *Ann. Statist.*, **9**, 501-513.

Yao, Y. and Au, S.T. (1989). Least-squares estimation of a step function. *Sankhya: The Indian Journal of Statistics, A*, **51**, 370-381.

Yeh, M.P., Gardner, R.M., Adams, T.D., Yanowitz, F.G., and Crapo, R.O. (1983). "Anaerobic threshold": Problems of determination and validation. *J. Apply. Physiol. Respirit. Envioron. Excercise Physiol.*, **55**, 1178-1186.

Zwiers, F. and Storch, H.V. (1990). Regime-dependent autoregressive time series modeling of the Southern Oscillation. *Journal of Climate*, **3**, 1347-1363.

Table 3.1: Frequency of correct identification of l^0 in 100 repetitions and the estimated thresholds for segmented regression models

(m, m_u, m_o are the frequencies of correct, under- and over-estimations of l^0)

$MIC : m(m_u, m_o)$	sample size			
$\hat{r}_1$ (SE)	30	50	100	200
$Model(a)$	79 (18, 3)	95 (4, 1)	100 (0, 0)	100 (0, 0)
	1.168 (1.500)	1.033 (1.353)	1.410 (0.984)	1.259 (0.665)
$Model(b)$	70 (21, 9)	86 (8, 6)	99 (0, 1)	100 (0, 0)
	1.022 (1.546)	1.220 (1.407)	1.432 (0.908)	1.245 (0.692)
$Model(c)$	80 (6, 14)	97 (1, 2)	100 (0, 0)	100 (0, 0)
	0.890 (0.737)	0.761 (0.502)	0.901 (0.221)	0.932 (0.151)
$Model(d)$	85 (8, 7)	99 (0, 1)	100 (0, 0)	100 (0, 0)
	0.791 (1.009)	0.860 (0.665)	0.971 (0.232)	0.963 (0.169)
$Model(e)$	68 (23, 9)	87 (12, 1)	100 (0, 0)	100 (0, 0)
	0.463 (1.735)	0.708 (1.332)	0.989 (0.923)	0.940 (0.707)

Table 3.2: Estimated regression coefficients and variances of noise and their standard errors with $n = 200$

(Conditional on $\hat{l} = 1$)

$\hat{\beta}_{ij}$ (SE)	Model (a)	Model (b)	Model (c)	Model (d)	Model (e)
β_{10}	-0.003 (0.145)	-0.018 (0.146)	0.004 (0.143)	-0.008 (0.154)	-0.059 (0.177)
β_{11}	1.001 (0.038)	0.995 (0.037)	1.000 (0.035)	0.995 (0.041)	0.985 (0.045)
β_{12}	1.000 (0.024)	0.996 (0.025)	-0.004 (0.025)	0.000 (0.024)	1.000 (0.025)
β_{13}	—	—	—	0.994 (0.023)	0.995 (0.025)
β_{20}	1.485 (0.345)	1.388 (0.332)	0.962 (0.243)	1.009 (0.225)	0.960 (0.283)
β_{21}	0.005 (0.063)	0.019 (0.067)	0.008 (0.055)	0.000 (0.049)	0.008 (0.057)
β_{23}	1.006 (0.034)	0.998 (0.034)	0.495 (0.032)	0.498 (0.032)	0.998 (0.036)
β_{24}	—	—	—	0.997 (0.034)	0.996 (0.036)
σ^2	0.948 (0.108)	0.950 (0.154)	0.956 (0.156)	0.953 (0.160)	0.944 (0.158)

Table 3.3: The empirical distribution of $\hat{l}$ in 100 repetitions by MIC, SC and YC for piecewise constant model

(n_0, n_1, n_2, n_3 are the frequencies of $\hat{l} = 0, 1, 2, 3$ respectively)

$MIC : n_0, n_1, n_2, n_3$ $YC : n_0, n_1, n_2, n_3$ $SC : n_0, n_1, n_2, n_3$	sample size		
	50	150	450
$Model(f)$	5, 30, 48, 17	0, 18, 79, 3	0, 0, 98, 2
	5, 36, 45, 14	0, 36, 64, 0	0, 9, 91, 0
	0, 17, 52, 31	0, 1, 64, 35	0, 0, 83, 17
$Model(g)$	5, 38, 51, 6	0, 23, 72, 5	0, 0, 99, 1
	7, 41, 48, 4	0, 46, 53, 1	0, 7, 93, 0
	3, 18, 56, 23	0, 2, 79, 19	0, 0, 87, 13
$Model(h)$	0, 3, 81, 16	0, 0, 96, 4	0, 0, 98, 2
	0, 3, 84, 13	0, 0, 100, 0	0, 0, 100, 0
	0, 0, 63, 37	0, 0, 82, 18	0, 0, 87, 13
$Model(i)$	0, 5, 85, 10	0, 0, 97, 3	0, 0, 100, 0
	0, 7, 86, 7	0, 0, 100, 0	0, 0, 100, 0
	0, 1, 73, 26	0, 0, 83, 17	0, 0, 93, 7

Table 3.4: The estimated thresholds and their standard errors for piecewise constant model

(Conditional on $\hat{l} = 2$)

$\hat{r}_1$, (SE) / $\hat{r}_2$, (SE)	sample size 50	150	450
$Model(f)$	0.335 (0.078)	0.338 (0.039)	0.334 (0.012)
	0.660 (0.032)	0.666 (0.008)	0.667 (0.003)
$Model(g)$	0.313 (0.076)	0.332 (0.032)	0.334 (0.013)
	0.656 (0.015)	0.669 (0.009)	0.667 (0.002)
$Model(h)$	0.316 (0.027)	0.334 (0.007)	0.333 (0.002)
	0.662 (0.030)	0.667 (0.006)	0.667 (0.003)
$Model(i)$	0.323 (0.023)	0.332 (0.010)	0.334 (0.004)
	0.661 (0.030)	0.666 (0.007)	0.667 (0.003)

Table 4.1: Frequency of correct identification of l^0 in 100 repetitions and the estimated thresholds for segmented regression models with two regimes

(m, m_u, m_o are the frequencies of correct, under- and over-estimations of l^0)

$MIC : m(m_u, m_o)$	sample size		
$\hat{\tau}_1$ (SE)	50	100	200
Model (a′)	95 (3, 2)	98 (0, 2)	99 (0, 1)
	1.322 (1.681)	1.412 (1.293)	1.223 (1.060)
Model (d′)	91 (1, 8)	95 (0, 5)	99 (0, 1)
	0.808 (0.545)	0.936 (0.256)	0.960 (0.109)
Model (e′)	94 (3, 3)	98 (0, 2)	99 (0, 1)
	0.693 (1.583)	1.088 (1.470)	1.175 (1.111)

Table 4.2: Estimated regression coefficients and variances of noise and their standard errors with $n = 200$

(Conditional on $\hat{l} = 1$)

$\hat{\beta}_{ij}$ (SE)	Model (a′)	Model (d′)	Model (e′)
β_{10}	-0.049 (0.247)	0.007 (0.190)	-0.056 (0.227)
β_{11}	0.993 (0.066)	0.998 (0.059)	0.985 (0.065)
β_{12}	1.003 (0.017)	-0.001 (0.020)	0.999 (0.019)
β_{13}	—	0.998 (0.018)	0.997 (0.018)
β_{20}	1.258 (0.730)	0.957 (0.461)	0.749 (0.596)
β_{21}	0.033 (0.129)	0.013 (0.107)	0.045 (0.126)
β_{23}	0.998 (0.033)	0.503 (0.029)	1.002 (0.030)
β_{24}	—	0.998 (0.026)	0.999 (0.029)
σ_1^2	0.656 (0.117)	0.639 (0.167)	0.634 (0.166)
σ_2^2	0.929 (0.271)	1.050 (0.391)	0.963 (0.361)

Table 4.3: Frequency of correct identification of l^0 in 100 repetitions and the estimated threshold for a segmented regression model with three regimes

(m, m_u, m_o are the frequencies of correct, under- and over-estimations of l^0)

$MIC : m(m_u, m_o)$ $\hat{\tau}_1$ (SE) $\hat{\tau}_2$ (SE)	sample size		
	50	100	200
Model (j)	62 (26, 12)	86 (6, 8)	95 (0, 5)
	-1.211 (0.251)	-1.051 (0.151)	-1.034 (0.078)
	1.046 (0.493)	1.060 (0.388)	0.974 (0.096)

Table 4.4: Estimated regression coefficients and noise variances and their standard errors with $n = 200$

(Conditional on $\hat{l} = 2$)

Model (j)	$j = 1$	$j = 2$	$j = 3$
$\hat{\beta}_{j1}$ (SE)	0.987 (0.290)	-0.029 (0.212)	0.454 (0.413)
$\hat{\beta}_{j2}$ (SE)	0.996 (0.062)	0.097 (0.480)	0.011 (0.092)
$\hat{\beta}_{j3}$ (SE)	-0.001 (0.017)	1.000 (0.032)	0.499 (0.028)
$\hat{\sigma}_j$ (SE)	0.511 (0.165)	0.681 (0.269)	1.002 (0.294)

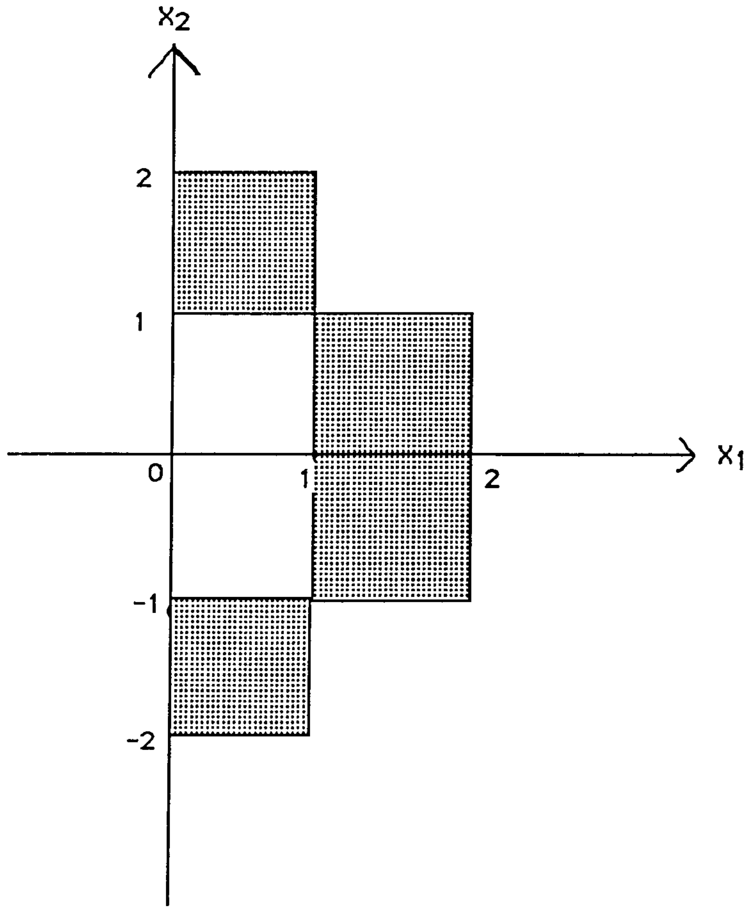

Figure 2.1 (x_1, x_2) uniformly distributed over the shaded area

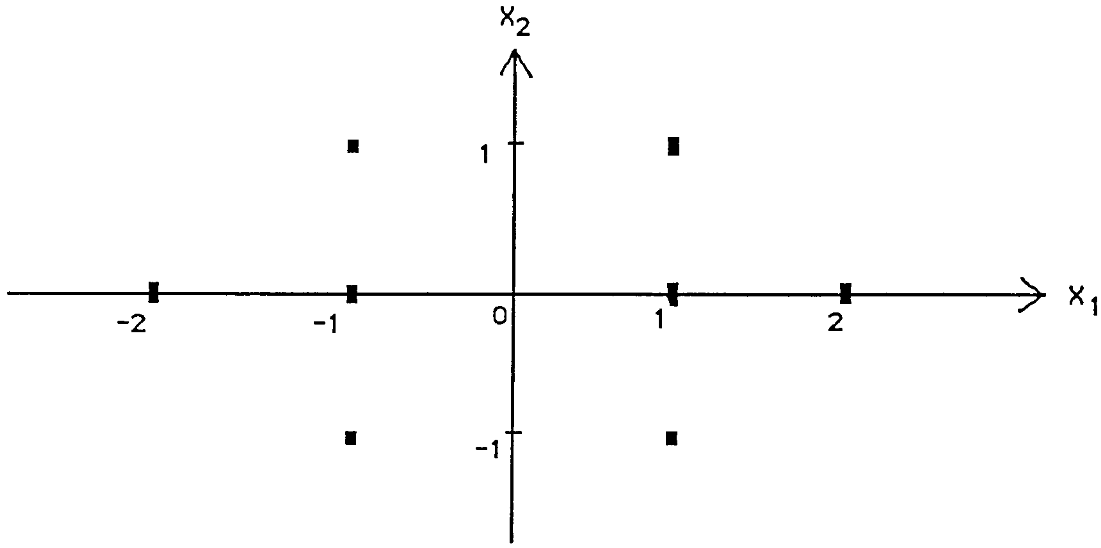

Figure 2.2 (x_1, x_2) uniformly distributed over the eight points

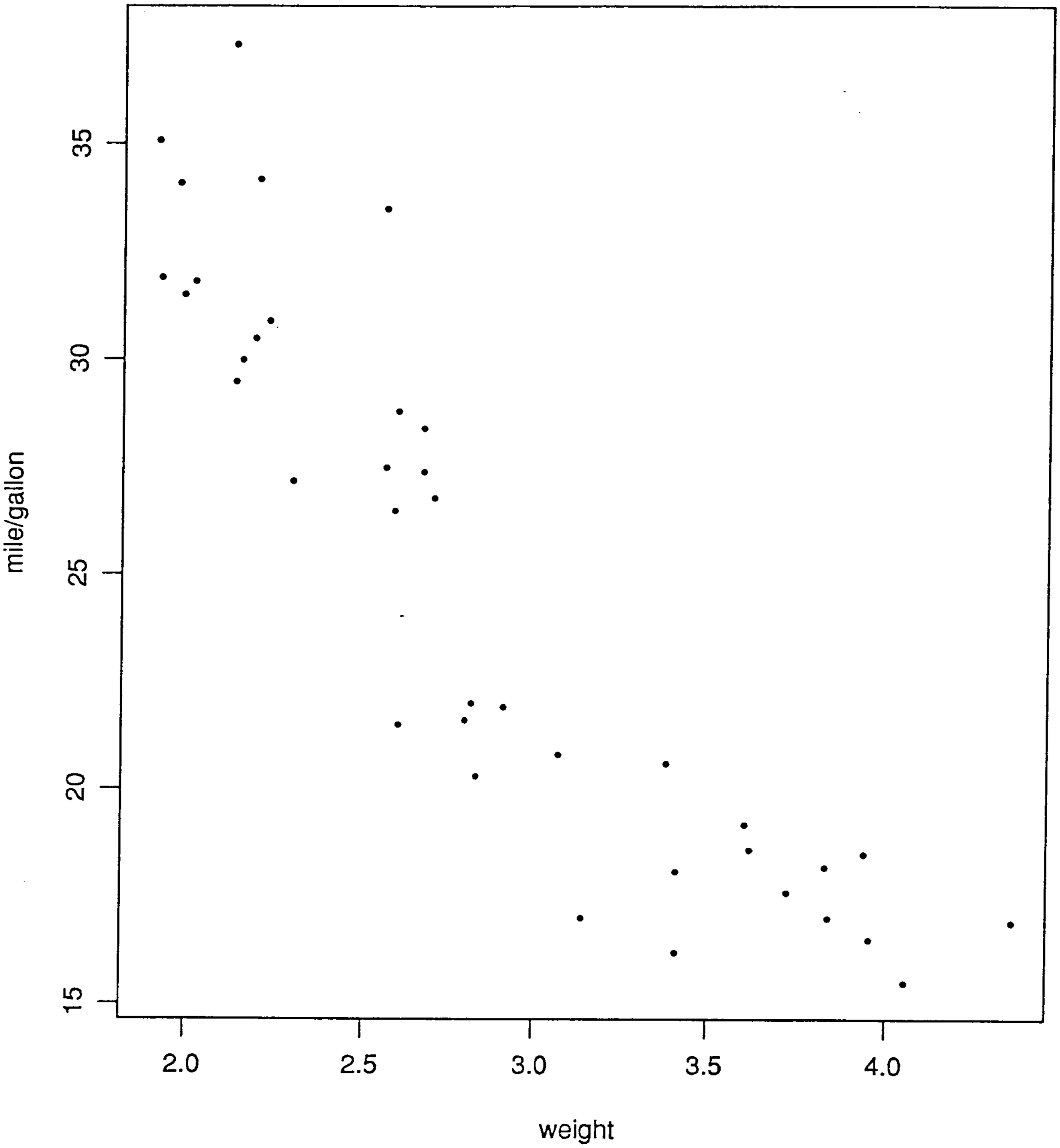

Figure 2.3 Mile per gallon vs. weight for 38 cars

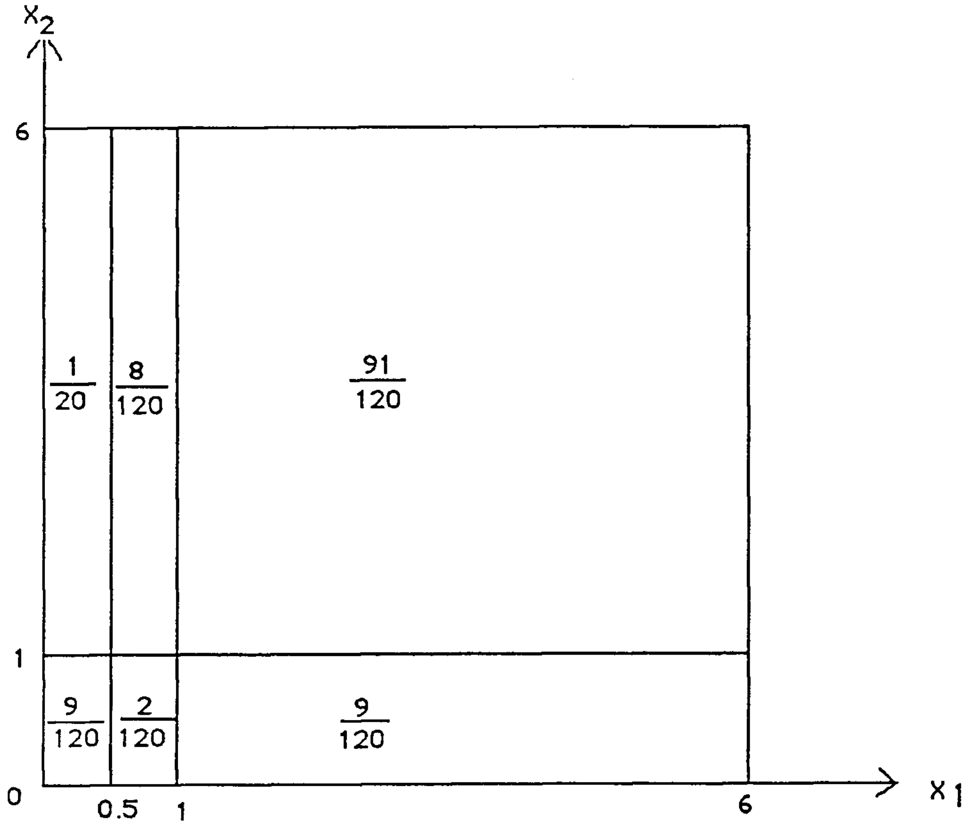

Figure 4.1 (x_1, x_2) uniformly distributed over each of six regions with indicated mass