Essays on functional coefficient models

Koo, Chao

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Essays on functional coefficient models

Proefschrift

ter verkrijging van de graad van doctor aan Tilburg University op gezag van de rector magnificus, prof. dr. E.H.L. Aarts, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de Ruth First zaal van de Universiteit op maandag 16 april 2018 om 14.00 uur door

Chao Hui Koo

geboren op 4 november 1985 te Rotterdam

Promotiecommissie:

Promotor:	prof. dr. Bas J.M. Werker
Copromotor:	dr. Pavel Čížek
Overige Leden:	prof. dr. John H.J. Einmahl
	prof. dr. Bertrand Melenberg
	prof. dr. Irène Gijbels

Acknowledgements

I would like to express my sincere gratitude to my supervisor dr. Pavel Čížek for the continuous support of my Ph.D study and related research. Besides my supervisor, I am grateful to my promotor Prof. Bas J.M. Werker and the rest of my thesis committee: Prof. John H.J. Einmahl, Prof. Bertrand Melenberg, and Prof. Irène Gijbels, for their insightful comments and suggestions. Last but not the least, I would like to thank my family: my parents, my sister, and my wife for supporting me throughout the years.

Shanghai, China
February 2018

Contents

Chapter 1

Introduction

Parametric regression modeling imposes strong restrictions on the functional form of regression. Even though their statistical properties are well established, the functional forms assumed in parametric models might be misspecified. Accordingly, many nonparametric and semiparametric regression models have been developed. In contrast to parametric modeling, nonparametric methods do not restrict the functional form, while semiparametric methods require only relatively weak prior restrictions. On the other hand, this flexibility can result in less precise estimation of parameters of interest.

Among semiparametric models, we study varying-coefficient models (also referred to as functional coefficient models) in the time series context (see Fan and Zhang, 2008, and Park et al., 2015, for an overview) which have the form:

$$y_t = x_t^\top a(z_t) + \varepsilon_t, \tag{1.1}$$

where y_t is a response, $a(z_t)$ is a vector of continuously differentiable functions of an observed transition variable z_t, x_t is a vector of covariates which might contain lagged responses, and ε_t is the error term satisfying $\mathrm{E}[\epsilon_t|x_t, z_t] = 0$. Model (1.1) can be treated as a linear model with interaction terms between the covariates x_t and transition variable z_t, where z_t is allowed to have a flexible form in the interaction term. In my dissertation, we introduce three new models based on model (1.1), and propose estimation procedures. In Chapter 2, we study the case that the coefficient functions $a(\cdot)$ are piecewise continuous. In Chapter 3, we restrict the coefficient functions to a parametric form: $a(\cdot) = \beta_1 w(\cdot) +$

$\beta_2\{1 - w(\cdot)\}$, where $w(\cdot)$ is an unknown smooth function of a scalar variable z_t, the so-called transition function, and β_1 and β_2 are slope parameters. In Chapter 4, we relax the zero conditional mean restriction $\mathrm{E}(\varepsilon_t|x_t, z_t) = 0$ such that the covariates x_t and transition variable z_t are allowed to be correlated with the error term ε_t. Summaries for each chapter are given below.

Chapter 2 considers a varying-coefficient model, where the coefficient functions $a(\cdot)$ are allowed to exhibit discontinuities at a finite set of points. We propose an estimation method builds upon the procedure in Gijbels et al. (2007). Contrary to Gijbels et al.'s nonparametric model with fixed regressors and independent homoscedastic errors, this chapter considers functional coefficient models in a random design and time-series context with serially correlated and heteroscedastic errors. Additionally, we consider two cases for the conditional variance function $\mathrm{E}(\varepsilon_t^2|z_t = z)$: one is continuous in the support of z_t; the other is discontinuous at a finite set of points. The consistency and asymptotic normality of the two proposed estimators are established in Theorems 2.5, 2.6, 2.9, and 2.10. The finite-sample performance is studied in a simulation study, showing that accounting for the discontinuity of the conditional variance is in general necessary for consistent estimation, but it does not worsen the performance of the estimators if the conditional variance is a continuous function of z_t.

Chapter 3 introduces a new semiparametric model – the semiparametric transition (SETR) model – that generalizes the models originally studied by Chan and Tong (1986) and Lin and Teräsvirta (1994) by letting the transition function $w(\cdot)$ to be of an unknown form. The estimation strategy is based on the iterative least squares. Consistency and the asymptotic distribution for the slope estimators of β_1 and β_2 are derived in Theorems 3.3 and 3.6, respectively. Monte Carlo simulations demonstrate that the proposed estimation of the SETR model provides precise estimates for many types of transition function, while the above mentioned parametric transition models can exhibit substantial biases.

Finally, Chapter 4 studies a functional coefficient instrumental variable model with endogenous x_t and z_t. Relying on the conditional mean-independence restriction (4.2), the functions in $a(\cdot)$ are identified according to Theorem 4.1. We propose a two-stage estimation procedure based upon local polynomial fitting and marginal integration techniques.

The estimator is shown to be consistent and asymptotically normal under weak dependence conditions in Theorem 4.4. Simulation evidence suggests the proposed estimator performs equally well as the two-stage estimator of Cai et al. (2006) in the case of an exogenous z_t. And our estimator also works for an endogenous transition variable z_t.

Chapter 2

Jump-Preserving Functional Coefficient Models for Nonlinear Time Series*

2.1 Introduction

The varying-coefficient models (VCM) form an important class of semiparametric models (see Hastie and Tibshirani, 1993; Cai et al., 2000) that assume the marginal effects of covariates to be an unknown function of an observable index variable. Practically, VCMs are formulated as linear models with coefficients being general functions of the index variable. Most existing literature assumes the coefficient functions to be continuous and smooth. In this chapter, we however allow coefficient functions to contain a finite set of discontinuities; additionally, discontinuities can be present also in the conditional error variance. This allows applying the flexible varying-coefficient modeling in parts of economics, biomedicine, epidemiology and other areas, where conditional expectations are known to exhibit jumps. For example, discontinuous coefficient functions are found by Čížek and Koo (2017b) in the dynamic models of GDP, by Zhao et al. (2017) in the time-varying capital asset pricing models, or by Bai and Perron (2003) and Zhao et al. (2016) in the models of inflation. Additionally, estimation of coefficient discontinuities lies at the

*This chapter is based on Čížek and Koo (2017a), Jump-preserving functional-coefficient models for nonlinear time series. CentER Discussion Paper 2017-017, Tilburg University.

core of the regression discontinuity designs (Lee and Lemieux, 2010), and although the location of the design discontinuity of often assumed, it is important to detect presence of other discontinuities if they exist. Besides that, Porter and Yu (2015) suggest the regression discontinuity modeling with an unknown location of the discontinuity point.

To the best of our knowledge, VCMs with discontinuities in coefficient functions have not been investigated before in heteroskedastic and time series setting. For independent and identically distributed data, Zhu et al. (2014) and Zhao et al. (2016) suggested methods for estimation of varying-coefficient models with discontinuities. On the other hand, there is a vast amount of literature on VCMs when coefficients are smooth continuous functions. Recent works include Hoover et al. (1998), Wu et al. (1998), and Fan and Zhang (2000) on longitudinal data analysis, Cai et al. (2000) and Huang and Shen (2004) on nonlinear time series, and Cai and Li (2008) and Sun et al. (2009) on panel data analysis. Additionally, hybrids of varying-coefficient models have also been developed: for example, partial linearly varying-coefficient models where some coefficient functions are constant (Zhang et al., 2002; Fan and Huang, 2005; Ahmad et al., 2005; Lee and Mammen, 2016), generalized linear models with varying coefficients (Cai et al., 2000), and varying-coefficient models in which the varying index is latent and estimated as a linear combination of several observed variables (Fan et al., 2003).

Although only a few studies on VCMs allow discontinuities in coefficient functions, literature on nonparametric estimation of discontinuous regression function is extensive. The classical estimation procedures usually consist of two stages. The locations of discontinuities are first estimated and then a conventional nonparametric estimator, which assumes the underlying function to be continuous, is used within each segment between two consecutive discontinuities to estimate the regression function itself. Examples of this approach include Müller (1992), Wu and Chu (1993), Kang et al. (2000), and Gijbels and Goderniaux (2004).

There are other techniques that do not estimate first the locations of discontinuities in a nonparametric regression; see, for example Godtliebsen et al. (1997) on nonlinear Gaussian filtering and Spokoiny (1998) and Polzehl and Spokoiny (2000) on adaptive weights smoothing. Besides these approaches, Gijbels et al. (2007) recently proposed an estimation method based on three local linear estimators in the framework of fixed design

and homoscedastic errors. At each design point z, they considered local linear estimates using data from the left-, right-, and two-sided neighborhoods of z. The final estimate of the conditional mean of the response equals one of these three local linear estimates chosen by comparing the weighted residual mean squared errors of three local linear fits. This approach was extended to conditional variance estimation by Casas and Gijbels (2012).

We generalize the estimation procedure by Gijbels et al. (2007) in two directions. First, we extend Gijbels et al. (2007) estimation method based on a comparison of the weighted residual mean squared errors to the VCMs, where discontinuities might occur only in one, few, or all coefficients. Although this has already been done by Zhao et al. (2016) in the case of independently and identically sampled observations, we analyze this method in the context of heteroskedastic and dependent data and provide additional asymptotic results such as the uniform convergence rate of the coefficient estimates. Second, as the method is shown to work well only if the conditional variance function of the error term is continuous, we propose an alternative measure of the three local linear fits based on the local Wald test statistics such that the proposed method is applicable even if the conditional variance function of the error term contains discontinuities.

This chapter is structured as follows. In Section 2.2, the VCM is introduced and the jump-preserving estimation procedure is introduced based on Gijbels et al. (2007) and Zhao et al. (2016). In Section 2.3, we establish the consistency and asymptotic normality of this estimator. In Section 2.4, an alternative estimator that does not require the continuity of conditional error variance is proposed and its asymptotic properties are derived. Finally, the finite sample properties of the two proposed estimators are investigated by means of a simulation study in Section 2.5. Proofs can be found in Sections 2.8 and 2.9.

2.2 The discontinuous varying-coefficient model

The varying-coefficient regression model takes the following form:

$$Y_i = X_i^\top a(Z_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where Y_i is the response variable, X_i is a $p \times 1$ vector of covariates, Z_i is a scalar index variable, $a(\cdot)$ is a $p \times 1$ vector of unspecified coefficient functions, and ε_i is an error term such that $\mathrm{E}[\varepsilon_i|X_i, Z_i] = 0$ and $\mathrm{E}[\varepsilon_i^2|X_i, Z_i] = \sigma^2(X_i, Z_i)$. Note that both X_i and Z_i can contain lagged values of Y_i. In this chapter, we consider piecewise-smooth coefficient functions $a(\cdot)$ that can exhibit a finite set of discontinuities located at points $\{s_q\}_{q=1}^{Q}$, where the number Q of jumps, the jump locations s_q, and the jump sizes d_q of the coefficient functions are all unknown. Contrary to Zhao et al. (2016), we assume that the conditional variance $\sigma^2(z) = \mathrm{E}[\sigma^2(X, Z)|Z = z]$ is not constant, but it is a continuous function of z in this section. The case with discontinuous $\sigma^2(z)$ will be investigated later in Section 2.4.

The semiparametric model (2.1) has been studied by Zhao et al. (2016) for the independent and identically distributed data, and in the present setting, it includes many popular time-series models. When X_i is a constant, the model is reduced to a nonparametric jump-preserving model in Gijbels et al. (2007). If all coefficient functions are constant, the model becomes a linear (possibly autoregressive) model. If the coefficient functions have the form: $a(\cdot) = \beta_1 w(\cdot) + \beta_2\{1 - w(\cdot)\}$ with $w(\cdot)$ being an unspecified scalar function, model (2.1) covers semiparametric transition models such as the one by Čížek and Koo (2017b), who estimated $w(\cdot)$ by a jump-preserving estimation proposed in this work. Moreover, model (2.1) includes the threshold autoregressive model and the smooth transition autoregressive model when $w(\cdot)$ takes a particular parametric form.

To define first the estimator of coefficient functions $a(\cdot)$ analogous to Gijbels et al. (2007) and Zhao et al. (2016), we let $K^{(c)}(\cdot)$ be a conventional bounded symmetric kernel function with a compact support $[-1, 1]$ and define $K^{(l)}(\cdot)$ and $K^{(r)}(\cdot)$ to be the corresponding left-sided and right-sided kernels, respectively, given by

$$K^{(l)}(v) = K^{(c)}(v) \cdot \mathbf{1}\{v \in [-1, 0)\} \quad \text{and} \quad K^{(r)}(v) = K^{(c)}(v) \cdot \mathbf{1}\{v \in [0, 1]\}, \tag{2.2}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. Using these kernels, we can define three pairs of local linear estimators $\hat{a}_n^{(\iota)}(z)$ and $\hat{b}_n^{(\iota)}(z)$ $(\iota = c, l, r)$ of coefficient functions $a(\cdot)$ and its

derivatives $a'(\cdot)$, respectively, at a fixed point z:

$$\left[\hat{a}_n^{(\iota)}(z), \hat{b}_n^{(\iota)}(z)\right] = \arg\min_{a,b} \sum_{i=1}^{n} \left\{Y_i - X_i^\top [a + b(Z_i - z)]\right\}^2 K_h^{(\iota)}(Z_i - z), \quad \iota = c, l, r, \tag{2.3}$$

where $K_h^{(\iota)}(\cdot) = h_n^{-1} K^{(\iota)}(\cdot/h_n)$, $h_n > 0$ is a bandwidth such that $h_n \to 0$ as $n \to \infty$ and the superscript $\iota = c, l, r$ indicates whether the conventional, left-sided, or right-sided kernel is used. Solving the least-squares minimization problem (2.3) for $\iota = c, l, r$ yields

$$\begin{pmatrix} \hat{a}_n^{(\iota)}(z) \\ \hat{b}_n^{(\iota)}(z) \end{pmatrix} = \left[\sum_{i=1}^{n} \begin{pmatrix} X_i \\ X_i(Z_i - z) \end{pmatrix} \begin{pmatrix} X_i \\ X_i(Z_i - z) \end{pmatrix}^\top K_h^{(\iota)}(Z_i - z)\right]^{-1} \sum_{i=1}^{n} \begin{pmatrix} X_i \\ X_i(Z_i - z) \end{pmatrix} Y_i K_h^{(\iota)}(Z_i - z). \tag{2.4}$$

To measure the quality of each local linear fit, Gijbels et al. (2007) and Zhao et al. (2016) advocate the use of the weighted residual mean squared error (WRMSE):

$$\Psi_n^{(\iota)}(z) = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_{n,i}^{(\iota)^2} K_h^{(\iota)}(Z_i - z)}{\sum_{i=1}^{n} K_h^{(\iota)}(Z_i - z)}, \quad \iota = c, l, r, \tag{2.5}$$

where the estimated residual $\hat{\varepsilon}_{n,i}^{(\iota)} = Y_i - X_i^\top \{\hat{a}_n^{(\iota)}(z) + \hat{b}_n^{(\iota)}(z)(Z_i - z)\}$. WRMSE is an estimator of conditional error variance $\sigma^2(z)$, which is similar to the one proposed in Fan and Yao (1998) except that the local constant fitting of $\hat{\varepsilon}_{n,i}^{(\iota)^2}$ and same bandwidth h_n for the conditional variance are used here. Although employing a different bandwidth for the conditional variance would improve the finite sample performance, our aim is to compare performance of the three local estimates of $a(z)$ rather than providing a good estimate of $\sigma^2(z)$. To avoid technical complexity in the proofs, the same bandwidth is therefore applied for the coefficient functions and WRMSE estimates.

The WRMSE introduced in (2.5) can be now used to select the consistent estimator out of (2.3) and thus to define the jump-preserving estimator of $a(z)$, which will be proved

consistent if the conditional error variance $\sigma^2(z)$ is continuous (cf. Zhao et al., 2016):

$$\check{a}_n(z) = \begin{cases} \hat{a}_n^{(c)}(z), & \text{if } \operatorname{diff}(z) \leq u_n, \\ \hat{a}_n^{(l)}(z), & \text{if } \operatorname{diff}(z) > u_n \text{ and } \Psi_n^{(l)}(z) < \Psi_n^{(r)}(z), \\ \hat{a}_n^{(r)}(z), & \text{if } \operatorname{diff}(z) > u_n \text{ and } \Psi_n^{(l)}(z) > \Psi_n^{(r)}(z), \\ \dfrac{\hat{a}_n^{(l)}(z) + \hat{a}_n^{(r)}(z)}{2}, & \text{if } \operatorname{diff}(z) > u_n \text{ and } \Psi_n^{(l)}(z) = \Psi_n^{(r)}(z), \end{cases} \tag{2.6}$$

where $\operatorname{diff}(z) = \Psi_n^{(c)}(z) - \min\{\Psi_n^{(l)}(z), \Psi_n^{(r)}(z)\}$ and the auxiliary parameter $u_n > 0$ is tending to zero, $u_n \to 0$ as $n \to \infty$. The intuition behing this proposal is based on the fact that the conventional local estimate $\hat{a}_n^{(c)}(z)$ should be the most precise one as it uses all observations in the interval $[z - h_n, z + h_n]$, but it is consistent only if there are no discontinuities in $(z-h_n, z+h_n)$. If $a(\cdot)$ is discontinuous at some point of $(z-h_n, z+h_n)$, $\hat{a}_n^{(c)}(z)$ is generally inconsistent (and the same can be also true in the case of $\hat{a}_n^{(l)}(z)$ or $\hat{a}_n^{(r)}(z)$), which leads to an increase of the corresponding WRMSE value in (2.5) as we confirm later in Section 2.3. Consequently, only a consistent estimator will minimize (2.5) asymptotically and will be thus selected in (2.6). The existence of a consistent estimator among $\hat{a}_n^{(c)}(z)$, $\hat{a}_n^{(l)}(z)$, and $\hat{a}_n^{(r)}(z)$ can be however assumed as bandwidth $h_n \to 0$ as $n \to \infty$ and the interval $(z - h_n, z + h_n)$ thus contains at most one point of discontinuity for any z and a sufficiently large n. See Zhao et al. (2016) for more details.

2.3 Asymptotic results

To derive the asymptotic properties of the proposed jump-preserving estimator, the assumptions about the data generating process (2.1) have to be detailed first. Later, the requirements on the kernel function and bandwidth are specified too.

Let us now define the α-mixing and the assumptions on the model (2.1). Suppose that $\mathcal{F}_a^b$ is the σ-algebra generated by $\{\xi_i; a \leq i \leq b\}$. The α-mixing coefficient of the process $\{\xi_i\}_{i=-\infty}^{\infty}$ is defined as

$$\alpha(m) = \sup\{|P(A \cap B) - P(A)P(B)| : A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_m^{\infty}\}.$$

If $\alpha(m) \to 0$ as $m \to \infty$, then the process $\{\xi_i\}_{i=-\infty}^{\infty}$ is called strong mixing or α-mixing. In the following assumptions, we additionally denote by $f(\cdot,\cdot)$ the joint probability density function of variables X_i and Z_i and by $f_Z(\cdot)$ the marginal density function of Z_i.

Assumption 2.A.

2.A1. The process $\{X_i, Z_i, \varepsilon_i\}$ is strictly stationary and strong mixing with α-mixing coefficients $\alpha(m)$, $m \in \mathbb{N}$, that satisfy $\alpha(m) \leq Cm^{-\gamma}$ with $0 < C < \infty$ and $\gamma > (2\delta - 2)/(\delta - 2)$ for some $\delta > 2$.

2.A2. There is a compact set $D = [s_0, s_{Q+1}]$ such that $\inf_{z\in D} f_Z(z) > 0$. The derivative of $f_Z(\cdot)$ is bounded and Lipschitz continuous for $z \in D$. The partial derivative of the joint density function $f(\cdot,\cdot)$ with respect to Z is bounded and continuous uniformly on the support of X and D except for the points $\{s_q\}_{q=0}^{Q+1}$, at which the left and right partial derivatives of $f(\cdot,\cdot)$ with respect to Z are bounded and left and right continuous, respectively.

2.A3. Let φ_i represent any element of matrix $X_iX_i^\top$, vector $X_i\varepsilon_i$, or variable ε_i^2. For δ given in Assumption 2.A1,

(i) $\mathrm{E}|\varphi_i|^\delta < \infty$,

(ii) $\sup_{z\in D} \mathrm{E}(|\varphi_i|^\delta | Z_i = z) f_Z(z) < \infty$,

(iii) for all integers $j > 1$,

$$\sup_{(z_1,z_j)\in D\times D} \mathrm{E}(|\varphi_1\varphi_j||Z_1 = z_1, Z_j = z_j) f_{Z_1Z_j}(z_1, z_j) < \infty,$$

where $f_{Z_1Z_j}(z_1, z_j)$ denotes the joint density of (Z_1, Z_j).

2.A4. The variance matrix $\Omega(z) = \mathrm{E}[XX^\top|Z = z]$ is bounded and positive definite uniformly on D except for the discontinuities $\{s_q\}_{q=0}^{Q+1}$, at which variance matrices $\Omega_-(s_q) = \lim_{z\uparrow s_q} \mathrm{E}[XX^\top|Z = z]$ and $\Omega_+(s_q) = \lim_{z\downarrow s_q} \mathrm{E}[XX^\top|Z = z]$ are bounded and positive definite.

2.A5. The second-order partial derivatives of $a(z)$ are bounded and Lipschitz continuous on D except for the discontinuities $\{s_q\}_{q=0}^{Q+1}$, at which $a(z)$ defined to be left and right continuous has the left and right second-order partial derivatives that are bounded and left and right Lipschitz continuous, respectively.

2.A6. The partial derivative of $\sigma^2(x,z)$ with respect to z is bounded and continuous on D.

Assumptions 2.A1–2.A5 are standard conditions for the VCMs with dependent data (see e.g. Conditions A.1 and A.2 in Cai et al. (2000) for the local linear estimation in VCMs and the assumptions in Hansen (2008) for a general nonparametric kernel estimator) adapted for discontinuities, at which we impose the corresponding conditions for the left and right limits. Further, Assumption 2.A6 imposes that the conditional variance $\sigma^2(z) = \mathrm{E}[\sigma^2(X,Z)|Z=z]$ is continuous; the case with discontinuous $\sigma^2(z)$ is investigated in Section 2.4.

The following assumptions about the kernel K, bandwidth h_n, auxiliary parameter u_n, and mixing exponent γ are also needed to show the asymptotic results for the jump-preserving estimator $\check{a}_n(z)$. First, standard assumptions on the kernel and bandwidth are given. After that, assumptions required by Hansen (2008) in the asymptotic analysis of the local linear regression estimators under dependence are introduced.

Assumption 2.B.

2.B1. The kernel $K^{(c)}(\cdot)$ is a bounded symmetric continuous density function and has a compact support $[-1,1]$. It is chosen so that the following constants are well defined and finite for $j=0,1,2$ and $\iota = c, r, l$:

$$\mu_j^{(\iota)} = \int_{-1}^{1} v^j K^{(\iota)}(v)\mathrm{d}v, \qquad \nu_j^{(\iota)} = \int_{-1}^{1} v^j K^{(\iota)2}(v)\mathrm{d}v,$$
$$c_0^{(\iota)} = \frac{\mu_2^{(\iota)}}{\mu_2^{(\iota)}\mu_0^{(\iota)} - \mu_1^{(\iota)2}}, \quad \text{and} \quad c_1^{(\iota)} = \frac{-\mu_1^{(\iota)}}{\mu_2^{(\iota)}\mu_0^{(\iota)} - \mu_1^{(\iota)2}}. \tag{2.7}$$

2.B2. The bandwidths h_n and u_n satisfy $u_n \to 0$, $h_n \to 0$, and $nh_n \to \infty$ as $n \to \infty$.

2.B3. Additionally, $nh_n^5 \to \bar{c} \in [0, +\infty)$ as $n \to \infty$, where $\bar{c}$ is some non-negative constant.

Assumption 2.C.

2.C1. The functions $\mathcal{K}_j^{(c)}(u) = u^j K^{(c)}(u)$ are Lipschitz continuous for all $j = 0,1,2,3$.

2.C2. For some $\varsigma \geq 1$, the strong mixing exponent γ given in Assumption 2.A1 satisfies

$$\gamma > \frac{1 + (\delta - 1)(2 + 1/\varsigma)}{\delta - 2}.$$

2.C3. The bandwidth h_n satisfies $\ln n/(nh_n^3) = \mathrm{o}(1)$ and $\ln n/(n^\theta h_n) = \mathrm{o}(1)$, where

$$\theta = \frac{\gamma - 2 - \frac{1}{\varsigma} - \frac{1+\gamma}{\delta - 1}}{\gamma + 2 - \frac{1+\gamma}{\delta - 1}}.$$

Note that the above assumptions impose that the bandwidth sequence $h_n \sim n^{-\alpha}$ for $\alpha \in [1/5, \min\{1/3, \theta\})$, where the upper bound depends on the mixing coefficient γ and the number of moments δ. For example for the exponentially mixing series, $\gamma = \infty$ and $\theta = 1 - (\delta - 1)/[\varsigma(\delta - 2)]$ can be made arbitrarily close to 1 for any δ by selecting a sufficiently large ς. If γ becomes finite and small, $\delta > 2$ will however have to be sufficiently large to ensure that $\theta > 1/5$.

Before providing the asymptotic properties of the jump-preserving estimator $\check{a}_n(z)$, we study the behavior of the three local linear estimators (2.3) in the continuous region and in the neighborhoods of discontinuities. The regions of continuity are defined by

$$D_{1n} = D_{1n}^{(c)} = D \setminus \bigcup_{q=0}^{Q+1} [s_q - h_n, s_q + h_n],$$
$$D_{1n}^{(l)} = D \setminus \bigcup_{q=0}^{Q+1} [s_q, s_q + h_n], \quad \text{and} \quad D_{1n}^{(r)} = D \setminus \bigcup_{q=0}^{Q+1} [s_q - h_n, s_q].$$

Theorem 2.1. *Under Assumptions 2.A1–2.A6, 2.B, and 2.C, it holds for $n \to \infty$ that*

$$\sup_{z \in D_{1n}^{(\iota)}} \left\| \hat{a}_n^{(\iota)}(z) - a(z) \right\| = \mathrm{O_p}\left(\sqrt{\frac{\ln n}{nh_n}} \right), \quad \iota = c, l, r.$$

Theorem 2.2. *If Assumptions 2.A1–2.A6 and 2.B are satisfied and a fixed point $z \in D_{1n}^{(\iota)}$ for some $n \in \mathbb{N}$ and $\iota = c, l, r$, it holds that*

$$\sqrt{nh_n} \left[\hat{a}_n^{(\iota)}(z) - a(z) - \frac{h_n^2}{2} \left(c_0^{(\iota)} \mu_2^{(\iota)} + c_1^{(\iota)} \mu_3^{(\iota)} \right) a''(z) \right] \xrightarrow{d} N\left(0, \Phi^{(\iota)}(z)\right)$$

as $n \to \infty$*, where*

$$\Phi^{(\iota)}(z) = \frac{c_0^{(\iota)2}\nu_0^{(\iota)} + 2c_0^{(\iota)}c_1^{(\iota)}\nu_1^{(\iota)} + c_1^{(\iota)2}\nu_2^{(\iota)}}{f_Z(z)} \cdot \Omega^{-1}(z)\Theta(z)\Omega^{-1}(z), \tag{2.8}$$

$\Omega(z) = \mathrm{E}[XX^\top|Z = z]$*, and* $\Theta(z) = \mathrm{E}[XX^\top\sigma^2(X, Z)|Z = z]$.

Theorem 2.1 establishes the uniform consistency of the three local linear estimators in their corresponding continuous regions. Theorem 2.2 then specifies the asymptotic distributions of the estimators $\hat{a}_n^{(c)}(z)$, $\hat{a}_n^{(l)}(z)$, and $\hat{a}_n^{(r)}(z)$ in the regions, where $a(\cdot)$ is continuous, left-continuous, and right-continuous around z, respectively. The stated bias term and asymptotic variance correspond to that derived in the iid case by Zhao et al. (2016) in their proof of Proposition 2.1. The asymptotic variance has the standard form of the local least-squares estimator except for the numerator of the fraction in (2.8), which however reduces to standard $c_0^{(\iota)2}\nu_0^{(\iota)}$ in the case of the centered estimation.

Since all three local linear estimators are consistent in their corresponding regions of continuity according to Theorem 2.1, it is easy to see that their corresponding WRMSE estimates (2.5) consistently converge to the conditional error variance $\sigma^2(z)$.

Theorem 2.3. *Let Assumptions 2.A1–2.A6 and 2.B hold. At any point* $z \in D_{1n}^{(\iota)}$ *for some* $n \in \mathbb{N}$ *and* $\iota = c, l, r$*, the mean squared error in (2.5) satisfies* $\Psi_n^{(\iota)}(z) \xrightarrow{P} \sigma^2(z)$ *as* $n \to \infty$.

Such a result does not however hold if the point z is close to a jump, that is, to a point of discontinuity. If a jump is located in the right neighborhood of z, only the left-sided local linear estimator $\hat{a}_n^{(l)}(z)$ is consistent. Similarly, the right-sided estimator $\hat{a}_n^{(r)}(z)$ is the only consistent estimator of $a(z)$ when there is a jump in the left neighborhood of z. Consequently, the three WRMSE estimates behave differently near a jump point. The next theorem describes the asymptotic behavior of WRMSE in a neighborhood of a jump s_q when the conditional error variance $\sigma^2(z)$ is continuous in z (cf. Zhao et al., 2016).

Theorem 2.4. *Let Assumptions 2.A1–2.A6 and 2.B hold. Then it holds as* $n \to \infty$ *that*

(i) *for any* $z = s_q + \tau h_n \in D$ *with* $q = 1, \ldots, Q+1$ *and* $\tau \in [-1, 0)$,

$$\begin{aligned}
\Psi_n^{(c)}(z) &\xrightarrow{P} \sigma^2(s_q) + d_q^\top C_\tau^{(c)} d_q, \\
\Psi_n^{(l)}(z) &\xrightarrow{P} \sigma^2(s_q), \\
\Psi_n^{(r)}(z) &\xrightarrow{P} \sigma^2(s_q) + d_q^\top C_\tau^{(r)} d_q.
\end{aligned}$$

(ii) *for any* $z = s_q + \tau h_n \in D$ *with* $q = 0, \ldots, Q$ *and* $\tau \in (0, 1]$,

$$\begin{aligned}
\Psi_n^{(c)}(z) &\xrightarrow{P} \sigma^2(s_q) + d_q^\top C_\tau^{(c)} d_q, \\
\Psi_n^{(l)}(z) &\xrightarrow{P} \sigma^2(s_q) + d_q^\top C_\tau^{(l)} d_q, \\
\Psi_n^{(r)}(z) &\xrightarrow{P} \sigma^2(s_q).
\end{aligned}$$

In both cases, $d_q = \lim_{z \downarrow s_q} a(z) - \lim_{z \uparrow s_q} a(z)$ *and* $C_\tau^{(\iota)}, \iota = c, l, r$, *represents a positive definite matrix defined in Section 2.8, equation* (2.40).

The above theorem shows that only the left-sided WRMSE is a consistent estimator of the conditional error variance $\sigma^2(z)$ if a jump in coefficients $a(z)$ occurs in the right neighborhood of z, while the other two WRMSE estimates contain strictly positive biases, which do not vanish asymptotically. Similarly, if a jump is in the left neighborhood of z, only the right-sided WRMSE leads to a consistent estimator of $\sigma^2(z)$. To sum up, the smallest WRMSE is – at least asymptotically – $\Psi_n^{(l)}(z)$ when a jump is in a right neighborhood of z and it is $\Psi_n^{(r)}(z)$ when a jump is in a left neighborhood of z. Hence, it is intuitively clear that the jump-preserving estimator $\check{a}_n(z)$ defined in (2.6) selects the appropriate local linear estimator at every point z for a sufficiently large n.

Based on this result, we will establish the consistency of $\check{a}_n(z)$ in the continuous region D_{1n}, in the neighborhoods of discontinuity points D_{2n}, and in the neighborhoods of discontinuity points excluding small regions around centers and around endpoints $D_{2n,\delta}$.

These regions are defined as follows:

$$D_{2n} = D \cap \bigcup_{q=0}^{Q+1} \{[s_q - h_n, s_q) \cup (s_q, s_q + h_n]\} \quad \text{and}$$
$$D_{2n,\delta} = D \cap \bigcup_{q=0}^{Q+1} \{[s_q - (1-\delta)h_n, s_q - \delta h_n] \cup [s_q + \delta h_n, s_q + (1-\delta)h_n]\} \tag{2.9}$$

for some $\delta \in (0, 1/2)$.

Theorem 2.5. *If Assumptions 2.A1–2.A6, 2.B, and 2.C are satisfied, it holds for $n \to \infty$ and some $\delta \in (0, 1/2)$ that*

(i)
$$\sup_{z \in D_{1n}} \|\check{a}_n(z) - a(z)\| = \mathrm{O_p}\left(\sqrt{\frac{\ln n}{n h_n}}\right),$$

(ii)
$$\sup_{z \in D_{2n,\delta}} \|\check{a}_n(z) - a(z)\| = \mathrm{O_p}\left(\sqrt{\frac{\ln n}{n h_n}}\right), \text{ and}$$

(iii) *for any $z \in D_{2n}$,*
$$\|\check{a}_n(z) - a(z)\| = \mathrm{O_p}\left(\sqrt{\frac{\ln n}{n h_n}}\right).$$

Theorem 2.5 states that the jump-preserving estimator $\check{a}_n(z)$ is uniformly consistent on D_{1n} and $D_{2n,\delta}$ for some $\delta \in (0, 1/2)$. At a point $z \in D_{2n}$ arbitrarily close to a point of discontinuity, $\check{a}_n(z)$ is only pointwise consistent.

The jump-preserving estimator $\check{a}_n(z)$ selects consistently (i.e., with probability approaching to 1) the appropriate local linear estimator on D excluding the jump points, where each of these local linear estimators is asymptotically normal at any point $z \in D \setminus \{s_q\}_{q=0}^{Q+1}$ according to Theorem 2.2. The following theorem can therefore establish the asymptotic normality of the jump-preserving estimator $\check{a}_n(z)$ at $z \in D \setminus \{s_q\}_{q=0}^{Q+1}$ (see also Casas and Gijbels, 2012; Zhao et al., 2016, Theorems 3.1).

Theorem 2.6. *If Assumptions 2.A1–2.A6, 2.B, and 2.C are satisfied and* $z \in D \setminus \{s_0, \ldots, s_{Q+1}\}$*, it holds that*

$$\sqrt{nh_n}\left[\check{a}_n(z) - a(z) - \frac{h_n^2}{2}\left(c_0^{(\iota)}\mu_2^{(\iota)} + c_1^{(\iota)}\mu_3^{(\iota)}\right)a''(z)\right] \xrightarrow{d} N\left(0, \Phi^{(\iota)}(z)\right)$$

as $n \to \infty$*, where* $\Phi^{(\iota)}(z)$ *is defined in equation* (2.8) *and*

$$\iota = \begin{cases} c, & \text{if } z \in D_{1n}, \\ l, & \text{if } z \in D \cap \bigcup_{q=0}^{Q+1} [s_q - h_n,\ s_q), \\ r, & \text{if } z \in D \cap \bigcup_{q=0}^{Q+1} (s_q,\ s_q + h_n]. \end{cases}$$

2.4 Discontinuous conditional variance function

In this section, the conditional variance function $\sigma^2(z)$ is also allowed to exhibit discontinuities. For this purpose, we replace Assumption 2.A6 by the following condition.

Assumption A6'. The partial derivative of $\sigma^2(x, z)$ with respect to z is bounded and continuous on D except for the points of discontinuity $\{\tilde{s}_q\}_{q=0}^{\tilde{Q}+1}$, at which $\sigma^2(x, z)$ defined to be left and right continuous has the left and right partial derivatives with respect to z that are bounded and left and right continuous, respectively.

Given the possibility of discontinuities of the variance functions $\sigma^2(z)$ and $\sigma^2(x, z)$ in Assumption A6', the subscripts '$-$' and '$+$' will now denote the corresponding left and right limits of these variance functions. Although the variance discontinuities introduced in Assumption A6' do not influence the consistency and convergence rates of the three local estimators (2.3), they can adversely affect the selection rule (2.6) based on a comparison of the three WRMSE estimates. In particular, if $\sigma^2(z)$ exhibits a jump at (or nearby) s_q, the error variances and thus WRMSE estimates are different in the left and right neighborhoods of the estimation point z. Hence, the limits of $\Psi_n^{(c)}(z)$, $\Psi_n^{(l)}(z)$, and $\Psi_n^{(r)}(z)$ in Theorem 2.4 contain different variances – error variance to the left of s_q, to the right of s_q, or a combination of those – and it is no longer possible to claim that $\Psi_n^{(l)}(z)$ is minimal in Theorem 2.4(i) or that $\Psi_n^{(r)}(z)$ is minimal in Theorem 2.4(ii). In such cases, the selection method (2.6) fails to detect and preserve jumps. On the other hand, if $\sigma^2(z)$

exhibits a jump in the continuity region D_1, all local linear estimates are consistent, but for the reason stated above, the selection method (2.6) can still fail to select the best (conventional) estimate. Thus the consistency is not violated, but the variance of estimates can increase and the asymptotic distribution in Theorem 2.6 becomes incorrect.

To deal with the discontinuity of $\sigma^2(z)$, we introduce now an alternative jump-preserving estimator which does not require the continuity of conditional error variance. Let the left-, right-, and two-sided h_n-neighborhood of z be

$$D_{zn}^{(l)} = [z - h_n, z], \quad D_{zn}^{(r)} = [z, z + h_n], \quad \text{and} \quad D_{zn}^{(c)} = [z - h_n, z + h_n],$$

respectively. To motivate an alternative to the selection method (2.6), we first suppose that s_q is in the right neighborhood of z, i.e., $s_q \in D_{zn}^{(r)}$. In such a case, only the left-sided local linear estimates $\hat{a}_n^{(l)}(z)$ and $\hat{b}_n^{(l)}(z)$ converge to the true parameter values $a^{(l)}(z) = a(z)$ and $b^{(l)}(z) = a'(z)$, respectively. (We are again implicitly assuming that bandwidth h_n is so small that there is at most one jump in $(z-h_n, z+h_n)$ for a sufficiently large n.) By the Taylor expansion and $\mathrm{E}[g(X_i)\varepsilon_i|Z_i] = \mathrm{E}[g(X_i)\mathrm{E}[\varepsilon_i|X_i, Z_i]|Z_i] = 0$ for any bounded non-zero function $g(\cdot)$, we have (under some regularity assumptions)

$$\begin{aligned}
&\mathrm{E}[g(X_i)\{Y_i - X_i^\top a^{(l)}(z)\}|Z_i] \\
&= \mathrm{E}\left[g(X_i)X_i^\top\{a(Z_i) - a^{(l)}(z)\}|Z_i\right] \\
&\le \mathrm{E}\left[\|g(X_i)X_i^\top\||Z_i\right] \mathrm{E}\left[\|a(Z_i) - a^{(l)}(z)\||Z_i\right] \\
&= \mathrm{O}(Z_i - z) = \mathrm{O}(h_n) = \mathrm{o}(1).
\end{aligned}$$

for $Z_i \in D_{zn}^{(l)}$. On the other hand, the above result does not hold for the limit values of the right-sided and two-sided local linear estimators, $a^{(c)}(z)$ and $a^{(r)}(z)$, which are different from $a(z)$. Thus as long as the coefficient functions $a(\cdot)$ are identified and $a^{(\iota)}(z) \neq a(z)$, $\iota = c, r$, it holds for $Z_i \in D_{zn}^{(\iota)}$ that $\mathrm{E}[g(X_i)\{Y_i - X_i^\top a^{(\iota)}(z)\}|Z_i] = \mathrm{E}[g(X_i)X_i^\top\{a(Z_i) - a^{(\iota)}(z)\}|Z_i] \neq \mathrm{o}(1)$ for a general $g(\cdot)$; in particular, it holds for $g(X_i) = X_i$, and if $\mathrm{E}(Z_iX_i^\top|Z_i = z)$ has the full rank, even for $g(X_i) = 1$. Analogous claims can be made if s_q is in the left neighborhood of z. Given the focus on time series models, $g(X_i) = 1$ is considered for the sake of simplicity.

Contrary to (2.6), the asymptotic conditional mean independence described above is a property independent of conditional error variance $\sigma^2(z)$. To select the consistent estimator out of the three local linear estimators (2.3), we therefore propose to test locally whether $\mathrm{E}[\varepsilon_i^{(\iota)}|Z_i] = 0$ for $Z_i \in D_{zn}^{(\iota)}$ and $\iota = c, l, r$, where $\varepsilon_i^{(\iota)} = Y_i - X_i^\top a^{(\iota)}(z)$:[†] rejection of $\mathrm{E}[\varepsilon_i^{(\iota)}|Z_i] = 0$ indicates that a given local linear estimator is not consistent and should not be used in a given neighborhood of z. According to Bierens (1982, Theorems 1 and 2), the conditional mean independence $\mathrm{E}[\varepsilon_i^{(\iota)}|Z_i] = 0$ is equivalent to zero correlation between $\varepsilon_i^{(\iota)}$ and $\exp(kZ_i)$ for all $k \in \mathbb{R}$, or alternatively, to zero correlation between $\varepsilon_i^{(\iota)}$ and Z_i^k for all $k \in \mathbb{N} \cup \{0\}$. To design a simple procedure with a good power, we therefore suggest to test zero correlation between $\varepsilon_i^{(\iota)}$ and Z_i^k for $k = 1, \ldots, m$, where m is a small finite number. Given the specific form of $\mathrm{E}[\varepsilon_i^{(\iota)}|Z_i] = \mathrm{E}[\varepsilon_i + X_i^\top\{a(z) - a^{(\iota)}(z)\}|Z_i]$ caused by an unaccounted discontinuity in $a(z)$, the cubic polynomial approximates this expectation well and $m = 3$ provides a sufficient power to detect its nonlinearity even in small intervals $(z - h_n, z + h_n)$; see Section 2.5.

To test for non-zero correlation of $\varepsilon_i^{(\iota)}$ and Z_i^j, $j = 1, \ldots, m$, we propose to regress the estimated residual $\tilde{\varepsilon}_{n,i}^{(\iota)} = Y_i - X_i^\top \hat{a}_n^{(\iota)}(z)$ on $\rho\left(\frac{Z_i - z}{h_n}\right)$ for $Z_i \in D_{zn}^{(\iota)}$, where $\rho(v) = (1, v, \cdots, v^m)^\top$. The corresponding ordinary least-squares slope estimates $\hat{\gamma}_n^{(\iota)}(z)$ will converge to $\gamma^{(\iota)}(z) = 0$ under the null hypothesis of $\mathrm{E}[\varepsilon_i^{(\iota)}|Z_i] = 0, Z_i \in D_{zn}^{(\iota)}$, and to $\gamma^{(\iota)}(z) \neq 0$ otherwise (for sufficiently large m and n); $\iota = c, l, r$. More specifically, we test significance of the slope estimates $\hat{\gamma}_n^{(\iota)}(z)$ that are the minimizers of the following least square problem:

$$\min_{\gamma} \sum_{i=1}^{n} \left\{ \tilde{\varepsilon}_{n,i}^{(\iota)} - \rho^\top\left(\frac{Z_i - z}{h_n}\right)\gamma \right\}^2 \tilde{K}_h^{(\iota)}(Z_i - z), \tag{2.10}$$

where $\tilde{K}_h^{(\iota)}(\cdot) = h_n^{-1}\tilde{K}^{(\iota)}(\cdot/h_n)$, $\tilde{K}^{(c)}(\cdot)$ is the uniform kernel function on $[-1, 1]$,

$$\tilde{K}^{(l)}(v) = \tilde{K}^{(c)}(v) \cdot \mathbf{1}\{v \in [-1, 0)\}, \quad \text{and} \quad \tilde{K}^{(r)}(v) = \tilde{K}^{(c)}(v) \cdot \mathbf{1}\{v \in [0, 1]\}.$$

[†] A similar result holds also if the local linear approximation, $\varepsilon_i^{(\iota)} = Y_i - X_i^\top\{a(z) + b(z)(Z_i - z)\}$, is used.

Solving the minimization (2.10) leads to estimate $\hat{\gamma}_n^{(\iota)}(z) = \tilde{S}_n^{(\iota)-1}(z)\tilde{T}_n^{(\iota)}(z)$, where

$$\tilde{S}_n^{(\iota)}(z) = \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{Z_i - z}{h_n}\right) \rho^\top\left(\frac{Z_i - z}{h_n}\right) \tilde{K}_h^{(\iota)}(Z_i - z) \quad \text{and}$$
$$\tilde{T}_n^{(\iota)}(z) = \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{Z_i - z}{h_n}\right) \tilde{K}_h^{(\iota)}(Z_i - z)\tilde{\varepsilon}_{n,i}^{(\iota)}.$$

In order to test the hypothesis $\gamma^{(\iota)}(z) = 0$, the Wald test statistics is used here, which forms an alternative measure $\tilde{\Psi}_n^{(\iota)}(z)$ to the WRMSE $\Psi_n^{(\iota)}(z)$ introduced in (2.5) and provides an indication about the dependence between estimated residual and Z_i:

$$\tilde{\Psi}_n^{(\iota)}(z) = \hat{\gamma}_n^{(\iota)\top}(z) \left(\frac{\tilde{S}_n^{(\iota)}(z)}{\tilde{N}_n^{(\iota)}(z)}\right) \hat{\gamma}_n^{(\iota)\top}(z), \tag{2.11}$$

where

$$\hat{e}_{n,i}^{(\iota)}(z) = \tilde{\varepsilon}_{n,i}^{(\iota)} - \rho^\top\left(\frac{Z_i - z}{h_n}\right) \hat{\gamma}_n^{(\iota)}(z) \quad \text{and}$$
$$\tilde{N}_n^{(\iota)}(z) = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_{n,i}^{(\iota)^2}(z)\tilde{K}_h^{(\iota)}(Z_i - z).$$

For this quantity (2.11), we derive now theorems analogous to Theorems 2.3 and 2.4 for the case of the Wald measure $\tilde{\Psi}_n^{(\iota)}(z)$ under the following condition.

Assumption 2.D.

2.D1. The uniform kernel $\tilde{K}^{(c)}(\cdot)$ has support $[-1, 1]$ and the kernel moment matrix $\tilde{M}^{(\iota)} = \int_{-1}^{1} \rho(u)\rho^\top(u)\tilde{K}^{(\iota)}(u)\mathrm{d}u$, $\iota = c, l, r$, is positive definite.

2.D2. The number m of powers used in the auxiliary regressions (2.10) is sufficiently large such that at least one of the slope coefficients $\gamma_{q,\tau}^{(\iota)}$, which has its explicit expression given in equation (2.68), is non-zero for $z = s_q + \tau h_n$, $q = 0, \ldots, Q+1$, for any given $\tau \in (-1, 0)$ and $\iota = c, r$ and $\tau \in (0, 1)$ and $\iota = c, l$.

Theorem 2.7. *Suppose that Assumptions 2.A1–2.A5, A6', 2.B, and 2.D hold. At any $z \in D_{1n}^{(\iota)}$ for some $n \in \mathbb{N}$ and $\iota = c, l, r$, it holds that $\tilde{\Psi}_n^{(\iota)}(z) \xrightarrow{P} 0$ as $n \to \infty$.*

Theorem 2.8. *If Assumptions 2.A1–2.A5, A6', 2.B, and 2.D are satisfied, the following results hold as $n \to \infty$.*

(i) *For any* $z = s_q + \tau h_n \in D$ *with* $q = 1, \ldots, Q+1$ *and* $\tau \in (-1, 0)$,

$$\begin{aligned}
\tilde{\Psi}_n^{(c)}(z) &\xrightarrow{P} \gamma_{q,\tau}^{(c)\top} \tilde{C}_\tau^{(c)} \gamma_{q,\tau}^{(c)}, \\
\tilde{\Psi}_n^{(l)}(z) &\xrightarrow{P} 0, \\
\tilde{\Psi}_n^{(r)}(z) &\xrightarrow{P} \gamma_{q,\tau}^{(r)\top} \tilde{C}_\tau^{(r)} \gamma_{q,\tau}^{(r)}.
\end{aligned}$$

(ii) *For any* $z = s_q + \tau h_n \in D$ *with* $q = 0, \ldots, Q$ *and* $\tau \in (0, 1)$,

$$\begin{aligned}
\tilde{\Psi}_n^{(c)}(z) &\xrightarrow{P} \gamma_{q,\tau}^{(c)\top} \tilde{C}_\tau^{(c)} \gamma_{q,\tau}^{(c)}, \\
\tilde{\Psi}_n^{(l)}(z) &\xrightarrow{P} \gamma_{q,\tau}^{(l)\top} \tilde{C}_\tau^{(l)} \gamma_{q,\tau}^{(l)}, \\
\tilde{\Psi}_n^{(r)}(z) &\xrightarrow{P} 0.
\end{aligned}$$

In both cases for $\iota = c, l, r$, $\tilde{C}_\tau^{(\iota)}$ *is a positive definite matrix defined in Section 2.8, equation (2.71), and the explicit form of* $\gamma_{q,\tau}^{(\iota)}$ *is given in Section 2.8, equation (2.68).*

Given the above results, we can use the Wald statistics $\tilde{\Psi}_n^{(\iota)}(z)$ to again distinguish which local estimators $\hat{a}_n^{(\iota)}(z)$ are consistent or inconsistent due to a discontinuity of coefficient functions, but now without requiring that the conditional variance $\sigma^2(z)$ is continuous. We thus propose a new jump-preserving estimator $\tilde{a}_n(z)$ of coefficient functions $a(z)$ when the conditional error variance contains a finite set of discontinuities:

$$\tilde{a}_n(z) = \begin{cases} \hat{a}_n^{(c)}(z), & \text{if } \tilde{\text{diff}}(z) \leq u_n, \\ \hat{a}_n^{(l)}(z), & \text{if } \tilde{\text{diff}}(z) > u_n \text{ and } \tilde{\Psi}_n^{(r)}(z) > \tilde{\Psi}_n^{(l)}(z), \\ \hat{a}_n^{(r)}(z), & \text{if } \tilde{\text{diff}}(z) > u_n \text{ and } \tilde{\Psi}_n^{(l)}(z) > \tilde{\Psi}_n^{(r)}(z), \\ \dfrac{\hat{a}_n^{(l)}(z) + \hat{a}_n^{(r)}(z)}{2}, & \text{if } \tilde{\text{diff}}(z) > u_n \text{ and } \tilde{\Psi}_n^{(l)}(z) = \tilde{\Psi}_n^{(r)}(z), \end{cases} \tag{2.12}$$

where the auxiliary parameter $u_n > 0$ is again tending to zero with increasing n and $\tilde{\text{diff}}(z) = \tilde{\Psi}_n^{(c)}(z) - \min\{\tilde{\Psi}_n^{(l)}(z), \tilde{\Psi}_n^{(r)}(z)\}$. The consistency and asymptotic normality of the proposed jump-preserving estimator $\tilde{a}_n(z)$ are established in the following theorems.

Theorem 2.9. *Under Assumptions 2.A1–2.A5, A6', 2.B, 2.C, and 2.D, it holds for* $n \to \infty$ *and some* $\delta \in (0, 1/2)$ *that*

(i)
$$\sup_{z\in D_{1n}} \|\tilde{a}_n(z) - a(z)\| = \mathrm{O_p}\left(\sqrt{\frac{\ln n}{nh_n}}\right),$$

(ii)
$$\sup_{z\in D_{2n,\delta}} \|\tilde{a}_n(z) - a(z)\| = \mathrm{O_p}\left(\sqrt{\frac{\ln n}{nh_n}}\right), \text{ and}$$

(iii) *for any given* $z \in D_{2n}$,
$$\|\tilde{a}_n(z) - a(z)\|\mathrm{O_p}\left(\sqrt{\frac{\ln n}{nh_n}}\right).$$

Theorem 2.10. *If Assumptions 2.A1–2.A5, A6', 2.B, 2.C, and 2.D are satisfied and a point* $z \in D \setminus \{s_0, \ldots, s_{Q+1}\}$, *it holds that*

$$\sqrt{nh_n}\left[\tilde{a}_n(z) - a(z) - \frac{h_n^2}{2}\left(c_0^{(\iota)}\mu_2^{(\iota)} + c_1^{(\iota)}\mu_3^{(\iota)}\right)a''(z)\right] \xrightarrow{d} N\left(0, \Phi_{lr}^{(\iota)}(z)\right)$$

as $n \to \infty$, *where* $\Phi_{lr}^{(\iota)}(z) = f_Z^{-1}(z)\Omega^{-1}(z)\left(\Phi_l^{(\iota)}(z) + \Phi_r^{(\iota)}(z)\right)\Omega^{-1}(z)$,

$$\Phi_l^{(\iota)}(z) = \Theta_-(z)\left[c_0^{(\iota)2}\nu_0^{(l)} + 2c_0^{(\iota)}c_1^{(\iota)}\nu_1^{(l)} + c_1^{(\iota)2}\nu_2^{(l)}\right]\mathbf{1}(\iota \in \{c, l\})$$
$$\Phi_r^{(\iota)}(z) = \Theta_+(z)\left[c_0^{(\iota)2}\nu_0^{(r)} + 2c_0^{(\iota)}c_1^{(\iota)}\nu_1^{(r)} + c_1^{(\iota)2}\nu_2^{(r)}\right]\mathbf{1}(\iota \in \{c, r\}),$$

$\Omega(z) = \mathrm{E}[XX^\top|Z = z]$, $\Theta_-(z) = \mathrm{E}[XX^\top\sigma_-^2(X, Z)|Z = z]$, $\Theta_+(z) = \mathrm{E}[XX^\top\sigma_+^2(X, Z)|Z = z]$, *and*

$$\iota = \begin{cases} c, & \text{if } z \in D_{1n}, \\ l, & \text{if } z \in D \cap \bigcup_{q=0}^{Q+1}[s_q - h_n,\ s_q), \\ r, & \text{if } z \in D \cap \bigcup_{q=0}^{Q+1}(s_q,\ s_q + h_n]. \end{cases}$$

If $\Theta_-(z) = \Theta_+(z)$, $\Phi_{lr}^{(\iota)}(z) = \Phi^{(\iota)}(z)$ *defined in Theorem 2.6.*

2.5 Simulations

In this section, we first discuss the selection procedure of the smoothing parameters h_n and u_n. Next, we examine the finite sample properties of the jump-preserving estimators $\check{a}_n(\cdot)$ defined in (2.6) and $\tilde{a}_n(\cdot)$ given in (2.12) using two simulated examples.

Among many bandwidth selection procedures for nonparametric models, we opt for the cross-validation method similarly to Zhao et al. (2016). When covariates X_i and Z_i do not contain any lagged dependent variables, we select the smoothing parameters by the leave-one-out cross-validation. The selected smoothing parameters $\hat{h}_n$ and $\hat{u}_n$ are thus determined by

$$(\hat{h}_n, \hat{u}_n) = \arg\min_{h_n, u_n} \sum_{i=1}^{n} \left[Y_i - X_i^\top \mathring{a}_{n,-i}(Z_i) \right]^2,$$

where $\mathring{a}_{n,-i}(Z_i)$ represents a jump-preserving estimate $\check{a}_n(\cdot)$ or $\tilde{a}_n(\cdot)$ based on all data except for the ith observation (Y_i, X_i, Z_i). If covariates X_i and Z_i do contain some lagged dependent variables with the lags up to order m, we suggest to apply the m-block-out cross-validation technique:

$$(\hat{h}_n, \hat{u}_n) = \arg\min_{h_n, u_n} \sum_{i=1}^{n} \left[Y_i - X_i^\top \mathring{a}_{n,-m_i}(Z_i) \right]^2,$$

where $\mathring{a}_{n,-m_i}(Z_i)$ is computed without using observations $\{Y_{i+j}, X_{i+j}, Z_{i+j}\}_{j=-m}^{m}$ (see Patton et al., 2009, for the data-dependent block-size selection).

To observe the estimation precision both in neighborhoods of change points and overall, we evaluate the performance of the proposed estimators via the global mean absolute deviation of errors (MADE) and local mean absolute deviation of errors ($\text{MADE}_{\text{local}}$):

$$\text{MADE} = \frac{1}{n_{\text{grid}}} \sum_{j=1}^{n_{\text{grid}}} \|\mathring{a}_n(z_j) - a(z_j)\|_1$$

and

$$\text{MADE}_{\text{local}} = \frac{1}{n_{\text{grid}}} \sum_{q=1}^{Q} \sum_{j=1}^{n_{\text{grid}}} \|\mathring{a}_n(z_j) - a(z_j)\|_1 \cdot \mathbf{1}\{z_j \in (s_q - 0.1, s_q + 0.1)\},$$

where $\mathring{a}_n(z_j)$ represents one of the considered estimators, $\{z_j\}_{j=1}^{n_{\text{grid}}}$ are the grid points, and $\|\cdot\|_1$ denotes the absolute value norm.

2.5.1 Experiment 1: Constant conditional variance function

First, we consider an AR(1) process:[‡]

$$X_t = a_0(Z_t) + a_1(Z_t)X_{t-1} + \sigma(Z_t)\varepsilon_t, \quad t = 1, \ldots, n, \tag{2.13}$$

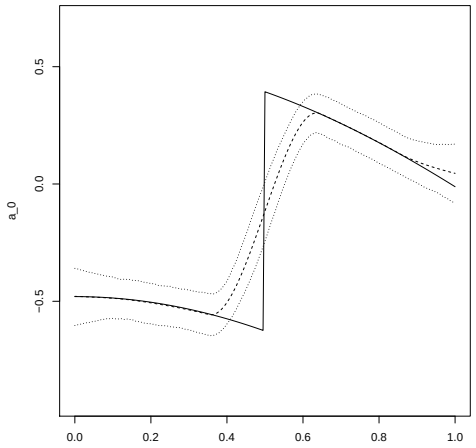
where the variable Z_t is drawn independently from the uniform distribution, $Z_t \sim U(0,1)$, the errors are independent standard normal, $\varepsilon_t \sim N(0,1)$, and the coefficient functions

$$a_0(Z_t) = 1.2\cos(Z_t) - 1.68 \cdot \mathbf{1}\{Z_t < 0.5\} - 0.66 \cdot \mathbf{1}\{Z_t \geq 0.5\} \quad \text{and}$$
$$a_1(Z_t) = \cos(Z_t) - \mathbf{1}\{Z_t < 0.5\} - 0.25 \cdot \mathbf{1}\{Z_t \geq 0.5\}.$$

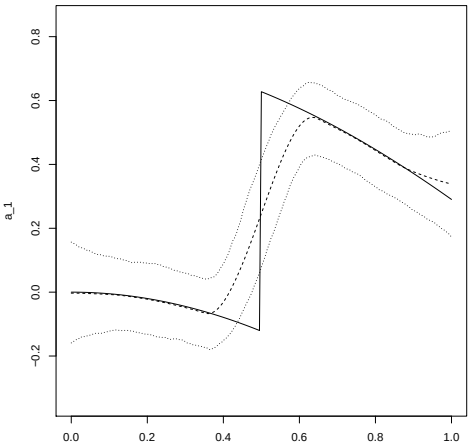
In this first simulation experiment, the variance function is constant: $\sigma^2(Z_t) = (0.6)^2$. The process (2.13) is evaluated at two sample sizes $n = 300$ and $n = 600$, and for each sample size, 1000 samples are simulated. We estimate the coefficient functions using local linear fitting on an equispaced grid of points $\{z_j\}_{j=1}^{n_{\text{grid}}}$ with $z_1 = 0$, $z_{n_{\text{grid}}} = 1$, and $n_{\text{grid}} = 200$. All nonparametric estimators employ the Epanechnikov kernel: $K^{(c)}(v) = 0.75(1 - v^2)\mathbf{1}\{|v| \leq 1\}$.

First, the bandwidth h_n is set to $0.54n^{-1/5}$ for all three local estimators, and u_n is selected by cross-validation. Figure 2.1 provides a graphical presentation of the performance of the two jump-preserving local linear estimators $\check{a}_n(z)$ (selection using WRMSE) and $\tilde{a}_n(z)$ (selection using the Wald statistics) and the conventional local linear estimator $\hat{a}_n^{(c)}(z)$ for $n = 600$. Both jump-preserving estimators track the true coefficient functions closely, while the conventional local linear estimator is inconsistent around the discontinuity $z = 0.5$ as the confidence intervals of $\hat{a}_n^{(c)}(z)$ do not contain the discontinuity. In addition, $\check{a}_n(z)$ compared to $\tilde{a}_n(z)$ has a wider confidence interval near the boundaries. The procedure of selecting the left-sided, right-sided, or conventional local estimators proposed for $\tilde{a}_n(z)$

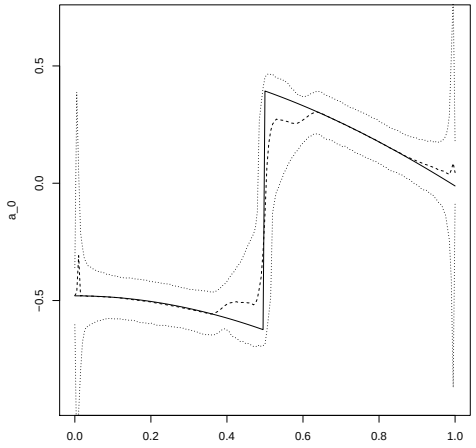
[‡]We have also studied the same AR(1) process (2.13) with coefficients that are functions of time t/n. Although using a linear time trend t/n as Z_t might violate Assumption 2.A1, the simulation results are similar to the case with a uniformly distributed Z_t.



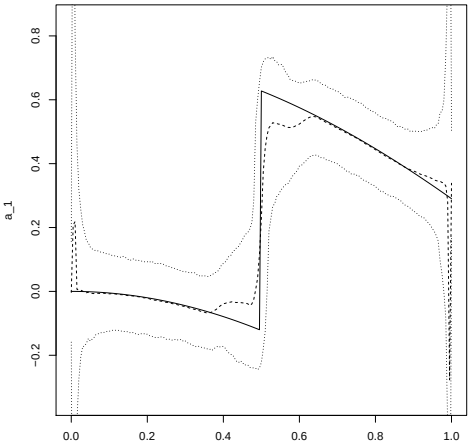
(A) Conventional $\hat{a}_{n,0}^{(c)}$



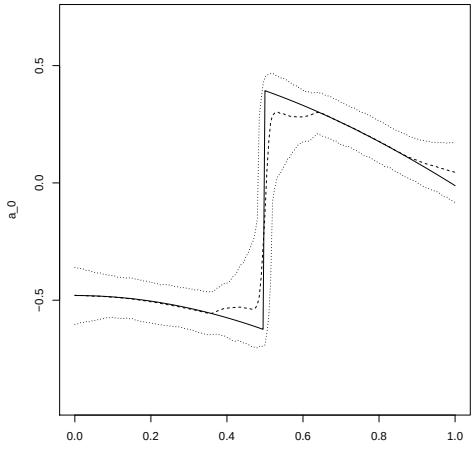
(B) Conventional $\hat{a}_{n,1}^{(c)}$



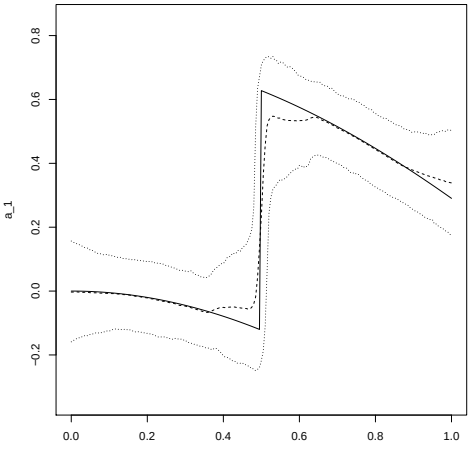
(C) Jump-preserving $\check{a}_{n,0}$ (WRMSE)



(D) Jump-preserving $\check{a}_{n,1}$ (WRMSE)

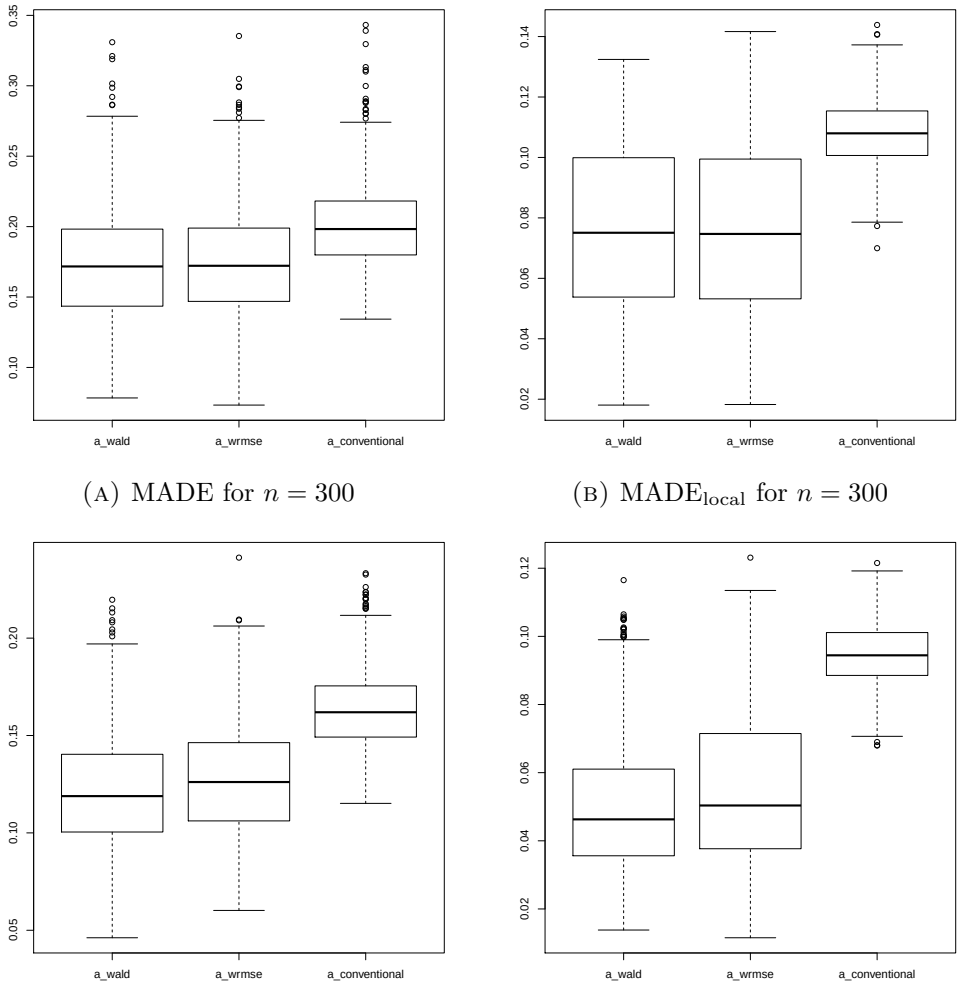


(E) Jump-preserving $\tilde{a}_{n,0}$ (Wald)



(F) Jump-preserving $\tilde{a}_{n,1}$ (Wald)

FIGURE 2.1: Homoscedastic model with the fixed bandwidth and $n = 600$: the solid lines represent the true coefficient functions, the dashed lines are the average varying coefficient estimates, and the dotted lines are the 95% confidence bands.



(A) MADE for $n = 300$

(B) $\text{MADE}_{\text{local}}$ for $n = 300$

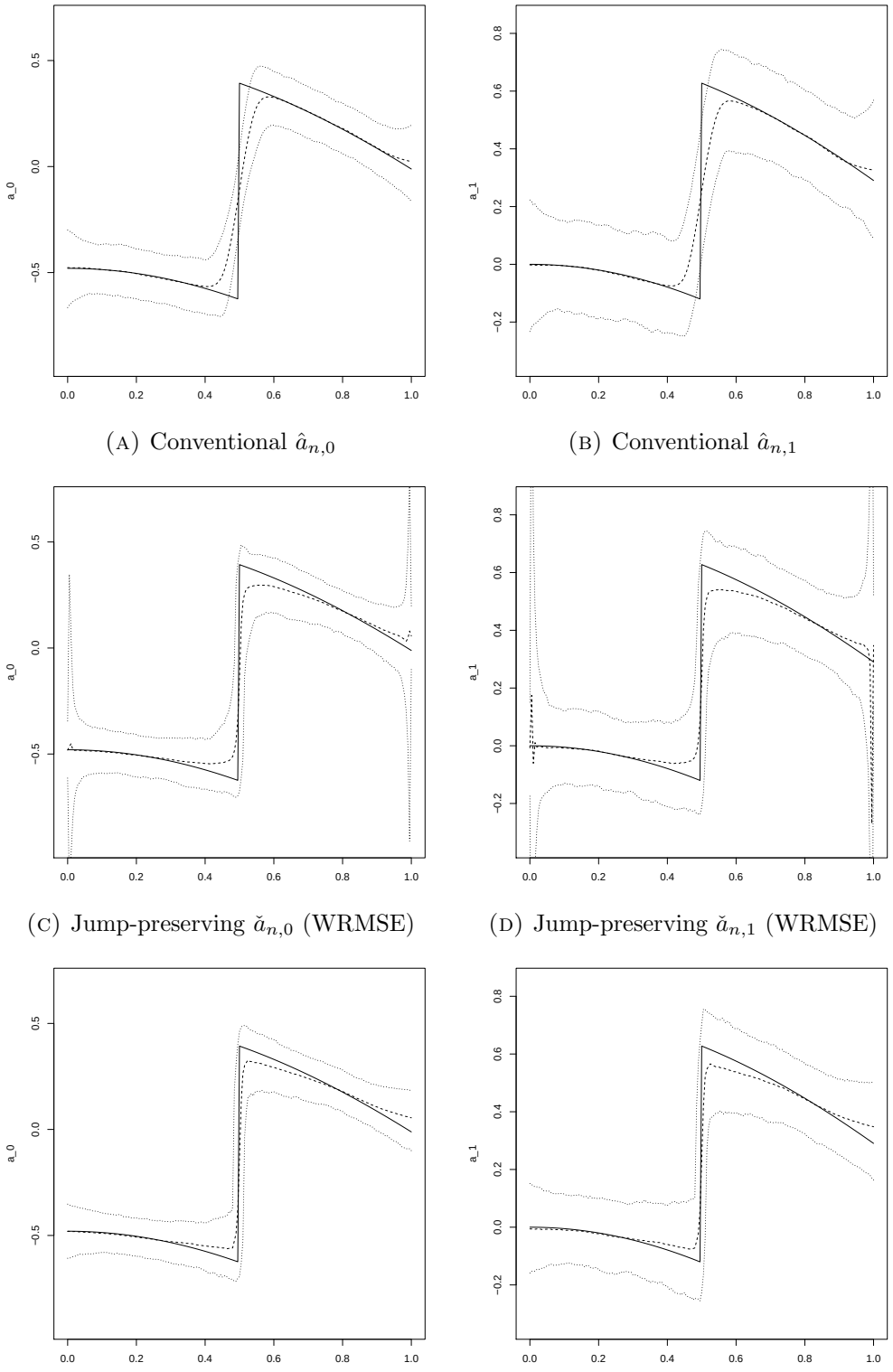
(C) MADE for $n = 600$

(D) $\text{MADE}_{\text{local}}$ for $n = 600$

FIGURE 2.2: Homoscedastic model with the fixed bandwidth: global and local mean absolute deviations of the estimates. Each plot contains boxplots for (from left to right) the jump-preserving estimator based on the Wald statistics, the jump-preserving estimator based on WRMSE, and the conventional estimator.

in Section 2.4 still chooses $\hat{a}_n^{(c)}(z)$ around the boundary points and is thus less affected by the boundaries than $\check{a}_n(z)$.

Due to strong boundary effects in $\check{a}_n(z)$, the 1000 global and local MADE values for each sample size are computed for $z \in [0.05, 0.95]$. The boxplots are shown in Figure 2.2. The conventional local linear estimator has higher global and local MADE values compared to the jump-preserving estimators $\check{a}_n(z)$ and $\tilde{a}_n(z)$, where there is no significant difference in the MADEs of $\check{a}_n(z)$ and $\tilde{a}_n(z)$. Both jump-preserving estimators thus perform well in the case of the process with homoscedastic error. When the sample size becomes larger, all global and local MADEs decrease proportionally for all estimators.

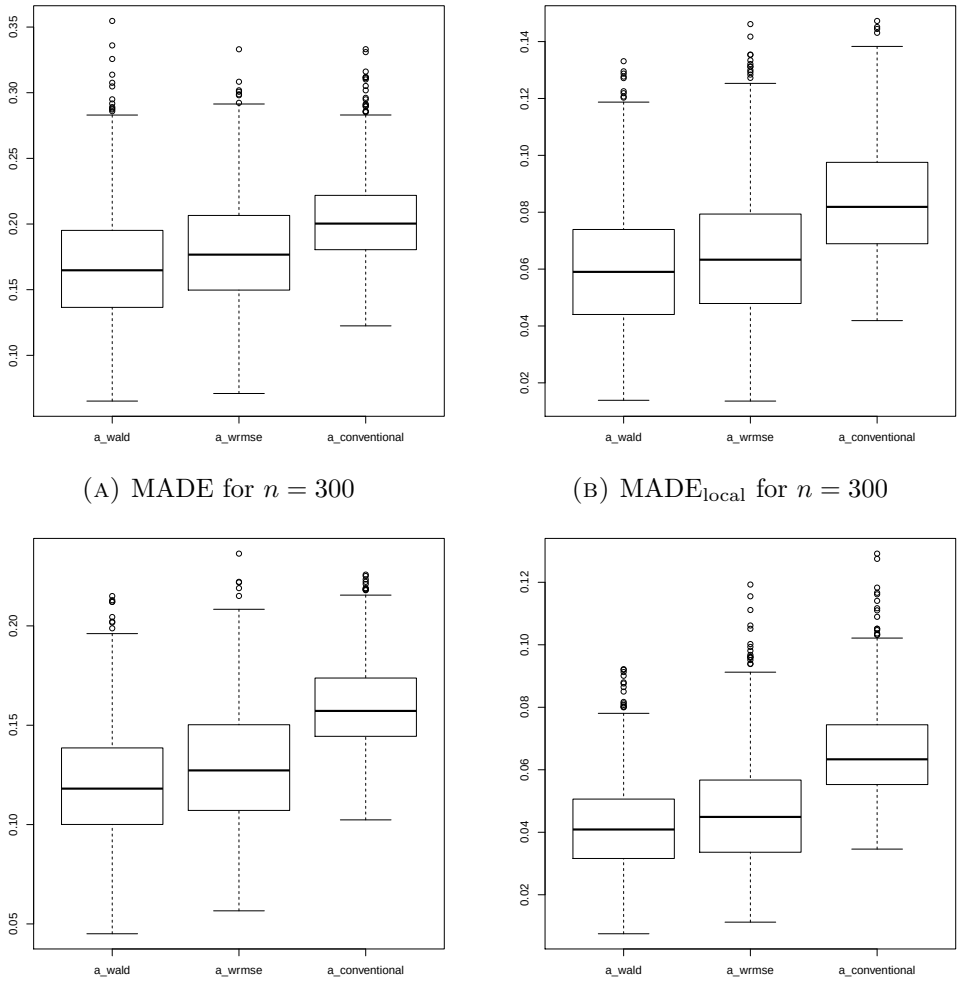


(A) Conventional $\hat{a}_{n,0}$ (B) Conventional $\hat{a}_{n,1}$

(C) Jump-preserving $\check{a}_{n,0}$ (WRMSE) (D) Jump-preserving $\check{a}_{n,1}$ (WRMSE)

(E) Jump-preserving $\tilde{a}_{n,0}$ (Wald) (F) Jump-preserving $\tilde{a}_{n,1}$ (Wald)

FIGURE 2.3: Homoscedastic model with the cross-validated bandwidth and $n = 600$: the solid lines represent the true coefficient functions, the dashed lines are the average varying coefficient estimates, and the dotted lines are the 95% confidence bands.



(A) MADE for $n = 300$

(B) $\text{MADE}_{\text{local}}$ for $n = 300$

(C) MADE for $n = 600$

(D) $\text{MADE}_{\text{local}}$ for $n = 600$

FIGURE 2.4: Homoscedastic model with the cross-validated bandwidth: global and local mean absolute deviations of the estimates. Each plot contains boxplots for (from left to right) the jump-preserving estimator based on the Wald statistics, the jump-preserving estimator based on WRMSE, and the conventional estimator.

Next, we repeat the experiment, but cross-validate both h_n and u_n for each replication; the results are shown in Figures 2.3 and 2.4. The interpretation of the results is similar as above. The main difference is that the MADE of the conventional local linear estimator is smaller than before since the bandwidth selected for $\hat{a}_n^{(c)}(z)$ is freely chosen and thus becomes smaller in an attempt to capture the discontinuity as good as possible, while decreasing the precision in the continuity region. Nevertheless, the discontinuity is not included in its confidence interval and its performance is still worse than that of the proposed jump-preserving estimators.

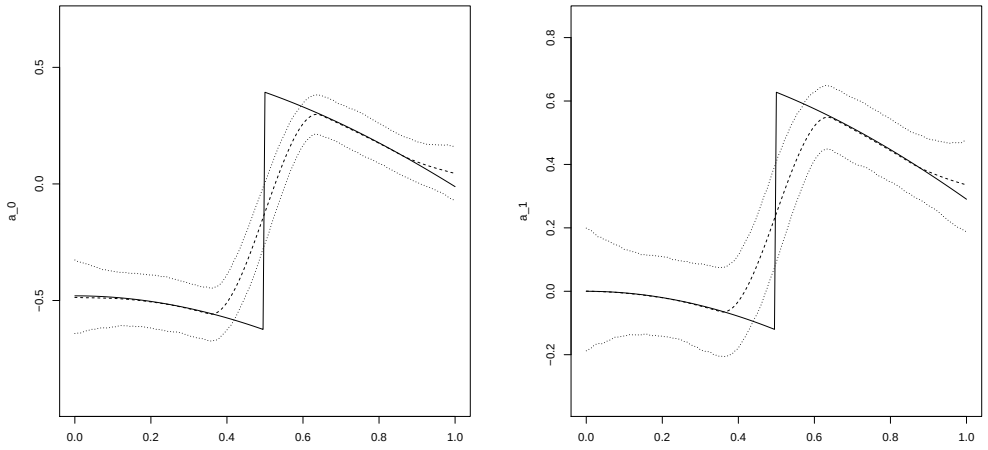
2.5.2 Experiment 2: discontinuous conditional variance function

Now we consider the same time-varying AR(1) process as in (2.13), but with a discontinuous conditional variance function:

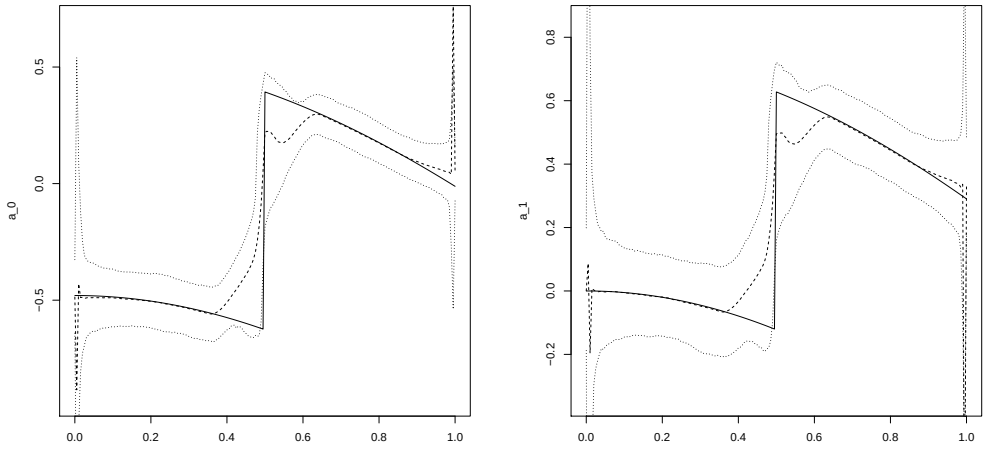
$$\sigma^2(Z_t) = \left(0.8 \cdot \mathbf{1}\{Z_t < 0.5\} + 0.6 \cdot \mathbf{1}\{Z_t \geq 0.5\}\right)^2 . \tag{2.14}$$

The evaluation is performed in the same way as in the previous section. Let us note that this experiment exhibits only one discontinuity in the variance function, which coincides with the discontinuity in the coefficient functions. Qualitatively similar results are also obtained if the variance discontinuity occurs at the points of continuity of the coefficient functions, see Sections 2.10, where we additionally compare the variance of the estimates obtained from the simulation and the asymptotic distribution, respectively.

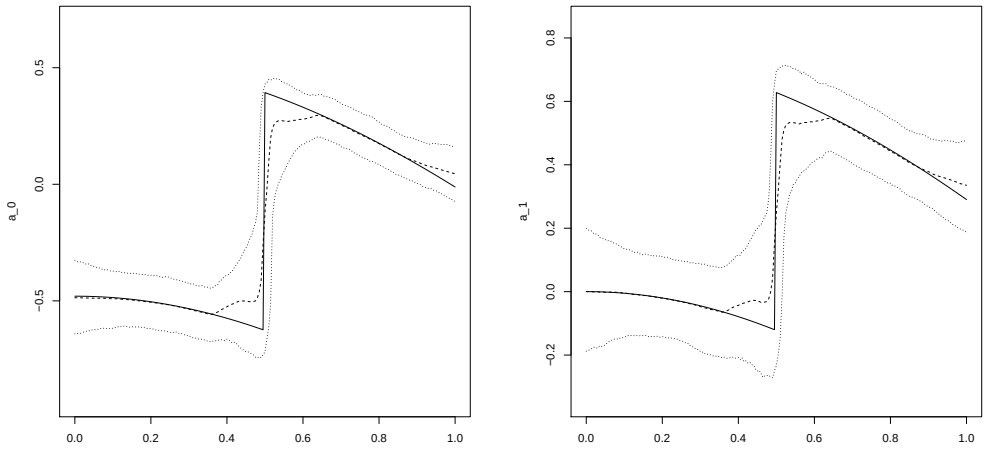
Figure 2.5 provides a graphical presentation of the performance of the convetional estimator $\hat{a}_n^{(c)}(z)$, jump-preserving estimator $\check{a}_n(z)$ based on WRMSE, and jump-preserving estimator $\tilde{a}_n(z)$ based on the Wald statistics with a fixed bandwidth $h_n = 0.54n^{-1/5}$, whereas the results using the cross-validated bandwidth h_u and u_n are presented in Figure 2.7. In this case, only the proposed jump-preserving estimators $\tilde{a}_n(z)$ based on the Wald statistics preserve the discontinuity, whereas $\hat{a}_n^{(c)}(z)$ and $\check{a}_n(z)$ are both inconsistent as their confidence intervals do not contain the discontinuity for z's near the jump point; note that this is true even for the jump-preserving method based on WRMSE. The corresponding boxplots with MADE are shown in Figures 2.6 and 2.8. The proposed estimator $\tilde{a}_n(z)$ based on the Wald statistics has the lowest global and local MADE values compared to the other jump-preserving estimator $\hat{a}_n(z)$ and to the conventional local linear estimator $\check{a}_n(z)$. The differences become a bit smaller when we cross-validate both the bandwidths h_n and u_n (see Figure 2.8). In both cases, the jump-preserving estimator $\tilde{a}_n(\cdot)$ in (2.12) outperforms the existing method $\hat{a}_n(\cdot)$ in (2.6) in the presence of the discontinuity of conditional variance function, in particular in terms of MADE.



(A) Conventional $\hat{a}_{n,0}$ (B) Conventional $\hat{a}_{n,1}$

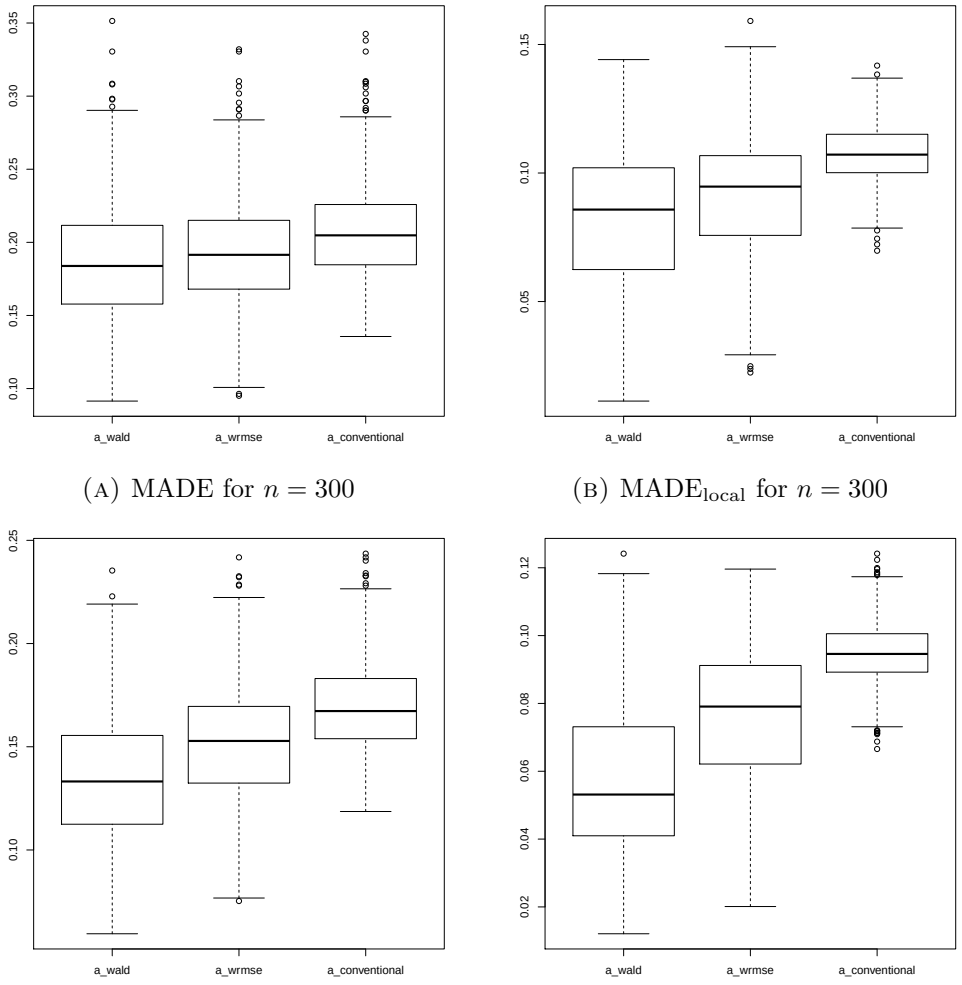


(C) Jump-preserving $\check{a}_{n,0}$ (WRMSE) (D) Jump-preserving $\check{a}_{n,1}$ (WRMSE)



(E) Jump-preserving $\tilde{a}_{n,0}$ (Wald) (F) Jump-preserving $\tilde{a}_{n,1}$ (Wald)

FIGURE 2.5: Heteroskedastic model with the fixed bandwidth: the solid lines represent the true coefficient functions, the dashed lines are the average varying coefficient estimates, and the dotted lines are the 95% confidence bands.



(A) MADE for $n = 300$

(B) $\text{MADE}_{\text{local}}$ for $n = 300$

(C) MADE for $n = 600$

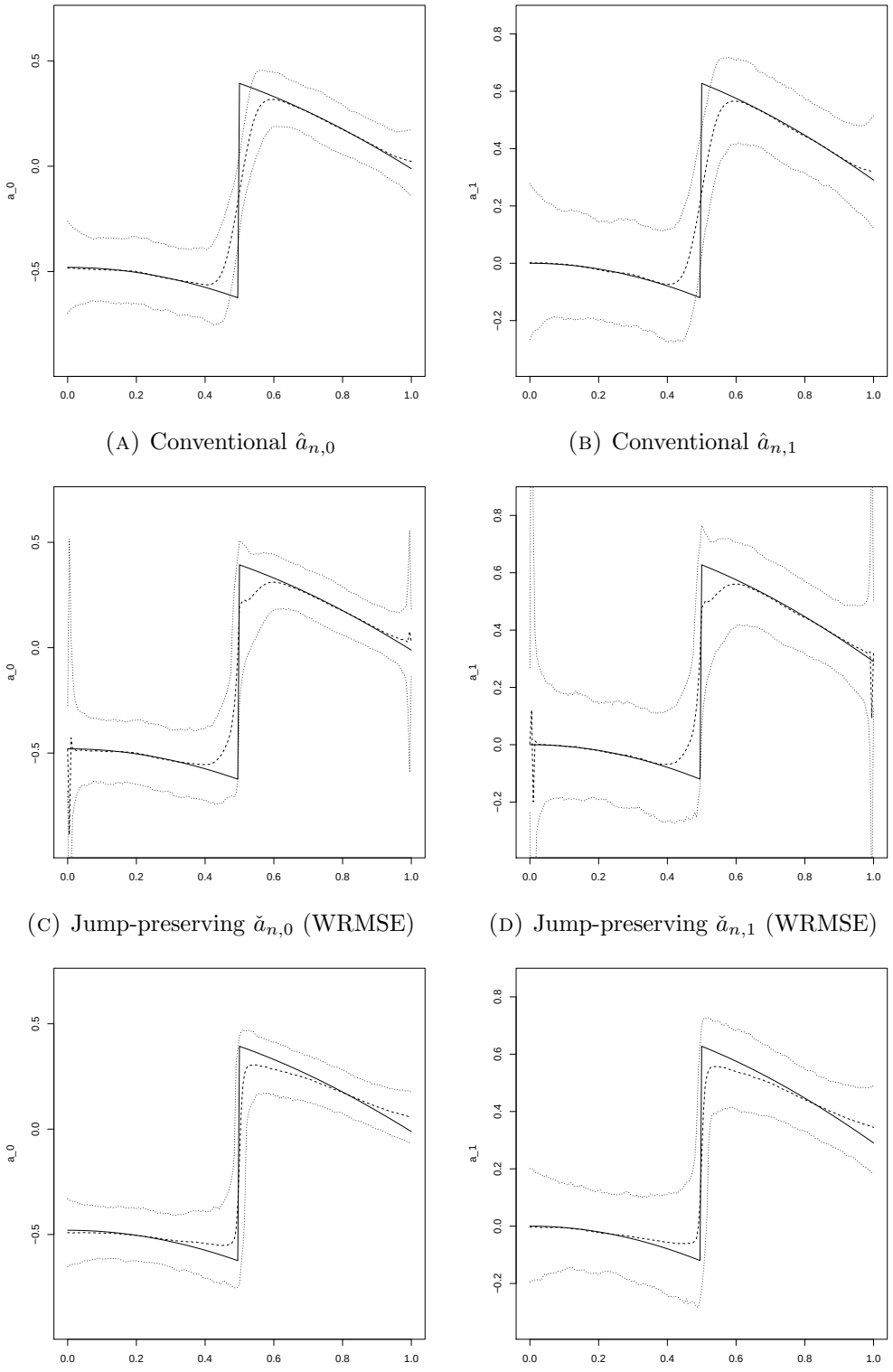
(D) $\text{MADE}_{\text{local}}$ for $n = 600$

FIGURE 2.6: Heteroskedastic model with the fixed bandwidth: global and local mean absolute deviations of the estimates. Each plot contains boxplots for (from left to right) the jump-preserving estimator based on the Wald statistics, the jump-preserving estimator based on WRMSE, and the conventional estimator.

2.6 Application

Nonlinearity in the US interest rate function has been documented in several studies, including Boldea and Hall (2013), who apply the smooth transition autoregressive model to the monthly US interest rates. Generalizing their model to the varying-coefficient setting leads to the interest rate model written in the following way:

$$r_t = \beta_0(z_t) + \beta_1(z_t)r_{t-1} + \beta_2(z_t)r_{t-2} + \beta_3(z_t)\pi_{t-1} + \beta_4(z_t)y_{t-1} + \varepsilon_t,$$

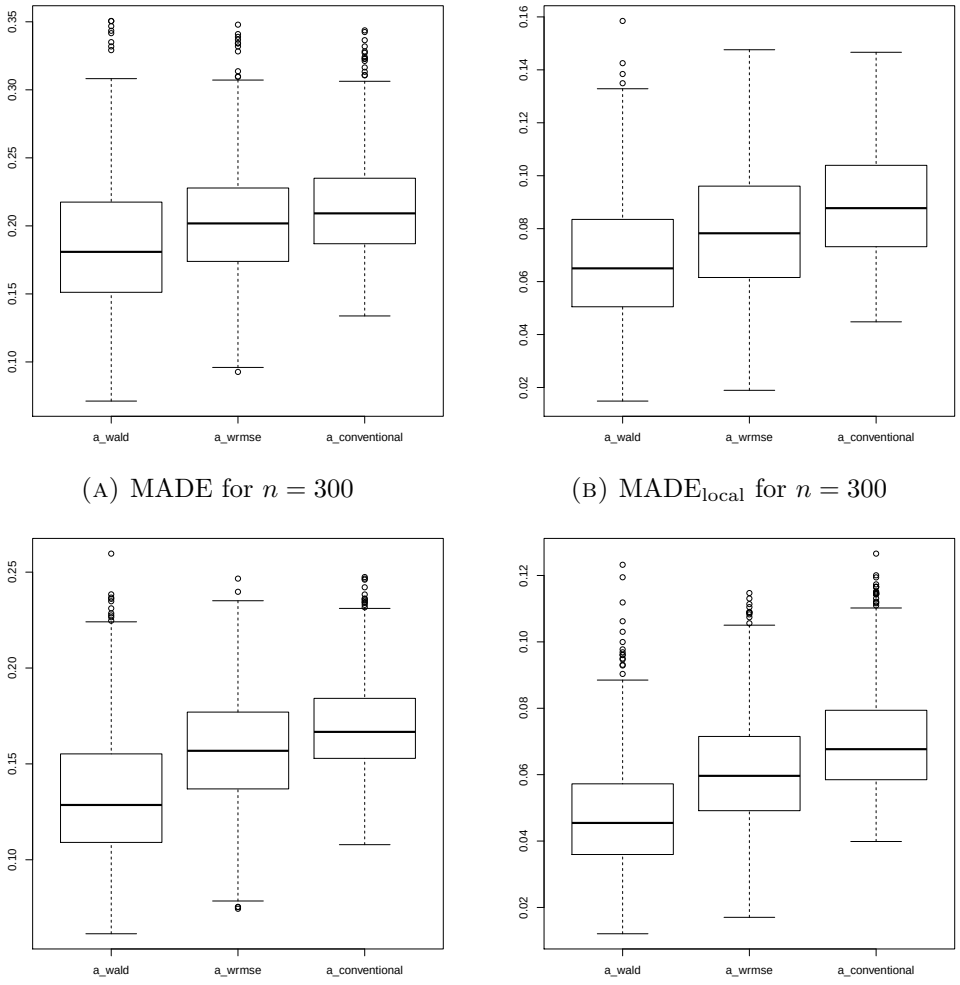


(A) Conventional $\hat{a}_{n,0}$ (B) Conventional $\hat{a}_{n,1}$

(C) Jump-preserving $\check{a}_{n,0}$ (WRMSE) (D) Jump-preserving $\check{a}_{n,1}$ (WRMSE)

(E) Jump-preserving $\tilde{a}_{n,0}$ (Wald) (F) Jump-preserving $\tilde{a}_{n,1}$ (Wald)

FIGURE 2.7: Heteroskedastic model with the cross-validated bandwidth: the solid lines represent the true coefficient functions, the dashed lines are the average varying coefficient estimates, and the dotted lines are the 95% confidence bands.



(A) MADE for $n = 300$

(B) $\text{MADE}_{\text{local}}$ for $n = 300$

(C) MADE for $n = 600$

(D) $\text{MADE}_{\text{local}}$ for $n = 600$

FIGURE 2.8: Heteroskedastic model with the cross-validated bandwidth: global and local mean absolute deviations of the estimates. Each plot contains boxplots for (from left to right) the jump-preserving estimator based on the Wald statistics, the jump-preserving estimator based on WRMSE, and the conventional estimator.

where r_t represents the monthly interest rate, π_t and y_t are the inflation and output gaps, and the index variable $z_t = r_{t-1} - r_{t-4}$. Due to the evidence of structural break in 1990 by Boldea and Hall (2013), we will estimate this model only using the data from 1991 till 2010 and compare the results to the smooth transition estimates.

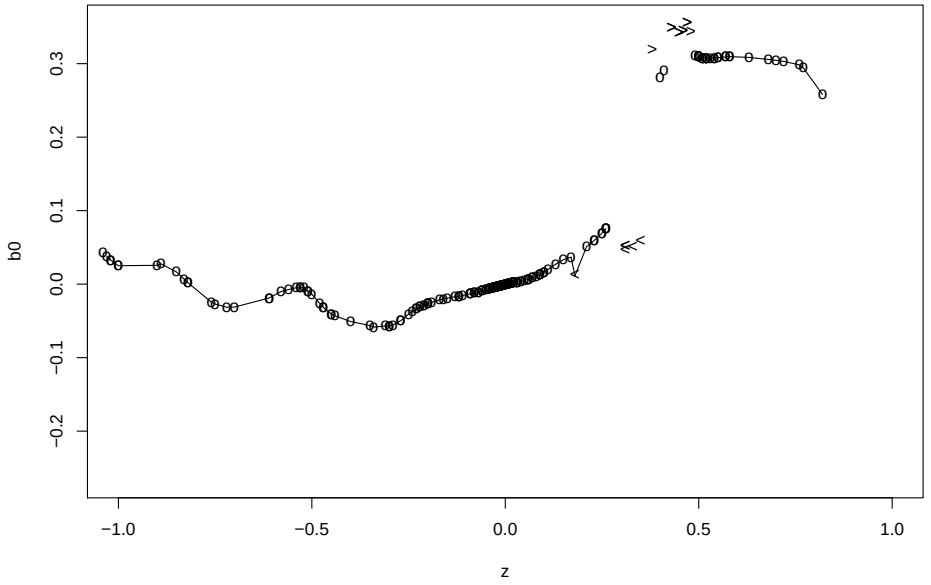
The estimation is performed by the method proposed in Section 2.4 using the Epanechnikov kernel and bandwidth equal to 0.33. The bandwidth was limited due to the fact that the support of the index variable is approximately $(-1, 1)$ and that the transition prior to 1990 seems to occur around z equal to 0.5 (Boldea and Hall, 2013); there are no

prior significant results about the point of transition in more recent data. The auxiliary parameter u_n in (2.12) was then obtained by the leave-one-out cross-validation.

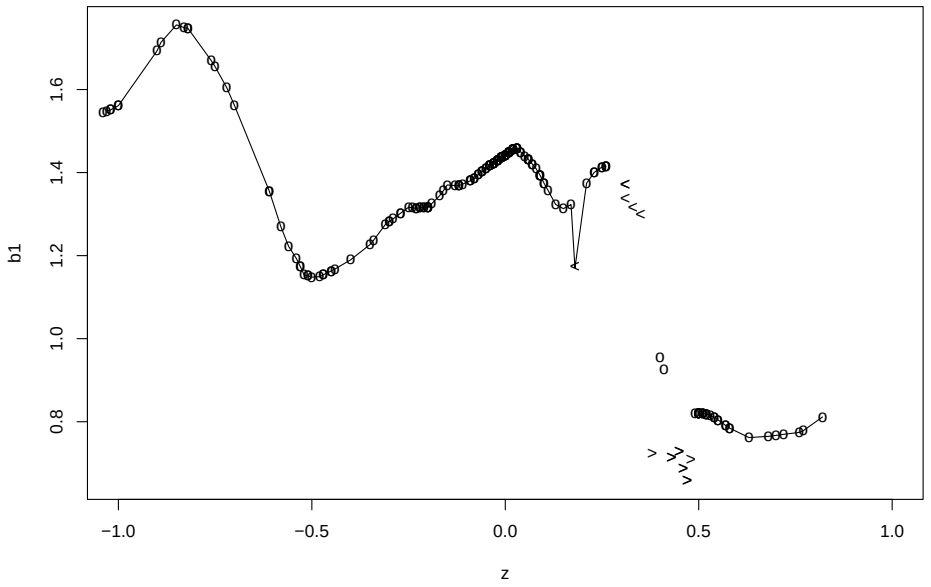
The estimates for the parameters $\beta_0(z_t), \ldots, \beta_3(z_t)$ are presented in Figures 2.9 and 2.10 (coefficient $\beta_4(z_t)$ is not displayed due to its insignificance in the original and present study). The magnitudes of the coefficients in the left parts of the graphs ($z < 0$) and in the right parts of the graphs ($z > 0.5$) are similar to the regime estimates obtained in Boldea and Hall (2013). Despite taking some fluctuations of the estimates due to a relatively small bandwidth into account, there is not a clear support for monotonicity of the coefficient functions imposed by the smooth transition models. Additionally, the proposed estimation method detects a jump around 0.4, which is close to Boldea and Hall (2013)'s findings regarding the location of the regime change. Altogether, the varying coefficient model provides more flexibility in modelling the interest rates and indicates the dynamics of the US interest rates changes substantially if their quarterly changes are 0.5 or higher.

2.7 Conclusions

In this paper, we propose estimators for varying-coefficient models with discontinuous coefficient functions. First, we adapt the local linear estimators of Gijbels et al. (2007) and Zhao et al. (2016), which select among the left-sided, right-sided, and conventional local linear estimators by comparing their weighted residual mean squared errors, to the time series setting. This approach works well when there are no discontinuities in the conditional error variance. To cope with the discontinuity problem in the conditional error variance, we propose a different "correctness" measure of the three local linear fits based on the Wald statistics. In all cases, the asymptotic properties including the uniform consistency and asymptotic normality are derived for both proposed estimators and their performance is tested with simulated examples.

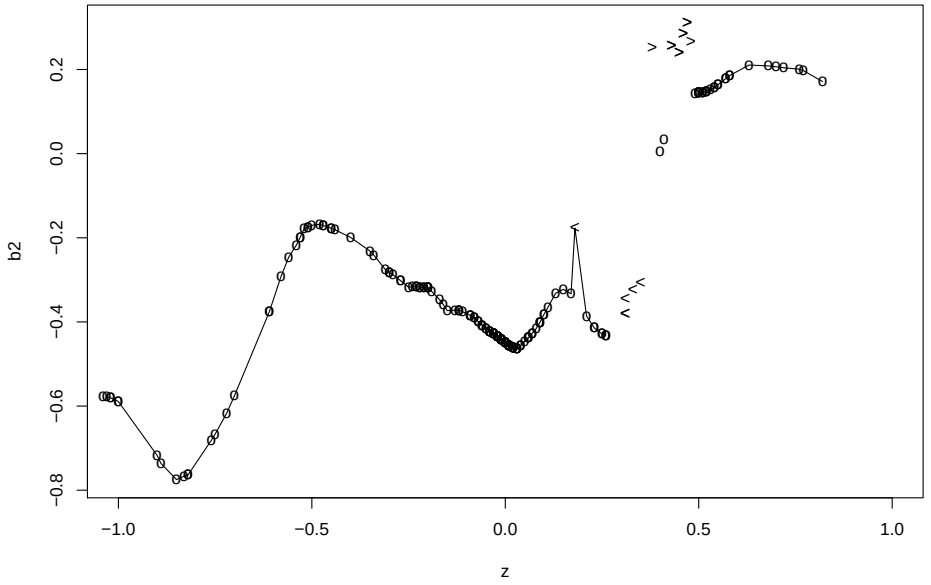


(A) Coefficient $\beta_0(z)$

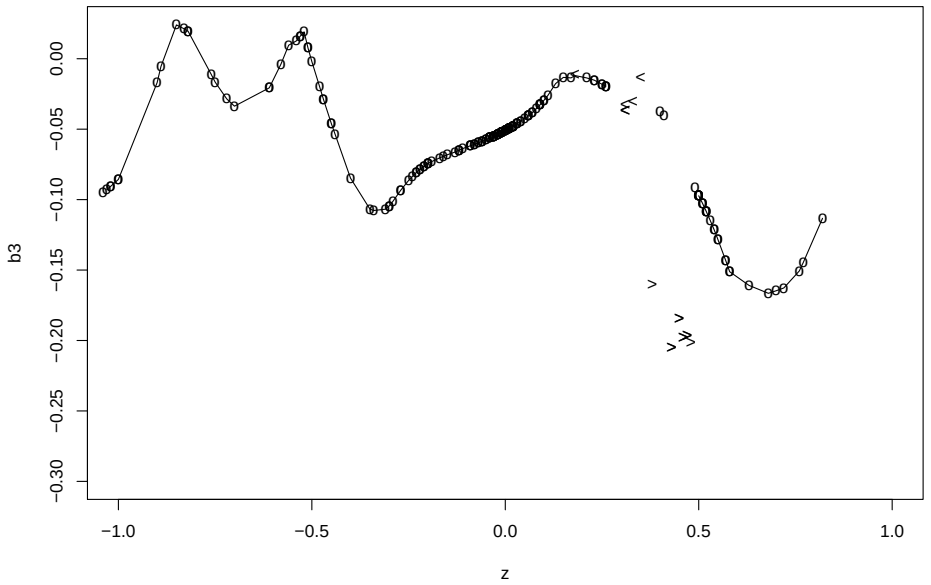


(B) Coefficient $\beta_1(z)$

FIGURE 2.9: The coefficient estimates obtained by the heteroscedasticity-robust jump-preserving estimator for the US interest rate reaction function. The use of the two-, left-, and right-sided kernel estimator is indicated by symbols 'o', '<', '>', respectively.



(A) Coefficient $\beta_2(z)$



(B) Coefficient $\beta_3(z)$

FIGURE 2.10: The coefficient estimates obtained by the heteroscedasticity-robust jump-preserving estimator for the US interest rate reaction function. The use of the two-, left-, and right-sided kernel estimator is indicated by symbols 'o', '<', '>', respectively.

2.8 Appendix: Proofs of the main results

In this section, we prove the theorems presented in Sections 2.3 and 2.4. Auxiliary lemmas are collected in Section 2.9. Throughout Sections 2.8 and 2.9, we let C be a generic positive constant, which may take different values at different places, and write $M \succ 0$ if matrix M is positive definite. All limiting expressions including $\mathrm{o_p}(\cdot)$ and $\mathrm{O_p}(\cdot)$ are taken for $n \to \infty$, unless stated otherwise. The dependence on z of the variables introduced in Sections 2.8 and 2.9 is kept implicit in order to shorten the length of proofs.

First, we introduce some notation. Denote

$$S_n^{(\iota)} = \begin{pmatrix} S_{n,0}^{(\iota)} & S_{n,1}^{(\iota)} \\ S_{n,1}^{(\iota)} & S_{n,2}^{(\iota)} \end{pmatrix}, \quad T_n^{(\iota)} = \begin{pmatrix} T_{n,0}^{(\iota)} \\ T_{n,1}^{(\iota)} \end{pmatrix}, \quad \text{and} \quad F_n^{(\iota)} = \begin{pmatrix} F_{n,0}^{(\iota)} \\ F_{n,1}^{(\iota)} \end{pmatrix},$$

where

$$S_{n,j}^{(\iota)} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top \left(\frac{Z_i - z}{h_n}\right)^j K_h^{(\iota)}(Z_i - z), \quad j = 0, 1, 2, 3, \tag{2.15}$$

$$T_{n,j}^{(\iota)} = \frac{1}{n}\sum_{i=1}^{n} X_i \left(\frac{Z_i - z}{h_n}\right)^j K_h^{(\iota)}(Z_i - z) Y_i, \quad j = 0, 1, \quad \text{and} \tag{2.16}$$

$$F_{n,j}^{(\iota)} = \frac{1}{n}\sum_{i=1}^{n} X_i \left(\frac{Z_i - z}{h_n}\right)^j K_h^{(\iota)}(Z_i - z) \varepsilon_i, \quad j = 0, 1. \tag{2.17}$$

Using the above notation, the local linear estimators of $a(\cdot)$ and $a'(\cdot)$ in (2.4) can be written as

$$\begin{aligned} \hat{\beta}_n^{(\iota)} =& \begin{pmatrix} \hat{a}_n^{(\iota)}(z) \\ \hat{b}_n^{(\iota)}(z) \end{pmatrix} \\ =& H_n^{-1} \left[\sum_{i=1}^{n} H_n^{-1} \begin{pmatrix} X_i \\ X_i(Z_i - z) \end{pmatrix} \begin{pmatrix} X_i \\ X_i(Z_i - z) \end{pmatrix}^\top H_n^{-1} K_h^{(\iota)}(Z_i - z)\right]^{-1} \\ & \sum_{i=1}^{n} H_n^{-1} \begin{pmatrix} X_i \\ X_i(Z_i - z) \end{pmatrix} Y_i K_h^{(\iota)}(Z_i - z) \\ =& H_n^{-1} S_n^{(\iota)-1} T_n^{(\iota)}, \end{aligned} \tag{2.18}$$

where H_n is a $2p \times 2p$ diagonal matrix with its first p diagonal elements equal to 1's and its last p elements equal to h_n's.

Since the coefficient functions $a(z)$ are twice continuously differentiable except for the discontinuities $\{s_q\}_{q=0}^{Q+1}$ (Assumption 2.A5), it follows from the Taylor expansion for $Z_i \in D_{zn}^{(\iota)}, z \in D_{1n}$, that

$$a(Z_i) = a(z) + h_n \left(\frac{Z_i - z}{h_n}\right) a'(z) + \frac{h_n^2}{2} \left(\frac{Z_i - z}{h_n}\right)^2 a''(z) + \mathrm{o}(Z_i - z)^2 \tag{2.19}$$

uniformly in $z \in D_{1n}^{(\iota)}$, which implies

$$\begin{aligned} T_{n,0}^{(\iota)} - F_{n,0}^{(\iota)} &= \frac{1}{n} \sum_{i=1}^{n} K_h^{(\iota)}(Z_i - z) X_i X_i^\top a(Z_i) \\ &= S_{n,0}^{(\iota)} a(z) + h_n S_{n,1}^{(\iota)} a'(z) + \frac{h_n^2}{2} S_{n,2}^{(\iota)} a''(z) + S_{n,0}^{(\iota)} \cdot \mathrm{o_p}(h_n^2) \end{aligned}$$

and

$$T_{n,1}^{(\iota)} - F_{n,1}^{(\iota)} = S_{n,1}^{(\iota)} a(z) + h_n S_{n,2}^{(\iota)} a'(z) + \frac{h_n^2}{2} S_{n,3}^{(\iota)} a''(z) + S_{n,1}^{(\iota)} \cdot \mathrm{o_p}(h_n^2).$$

Consequently for $\beta = [a^\top(z), a'^\top(z)]^\top$, it holds that

$$T_n^{(\iota)} - F_n^{(\iota)} = S_n^{(\iota)} H_n \beta + \frac{h_n^2}{2} \begin{pmatrix} S_{n,2}^{(\iota)} \\ S_{n,3}^{(\iota)} \end{pmatrix} a''(z) + \begin{pmatrix} S_{n,0}^{(\iota)} \\ S_{n,1}^{(\iota)} \end{pmatrix} \cdot \mathrm{o_p}(h_n^2). \tag{2.20}$$

Using (2.18), (2.20), and Lemma 2.18(ii), we finally obtain

$$\begin{aligned} H_n(\hat{\beta}_n^{(\iota)} - \beta) =& S_n^{(\iota)-1} T_n^{(\iota)} - H_n \beta \\ =& S_n^{(\iota)-1} F_n^{(\iota)} + \frac{h_n^2}{2} S_n^{(\iota)-1} \begin{pmatrix} S_{n,2}^{(\iota)} \\ S_{n,3}^{(\iota)} \end{pmatrix} a''(z) + \mathrm{o_p}(h_n^2) \end{aligned} \tag{2.21}$$

uniformly in $z \in D_{1n}^{(\iota)}$.

Proof of Theorem 2.1.

According to Lemma 2.13, the terms $S_{n,j}^{(\iota)}$, $S_n^{(\iota)-1}$, and $F_{n,j}^{(\iota)}$ uniformly converge on $D_{1n}^{(\iota)}$ to their corresponding expected values at rates $(nh_n/\ln n)^{-1/2} + h_n$ and $(nh_n/\ln n)^{-1/2}$, respectively. It follows from (2.21) and Assumptions 2.A2, 2.A3(ii), and 2.A4 that

$$
\begin{aligned}
&\sup_{z\in D_{1n}^{(\iota)}} \left\| H_n(\hat{\beta}_n^{(\iota)} - \beta) \right\| \\
&\leq \sup_{z\in D_{1n}^{(\iota)}} \left\| S_n^{(\iota)-1} \right\| \left\{ \sup_{z\in D_{1n}^{(\iota)}} \left\| F_n^{(\iota)} \right\| + \sup_{z\in D_{1n}^{(\iota)}} \left\| \frac{h_n^2}{2} \begin{pmatrix} S_{n,2}^{(\iota)} \\ S_{n,3}^{(\iota)} \end{pmatrix} \right\| \right\} \max_{z\in D_{1n}^{(\iota)}} \|a''(z)\| + \mathrm{o_p}(h_n^2) \\
&\leq C_1 \cdot \frac{\sup_{z\in D_{1n}^{(\iota)}} \|\Omega^{-1}(z)\|}{\inf_{z\in D} f_Z(z)} \left\{ 1 + \mathrm{O_p}\left(\sqrt{\frac{\ln n}{nh_n}} + h_n \right) \right\} \\
&\quad \cdot \left[\mathrm{O_p}\left(\sqrt{\frac{\ln n}{nh_n}} \right) + C_2 h_n^2 \left\{ \sup_{z\in D_{1n}^{(\iota)}} \|f_Z(z)\Omega(z)\| + \mathrm{O_p}\left(\sqrt{\frac{\ln n}{nh_n}} + h_n \right) \right\} \right] + \mathrm{o_p}(h_n^2) \\
&\leq C_3 \cdot \left\{ 1 + \mathrm{O_p}\left(\sqrt{\frac{\ln n}{nh_n}} + h_n \right) \right\} \cdot \mathrm{O_p}\left(\sqrt{\frac{\ln n}{nh_n}} + h_n^2 + h_n^3 \right) + \mathrm{o_p}(h_n^2) \\
&= \mathrm{O_p}\left(\sqrt{\frac{\ln n}{nh_n}} \right) + \mathrm{O_p}(h_n^2), \quad \iota = c, l, r,
\end{aligned}
$$

where C_1, C_2, and C_3 represent some positive constants and $\Omega(z) = \mathrm{E}[XX^\top | Z = z]$. As a result, we have

$$
\sup_{z\in D_{1n}^{(\iota)}} \left\| \hat{a}_n^{(\iota)}(z) - a(z) \right\| = \mathrm{O_p}\left(\sqrt{\frac{\ln n}{nh_n}} \right) + \mathrm{O_p}(h_n^2), \quad \iota = c, l, r,
$$

and

$$
\sup_{z\in D_{1n}^{(\iota)}} \left\| \hat{b}_n^{(\iota)}(z) - a'(z) \right\| = \mathrm{O_p}\left(h_n^{-1} \sqrt{\frac{\ln n}{nh_n}} \right) + \mathrm{O_p}(h_n), \quad \iota = c, l, r.
$$

The claim follows by noting that $h_n^2 = \mathrm{o}(\sqrt{\ln n/(nh_n)})$ by Assumption 2.B3. □

Proof of Theorem 2.2.

By the weak convergence results for $S_{n,j}^{(\iota)}$ and $S_n^{(\iota)-1}$ in Lemmas 2.11(i) and 2.11(ii) and equation (2.21),

$$\hat{a}_n^{(\iota)}(z) - a(z) = \left[\frac{\Omega^{-1}(z)}{f_Z(z)}\left(c_0^{(\iota)}F_{n,0}^{(\iota)} + c_1^{(\iota)}F_{n,1}^{(\iota)}\right) + \frac{h_n^2}{2}\left(c_0^{(\iota)}\mu_2^{(\iota)} + c_1^{(\iota)}\mu_3^{(\iota)}\right)a''(z)\right] \cdot (1 + \mathrm{o_p}(1)) + \mathrm{o_p}(h_n^2), \tag{2.22}$$

where $c_j^{(\iota)}$ and $\mu_j^{(\iota)}$ are defined in (2.7). The stochastic term in (2.22) can be analyzed in the following way. Let

$$U_n^{(\iota)} = c_0^{(\iota)}F_{n,0}^{(\iota)} + c_1^{(\iota)}F_{n,1}^{(\iota)} = \frac{1}{n}\sum_{i=1}^{n} W_i^{(\iota)}, \tag{2.23}$$

where

$$W_i^{(\iota)} = X_i\left[c_0^{(\iota)} + c_1^{(\iota)}\left(\frac{Z_i - z}{h_n}\right)\right]K_h^{(\iota)}(Z_i - z_0)\varepsilon_i. \tag{2.24}$$

By applying the central limit theorem for strong mixing process (Fan and Yao, 2003, Theorem 2.21) under the mixing condition in Assumption 2.A1 and the moment condition in Assumption 2.A3(i), $\sqrt{nh_n}U_n^{(\iota)}$ is asymptotically normal with mean 0 (due to the law of iterated expectation) and variance (by Lemma 2.12)

$$nh_n\mathrm{var}(U_n^{(\iota)}) = f_Z(z)\Theta(z)\left[c_0^{(\iota)2}\nu_0^{(\iota)} + 2c_0^{(\iota)}c_1^{(\iota)}\nu_1^{(\iota)} + c_1^{(\iota)2}\nu_2^{(\iota)}\right] + \mathrm{o}(1),$$

where $\Theta(z) = \mathrm{E}[XX^\top\sigma^2(X,Z)|Z = z]$. As the remaining term in (2.22) is deterministic, we obtain

$$\sqrt{nh_n}\left[\hat{a}_n^{(\iota)}(z) - a(z) - \frac{h_n^2}{2}\left(c_0^{(\iota)}\mu_2^{(\iota)} + c_1^{(\iota)}\mu_3^{(\iota)}\right)a''(z)\right] = \frac{\Omega^{-1}(z)}{f_Z(z)}\sqrt{nh_n}U_n^{(\iota)} + \mathrm{o_p}(1),$$

where the leading term is asymptotically normal with mean 0 and variance $\Phi^{(\iota)}(z)$ given in Theorem 2.2. □

Proof of Theorem 2.3.

It follows from the definition of WRMSE $\Psi_n^{(\iota)}(z)$ in (2.5) that

$$\Psi_n^{(\iota)}(z) = \frac{N_n^{(\iota)}}{K_n^{(\iota)}},$$

where the denominator

$$K_n^{(\iota)} = \frac{1}{n}\sum_{i=1}^{n} K_h^{(\iota)}(Z_i - z) \tag{2.25}$$

and the numerator $N_n^{(\iota)}$, which can be decomposed into three terms, is given by

$$\begin{aligned}
N_n^{(\iota)} =& \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_{n,i}^{(\iota)2} K_h^{(\iota)}(Z_i - z) \\
=& \frac{1}{n}\sum_{i=1}^{n} \left[Y_i - X_i^\top \{\hat{a}_n^{(\iota)}(z) + \hat{b}_n^{(\iota)}(z)(Z_i - z)\}\right]^2 K_h^{(\iota)}(Z_i - z) \\
=& \frac{1}{n}\sum_{i=1}^{n} \left[\varepsilon_i + X_i^\top \{a(Z_i) - \hat{a}_n^{(\iota)}(z) - \hat{b}_n^{(\iota)}(z)(Z_i - z)\}\right]^2 K_h^{(\iota)}(Z_i - z) \\
=& \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 K_h^{(\iota)}(Z_i - z) \\
&+ \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i \left[X_i^\top \{a(Z_i) - \hat{a}_n^{(\iota)}(z) - \hat{b}_n^{(\iota)}(z)(Z_i - z)\}\right] K_h^{(\iota)}(Z_i - z) \\
&+ \frac{1}{n}\sum_{i=1}^{n} \left[X_i^\top \{a(Z_i) - \hat{a}_n^{(\iota)}(z) - \hat{b}_n^{(\iota)}(z)(Z_i - z)\}\right]^2 K_h^{(\iota)}(Z_i - z) \\
=& N_{n,1}^{(\iota)} + N_{n,2}^{(\iota)} + N_{n,3}^{(\iota)}
\end{aligned} \tag{2.26}$$

with $N_{n,1}^{(\iota)}$, $N_{n,2}^{(\iota)}$, and $N_{n,3}^{(\iota)}$ being the first, second, and third terms in (2.26), respectively. According to Lemmas 2.11(iv) and 2.11(v), $N_{n,1}^{(\iota)}/K_n^{(\iota)} = \sigma^2(z) + \mathrm{o_p}(1)$ for $z \in D_{1n}^{(\iota)}$. It remains to show $N_{n,2}^{(\iota)} = \mathrm{o_p}(1)$ and $N_{n,3}^{(\iota)} = \mathrm{o_p}(1)$. By the Taylor expansion of $a(Z_i)$ and the weak convergence results for $F_{n,j}^{(\iota)}$, $\hat{a}_n^{(\iota)}(z)$, and $\hat{b}_n^{(\iota)}(z)$ in Lemmas 2.11(iii), 2.11(vi),

and 2.11(vii), respectively, we have

$$\begin{aligned}
N_{n,2}^{(\iota)} =& \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i\left[X_i^\top\{a(z)+a'(z)(Z_i-z)+\mathrm{o}(Z_i-z)\}\right.\\
&\left.-X_i^\top\{\hat{a}_n^{(\iota)}(z)+\hat{b}_n^{(\iota)}(z)(Z_i-z)\}\right]K_h^{(\iota)}(Z_i-z)\\
=& 2\{a(z)-\hat{a}_n^{(\iota)}(z)\}^\top F_{n,0}^{(\iota)}+2h_n\{a'(z)-\hat{b}_n^{(\iota)}(z)\}^\top F_{n,1}^{(\iota)}+\mathrm{o_p}(h_n)\\
=& 2\mathrm{o_p}(1)\cdot\mathrm{o_p}(1)+2h_n\cdot\mathrm{o_p}(h_n^{-1})\cdot\mathrm{o_p}(1)+\mathrm{o_p}(h_n)\\
=& \mathrm{o_p}(1).
\end{aligned}$$

Similarly by the Taylor expansion of $a(Z_i)$, Lemmas 2.11(i), 2.11(vi), and 2.11(vii), and the boundedness condition on $f_Z(z)\Omega(z)$ in Assumption 2.A3(ii), it follows that

$$\begin{aligned}
N_{n,3}^{(\iota)} =& \frac{1}{n}\sum_{i=1}^{n}\left[X_i^\top\{a(z)+a'(z)(Z_i-z)+\mathrm{o}(Z_i-z)\}\right.\\
&\left.-X_i^\top\{\hat{a}_n^{(\iota)}(z)+\hat{b}_n^{(\iota)}(z)(Z_i-z)\}\right]^2 K_h^{(\iota)}(Z_i-z)\\
=& \{a(z)-\hat{a}_n^{(\iota)}(z)\}^\top S_{n,0}^{(\iota)}\{a(z)-\hat{a}_n^{(\iota)}(z)\}\\
&+2h_n\{a(z)-\hat{a}_n^{(\iota)}(z)\}^\top S_{n,1}^{(\iota)}\{a'(z)-\hat{b}_n^{(\iota)}(z)\}\\
&+h_n^2\{a'(z)-\hat{b}_n^{(\iota)}(z)\}^\top S_{n,2}^{(\iota)}\{a'(z)-\hat{b}_n^{(\iota)}(z)\}+\mathrm{o_p}(h_n)\\
\le& \mathrm{o_p}(1)\cdot\mathrm{O}\left\{\sup_{z\in D}\|f_Z(z)\Omega(z)\|+\mathrm{o_p}(1)\right\}\cdot\mathrm{o_p}(1)\\
&+2h_n\cdot\mathrm{o_p}(1)\cdot\mathrm{O}\left\{\sup_{z\in D}\|f_Z(z)\Omega(z)\|+\mathrm{o_p}(1)\right\}\cdot\mathrm{o_p}(h_n^{-1})\\
&+h_n^2\cdot\mathrm{o_p}(h_n^{-1})\cdot\mathrm{O}\left\{\sup_{z\in D}\|f_Z(z)\Omega(z)\|+\mathrm{o_p}(1)\right\}\cdot\mathrm{o_p}(h_n^{-1})+\mathrm{o_p}(h_n)\\
=& \mathrm{o_p}(1).
\end{aligned}$$

This completes the proof of Theorem 2.3. □

Before investigating the limiting behavior of the jump-preserving estimator, we introduce additional notation. For any $z = s_q + \tau h_n$ with $\tau \in (-1, 1)$, we denote random variables

$$\acute{S}_{n,j}^{(\iota)} = \frac{1}{n}\sum_{i:Z_i<s_q} X_iX_i^\top\left(\frac{Z_i-z}{h_n}\right)^j K_h^{(\iota)}(Z_i-z), \quad j=0,1,2, \tag{2.27}$$

$$\grave{S}_{n,j}^{(\iota)} = \frac{1}{n} \sum_{i:Z_i \geq s_q} X_i X_i^\top \left(\frac{Z_i - z}{h_n} \right)^j K_h^{(\iota)}(Z_i - z), \quad j = 0, 1, 2, \tag{2.28}$$

$$\acute{F}_{n,j}^{(\iota)} = \frac{1}{n} \sum_{i:Z_i < s_q} X_i \left(\frac{Z_i - z}{h_n} \right)^j K_h^{(\iota)}(Z_i - z)\varepsilon_i, \quad j = 0, 1, \tag{2.29}$$

and

$$\grave{F}_{n,j}^{(\iota)} = \frac{1}{n} \sum_{i:Z_i \geq s_q} X_i \left(\frac{Z_i - z}{h_n} \right)^j K_h^{(\iota)}(Z_i - z)\varepsilon_i, \quad j = 0, 1. \tag{2.30}$$

Further, let

$$\acute{\mu}_{j,\tau}^{(\iota)} = \int_{-1}^{-\tau} u^j K^{(\iota)}(u)\mathrm{d}u, \quad \grave{\mu}_{j,\tau}^{(\iota)} = \int_{-\tau}^{1} u^j K^{(\iota)}(u)\mathrm{d}u, \tag{2.31}$$

$$\Omega_-(s_q) = \lim_{z \uparrow s_q} \mathrm{E}[XX^\top | Z = z], \quad \Omega_+(s_q) = \lim_{z \downarrow s_q} \mathrm{E}[XX^\top | Z = z], \tag{2.32}$$

$$\acute{\Omega}_{-,\tau}^{(\iota)}(s_q) = \begin{pmatrix} \acute{\mu}_{0,\tau}^{(\iota)}\Omega_-(s_q) & \acute{\mu}_{1,\tau}^{(\iota)}\Omega_-(s_q) \\ \acute{\mu}_{1,\tau}^{(\iota)}\Omega_-(s_q) & \acute{\mu}_{2,\tau}^{(\iota)}\Omega_-(s_q) \end{pmatrix},$$

$$\grave{\Omega}_{+,\tau}^{(\iota)}(s_q) = \begin{pmatrix} \grave{\mu}_{0,\tau}^{(\iota)}\Omega_+(s_q) & \grave{\mu}_{1,\tau}^{(\iota)}\Omega_+(s_q) \\ \grave{\mu}_{1,\tau}^{(\iota)}\Omega_+(s_q) & \grave{\mu}_{2,\tau}^{(\iota)}\Omega_+(s_q) \end{pmatrix},$$

$$a_-(s_q) = \lim_{z \uparrow s_q} a(z), \quad \text{and} \quad a_+(s_q) = \lim_{z \downarrow s_q} a(z) = a_-(s_q) + d_q.$$

Without loss of generality, we assume that $a(\cdot)$ is right continuous, i.e., $a(s_q) = a_+(s_q)$ for $q = 0, \ldots, Q$. By the mean value theorem and boundedness of the (left) partial derivatives of $a(\cdot)$ (Assumption 2.A5), it holds for $Z_i \in [s_q - (1-\tau)h_n, s_q)$ that

$$a(Z_i) = a_-(s_q) + \mathrm{O}(Z_i - s_q). \tag{2.33}$$

Similarly, we have for $Z_i \in (s_q, s_q + (1+\tau)h_n]$,

$$a(Z_i) = a_+(s_q) + \mathrm{O}(Z_i - s_q) = a_-(s_q) + d_q + \mathrm{O}(Z_i - s_q). \tag{2.34}$$

Using equations (2.33) and (2.34) and the limiting results for $\acute{F}_{n,j}^{(\iota)}$, $\grave{F}_{n,j}^{(\iota)}$, $\acute{S}_{n,j}^{(\iota)}$, and $\grave{S}_{n,j}^{(\iota)}$ in Lemmas 2.14(i) and 2.14(ii), we have for $j = 0, 1$,

$$\begin{aligned}
T_{n,j}^{(\iota)} =& \frac{1}{n}\sum_{i=1}^{n} X_i \left[X_i^\top a(Z_i) + \varepsilon_i\right] \left(\frac{Z_i - z}{h_n}\right)^j K_h^{(\iota)}(Z_i - z) \\
=& \frac{1}{n}\sum_{i:Z_i<s_q} X_i X_i^\top \left(\frac{Z_i - z}{h_n}\right)^j K_h^{(\iota)}(Z_i - z) a(Z_i) + \acute{F}_{n,j}^{(\iota)} \\
&+ \frac{1}{n}\sum_{i:Z_i\geq s_q} X_i X_i^\top \left(\frac{Z_i - z}{h_n}\right)^j K_h^{(\iota)}(Z_i - z) a(Z_i) + \grave{F}_{n,j}^{(\iota)} \\
=& \frac{1}{n}\sum_{i:Z_i<s_q} X_i X_i^\top \left(\frac{Z_i - z}{h_n}\right)^j K_h^{(\iota)}(Z_i - z) \left\{a_-(s_q) + \mathrm{O}(Z_i - s_q)\right\} + \mathrm{o_p}(1) \\
&+ \frac{1}{n}\sum_{i:Z_i\geq s_q} X_i X_i^\top \left(\frac{Z_i - z}{h_n}\right)^j K_h^{(\iota)}(Z_i - z) \left\{a_+(s_q) + \mathrm{O}(Z_i - s_q)\right\} + \mathrm{o_p}(1) \\
=& \acute{S}_{n,j}^{(\iota)} a_-(s_q) + \grave{S}_{n,j}^{(\iota)} \left\{a_-(s_q) + d_q\right\} + \mathrm{O_p}(h_n) + \mathrm{o_p}(1) \\
=& f_Z(s_q) \left[\left\{\acute{\mu}_{j,\tau}^{(\iota)}\Omega_-(s_q) + \grave{\mu}_{j,\tau}^{(\iota)}\Omega_+(s_q)\right\} a_-(s_q) + \grave{\mu}_{j,\tau}^{(\iota)}\Omega_+(s_q) d_q\right] + \mathrm{o_p}(1).
\end{aligned}$$

Hence, by Lemmas 2.14(i), 2.18(ii), and 2.19, the local linear estimator in (2.18) can be expressed for $z = s_q + \tau h_n$ with $\tau \in (-1, 1)$ as

$$\begin{aligned}
H_n \hat{\beta}_n^{(\iota)} =& {S_n^{(\iota)}}^{-1} T_n^{(\iota)} \\
=& \left[\acute{\Omega}_{-,\tau}^{(\iota)}(s_q) + \grave{\Omega}_{+,\tau}^{(\iota)}(s_q)\right]^{-1} (1 + \mathrm{o_p}(1)) \\
& \cdot \left[\begin{pmatrix} \acute{\mu}_{0,\tau}^{(\iota)}\Omega_-(s_q) + \grave{\mu}_{0,\tau}^{(\iota)}\Omega_+(s_q) \\ \acute{\mu}_{1,\tau}^{(\iota)}\Omega_-(s_q) + \grave{\mu}_{1,\tau}^{(\iota)}\Omega_+(s_q) \end{pmatrix} a_-(s_q) + \begin{pmatrix} \grave{\mu}_{0,\tau}^{(\iota)}\Omega_+(s_q) \\ \grave{\mu}_{1,\tau}^{(\iota)}\Omega_+(s_q) \end{pmatrix} d_q + \mathrm{o_p}(1)\right] \\
=& \begin{pmatrix} I_p \\ 0_p \end{pmatrix} a_-(s_q) + \begin{pmatrix} \Xi_{0,\tau}^{(\iota)} \\ \Xi_{1,\tau}^{(\iota)} \end{pmatrix} d_q + \mathrm{o_p}(1),
\end{aligned} \tag{2.35}$$

where I_p is the $p \times p$ identity matrix, 0_p is the null matrix of size $p \times p$, and

$$\begin{pmatrix} \Xi_{0,\tau}^{(\iota)} \\ \Xi_{1,\tau}^{(\iota)} \end{pmatrix} = \left[\acute{\Omega}_{-,\tau}^{(\iota)}(s_q) + \grave{\Omega}_{+,\tau}^{(\iota)}(s_q)\right]^{-1} \begin{pmatrix} \grave{\mu}_{0,\tau}^{(\iota)}\Omega_+(s_q) \\ \grave{\mu}_{1,\tau}^{(\iota)}\Omega_+(s_q) \end{pmatrix}. \tag{2.36}$$

Note that, according to the definition of the right-sided kernel $K^{(r)}(\cdot)$ in (2.2), one has for $\tau \in (0,1)$,

$$\acute{\mu}_{j,\tau}^{(r)} = \int_{-1}^{-\tau} u^j K^{(c)}(u)\mathbf{1}\{u \geq 0\}\,\mathrm{d}u = 0, \tag{2.37}$$

which implies that $\acute{\Omega}_{-,\tau}^{(r)}(s_q) = 0_{2p}$ and

$$\begin{pmatrix} \Xi_{0,\tau}^{(r)} \\ \Xi_{1,\tau}^{(r)} \end{pmatrix} = \grave{\Omega}_{+,\tau}^{(r)-1}(s_q) \begin{pmatrix} \grave{\mu}_{0,\tau}^{(r)}\Omega_+(s_q) \\ \grave{\mu}_{1,\tau}^{(r)}\Omega_+(s_q) \end{pmatrix} = \begin{pmatrix} I_p \\ 0_p \end{pmatrix} \tag{2.38}$$

due to Lemma 2.18(ii). Similarly, for $\tau \in (-1,0)$ and the left-sided kernel $K^{(l)}(\cdot)$, we obtain

$$\grave{\mu}_{j,\tau}^{(l)} = 0, \quad \Xi_{0,\tau}^{(l)} = 0_p, \quad \text{and} \quad \Xi_{1,\tau}^{(l)} = 0_p. \tag{2.39}$$

Proof of Theorem 2.4.

In order to prove Theorem 2.4 for continuous conditional error variance function $\sigma^2(z)$ (Assumption 2.A6), we analyze the limiting properties of each term of the decomposition of $N_n^{(\iota)}$ in (2.26). First, by Lemma 2.14(iv), $N_{n,1}^{(\iota)} = f_Z(s_q)\mu_0^{(\iota)}\sigma^2(s_q) + \mathrm{o_p}(1)$. Using equations (2.33)–(2.35), one obtains

$$\begin{aligned}
N_{n,2}^{(\iota)} =& \frac{2}{n}\sum_{i=1}^{n}\left[a(Z_i) - \hat{a}_n^{(\iota)}(z) - h_n\hat{b}_n^{(\iota)}(z)\left(\frac{Z_i - z}{h_n}\right)\right]^\top X_i\varepsilon_i K_h^{(\iota)}(Z_i - z) \\
=& \frac{2}{n}\sum_{i:Z_i<s_q}\left[\underbrace{a(Z_i) - a_-(s_q)}_{\mathrm{O}(Z_i - s_q)} - \Xi_{0,\tau}^{(\iota)}d_q - \left(\frac{Z_i - z}{h_n}\right)\Xi_{1,\tau}^{(\iota)}d_q\right]^\top X_i\varepsilon_i K_h^{(\iota)}(Z_i - z) \\
&+ \frac{2}{n}\sum_{i:Z_i\geq s_q}\left[\underbrace{a(Z_i) - a_-(s_q)}_{d_q + \mathrm{O}(Z_i - s_q)} - \Xi_{0,\tau}^{(\iota)}d_q - \left(\frac{Z_i - z}{h_n}\right)\Xi_{1,\tau}^{(\iota)}d_q\right]^\top X_i\varepsilon_i K_h^{(\iota)}(Z_i - z) \\
&+ \mathrm{o_p}(1) \\
=& -2[\Xi_{0,\tau}^{(\iota)}d_q]^\top\acute{F}_{n,0}^{(\iota)} - 2[\Xi_{1,\tau}^{(\iota)}d_q]^\top\acute{F}_{n,1}^{(\iota)} - 2[(\Xi_{0,\tau}^{(\iota)} - I_p)d_q]^\top\grave{F}_0^{(\iota)} - 2[\Xi_{1,\tau}^{(\iota)}d_q]^\top\grave{F}_{n,1}^{(\iota)} \\
&+ \mathrm{O_p}(h_n) + \mathrm{o_p}(1).
\end{aligned}$$

Hence, $N_{n,2}^{(\iota)} = \mathrm{o_p}(1)$ due to the limiting results for $\acute{F}_{n,j}^{(\iota)}$ and $\grave{F}_{n,j}^{(\iota)}$ in Lemma 2.14(ii). Again, it follows from (2.33)–(2.35) that

$$\begin{aligned}
N_{n,3}^{(\iota)} =& \frac{1}{n}\sum_{i=1}^{n}\left[X_i^\top a(Z_i) - X_i^\top\left\{\hat{a}_n^{(\iota)}(z) + h_n\hat{b}_n^{(\iota)}(z)\left(\frac{Z_i - z}{h_n}\right)\right\}\right]^2 K_h^{(\iota)}(Z_i - z)\\
=& \frac{1}{n}\sum_{i:Z_i<s_q}\left[X_i^\top\left\{\underbrace{a(Z_i) - a_-(s_q)}_{\mathrm{O}(Z_i - s_q)} - \Xi_{0,\tau}^{(\iota)} d_q - \left(\frac{Z_i - z}{h_n}\right)\Xi_{1,\tau}^{(\iota)} d_q\right\}\right]^2 K_h^{(\iota)}(Z_i - z)\\
&+ \frac{1}{n}\sum_{i:Z_i\geq s_q}\left[X_i^\top\left\{\underbrace{a(Z_i) - a_-(s_q)}_{d_q + \mathrm{O}(Z_i - s_q)} - \Xi_{0,\tau}^{(\iota)} d_q - \left(\frac{Z_i - z}{h_n}\right)\Xi_{1,\tau}^{(\iota)} d_q\right\}\right]^2 K_h^{(\iota)}(Z_i - z)\\
&+ \mathrm{o_p}(1)\\
=& d_q^\top {\Xi_{0,\tau}^{(\iota)}}^\top \acute{S}_{n,0}^{(\iota)}\Xi_{0,\tau}^{(\iota)} d_q + 2d_q^\top {\Xi_{0,\tau}^{(\iota)}}^\top \acute{S}_{n,1}^{(\iota)}\Xi_{1,\tau}^{(\iota)} d_q + d_q^\top {\Xi_{1,\tau}^{(\iota)}}^\top \acute{S}_{n,2}^{(\iota)}\Xi_{1,\tau}^{(\iota)} d_q\\
&+ d_q^\top[\Xi_{0,\tau}^{(\iota)} - I_p]^\top \grave{S}_{n,0}^{(\iota)}[\Xi_{0,\tau}^{(\iota)} - I_p] d_q + 2d_q^\top[\Xi_{0,\tau}^{(\iota)} - I_p]^\top \grave{S}_{n,1}^{(\iota)}\Xi_{1,\tau}^{(\iota)} d_q\\
&+ d_q^\top {\Xi_{1,\tau}^{(\iota)}}^\top \grave{S}_{n,2}^{(\iota)}\Xi_{1,\tau}^{(\iota)} d_q + \mathrm{O_p}(h_n) + \mathrm{o_p}(1)\\
=& d_q^\top \begin{pmatrix}\Xi_{0,\tau}^{(\iota)}\\ \Xi_{1,\tau}^{(\iota)}\end{pmatrix}^\top \begin{bmatrix}\acute{S}_{n,0}^{(\iota)} & \acute{S}_{n,1}^{(\iota)}\\ \acute{S}_{n,1}^{(\iota)} & \acute{S}_{n,2}^{(\iota)}\end{bmatrix}\begin{pmatrix}\Xi_{0,\tau}^{(\iota)}\\ \Xi_{1,\tau}^{(\iota)}\end{pmatrix} d_q\\
&+ d_q^\top \begin{pmatrix}\Xi_{0,\tau}^{(\iota)} - I_p\\ \Xi_{1,\tau}^{(\iota)}\end{pmatrix}^\top \begin{bmatrix}\grave{S}_{n,0}^{(\iota)} & \grave{S}_{n,1}^{(\iota)}\\ \grave{S}_{n,1}^{(\iota)} & \grave{S}_{n,2}^{(\iota)}\end{bmatrix}\begin{pmatrix}\Xi_{0,\tau}^{(\iota)} - I_p\\ \Xi_{1,\tau}^{(\iota)}\end{pmatrix} d_q + \mathrm{O_p}(h_n) + \mathrm{o_p}(1).
\end{aligned}$$

It follows from the convergence results for $\acute{S}_{n,j}^{(\iota)}$ and $\grave{S}_{n,j}^{(\iota)}$ in Lemma 2.14(i) that

$$\begin{aligned}
N_{n,3}^{(\iota)} =& f_Z(s_q) d_q^\top\left\{\begin{pmatrix}\Xi_{0,\tau}^{(\iota)}\\ \Xi_{1,\tau}^{(\iota)}\end{pmatrix}^\top \acute{\Omega}_{-,\tau}^{(\iota)}(s_q)\begin{pmatrix}\Xi_{0,\tau}^{(\iota)}\\ \Xi_{1,\tau}^{(\iota)}\end{pmatrix}\right.\\
&\left.+ \begin{pmatrix}\Xi_{0,\tau}^{(\iota)} - I_p\\ \Xi_{1,\tau}^{(\iota)}\end{pmatrix}^\top \grave{\Omega}_{+,\tau}^{(\iota)}(s_q)\begin{pmatrix}\Xi_{0,\tau}^{(\iota)} - I_p\\ \Xi_{1,\tau}^{(\iota)}\end{pmatrix}\right\} d_q + \mathrm{o_p}(1)\\
=& f_Z(s_q) d_q^\top \mu_0^{(\iota)} C_\tau^{(\iota)} d_q + \mathrm{o_p}(1),
\end{aligned}$$

where

$$C_\tau^{(\iota)} = \frac{1}{\mu_0^{(\iota)}} \begin{pmatrix} \Xi_{0,\tau}^{(\iota)} \\ \Xi_{1,\tau}^{(\iota)} \end{pmatrix}^\top \acute{\Omega}_{-,\tau}^{(\iota)}(s_q) \begin{pmatrix} \Xi_{0,\tau}^{(\iota)} \\ \Xi_{1,\tau}^{(\iota)} \end{pmatrix} + \frac{1}{\mu_0^{(\iota)}} \begin{pmatrix} \Xi_{0,\tau}^{(\iota)} - I_p \\ \Xi_{1,\tau}^{(\iota)} \end{pmatrix}^\top \grave{\Omega}_{+,\tau}^{(\iota)}(s_q) \begin{pmatrix} \Xi_{0,\tau}^{(\iota)} - I_p \\ \Xi_{1,\tau}^{(\iota)}. \end{pmatrix}. \tag{2.40}$$

For $\tau \in (0,1)$ and $\iota = c$ or l, $\acute{\mu}_{0,\tau}^{(\iota)}$ and $\grave{\mu}_{0,\tau}^{(\iota)}$ are nonzero. According to Lemma 2.20, both matrices $\Xi_{0,\tau}^{(\iota)}$ and $\Xi_{0,\tau}^{(\iota)} - I_p$ have rank p. Hence, $[\Xi_{0,\tau}^{(\iota)} \;\; \Xi_{1,\tau}^{(\iota)}]^\top$ and $[\Xi_{0,\tau}^{(\iota)} - I_p \;\; \Xi_{1,\tau}^{(\iota)}]^\top$ have the same rank p, thus full column rank. By the result that $\acute{\Omega}_{-,\tau}^{(\iota)}(s_q)$ and $\grave{\Omega}_{+,\tau}^{(\iota)}(s_q)$ are positive definite for $\tau \in (0,1)$ and $\iota = c$ or l in Lemma 2.19, the property that $A + B \succ 0$ for any $A \succ 0$ and $B \succ 0$, and the fact that $A^\top B A \succ 0$ if $B \succ 0$ and A has full column rank, we conclude that matrices $C_\tau^{(c)}$ and $C_\tau^{(l)}$ are positive definite for $\tau \in (0,1)$.

For $\tau \in (0,1)$ and $\iota = r$, it follows from equation (2.37): $\acute{\mu}_{j,\tau}^{(r)} = 0$ that

$$C_\tau^{(r)} = \frac{1}{\mu_0^{(r)}} \begin{pmatrix} \Xi_{0,\tau}^{(r)} - I_p \\ \Xi_{1,\tau}^{(r)} \end{pmatrix}^\top \begin{bmatrix} \grave{\mu}_{0,\tau}^{(r)}\Omega_+(s_q) & \grave{\mu}_{1,\tau}^{(r)}\Omega_+(s_q) \\ \grave{\mu}_{1,\tau}^{(r)}\Omega_+(s_q) & \grave{\mu}_{2,\tau}^{(r)}\Omega_+(s_q) \end{bmatrix} \begin{pmatrix} \Xi_{0,\tau}^{(r)} - I_p \\ \Xi_{1,\tau}^{(r)} \end{pmatrix}.$$

Since $\Xi_{0,\tau}^{(r)} = I_p$ and $\Xi_{1,\tau}^{(r)} = 0_p$ (equation (2.38)), $C_\tau^{(r)}$ is a null matrix for $\tau \in (0,1)$.

Similarly, for $\tau \in (-1,0)$, we have positive definite matrices $C_\tau^{(c)} \succ 0$ and $C_\tau^{(r)} \succ 0$ and the null matrix $C_\tau^{(l)} = 0_p$. Combining the limiting results of $N_{n,1}^{(\iota)}$, $N_{n,2}^{(\iota)}$, $N_{n,3}^{(\iota)}$, and $K_n^{(\iota)}$ (due to Lemma 2.14(iii)) yields Theorem 2.4. □

Proof of Theorem 2.5.

Following the proof of Theorem 3.2 in Gijbels et al. (2007), we write the jump-preserving estimator $\check{a}_n(z)$ as

$$\check{a}_n(z) = \hat{a}_n^{(c)}(z)\mathbf{1}\{A_n(z)\} + \hat{a}_n^{(l)}(z)\mathbf{1}\{B_n(z)\} + \hat{a}_n^{(r)}(z)\mathbf{1}\{C_n(z)\} + \frac{\hat{a}_n^{(l)}(z) + \hat{a}_n^{(r)}(z)}{2}\mathbf{1}\{BC_n(z)\},$$

in which $A_n(z)$, $B_n(z)$, $C_n(z)$, and $BC_n(z)$ correspond to the inequalities in (2.6) from top to bottom, respectively. Apparently, these sets are mutually exclusive, and for any $z \in D$,

$$\mathbf{1}\{A_n(z)\} + \mathbf{1}\{B_n(z)\} + \mathbf{1}\{C_n(z)\} + \mathbf{1}\{BC_n(z)\} = 1. \tag{2.41}$$

The rest of the proof is separated into three parts, which correspond to the regions D_{1n}, $D_{2n,\delta}$ for some $\delta \in (0, 1/2)$, and D_{2n} given in equation (2.9).

Part (i)

First, we consider z in the continuous region D_{1n}. According to Theorem 2.1, there exist a positive integer $n^{(\iota)}$ and a positive constant $C^{(\iota)} > 0$ such that for $n > n^{(\iota)}$,

$$\sup_{z \in D_{1n}} \sqrt{\frac{nh_n}{\ln n}} \left\| \hat{a}_n^{(\iota)}(z) - a(z) \right\| \leq C^{(\iota)}, \quad \iota = c, l, r,$$

with probability approaching to 1. Take $\zeta = \max_{\iota=\{c,l,r\}} C^{(\iota)}$; for $n > \max_{\iota=\{c,l,r\}} n^{(\iota)}$, it follows that

$$\begin{aligned}
\sup_{z \in D_{1n}} \sqrt{\frac{nh_n}{\ln n}} \left\| \check{a}_n(z) - a(z) \right\| = & \sup_{z \in D_{1n}} \sqrt{\frac{nh_n}{\ln n}} \left\| \hat{a}_n^{(c)}(z) - a(z) \right\| \mathbf{1}\{A_n(z)\} \\
& + \sup_{z \in D_{1n}} \sqrt{\frac{nh_n}{\ln n}} \left\| \hat{a}_n^{(l)}(z) - a(z) \right\| \mathbf{1}\{B_n(z)\} \\
& + \sup_{z \in D_{1n}} \sqrt{\frac{nh_n}{\ln n}} \left\| \hat{a}_n^{(r)}(z) - a(z) \right\| \mathbf{1}\{C_n(z)\} \\
& + \sup_{z \in D_{1n}} \sqrt{\frac{nh_n}{\ln n}} \left\| \frac{\hat{a}_n^{(l)}(z) + \hat{a}_n^{(r)}(z)}{2} - a(z) \right\| \mathbf{1}\{BC_n(z)\} \\
\leq & \zeta
\end{aligned}$$

with probability approaching to 1, which implies that

$$\sup_{z \in D_{1n}} \sqrt{\frac{nh_n}{\ln n}} \left\| \check{a}_n(z) - a(z) \right\| = \mathrm{O_p}(1).$$

Part (ii)

Next, we prove the uniform consistency for $\check{a}_n(z)$ in the region $D_{2n,\delta}$ for some $\delta \in (0, 1/2)$,

which contains neighborhoods of discontinuities excluding any small regions around centers of s_q and around end points $s_q - h_n$ and $s_q + h_n$. For some $\delta \in (0, 1/2)$, the region $D_{2n,\delta}$ consists of two disjoint sets:

$$\acute{D}_{2n,\delta} = D \cap \bigcup_{q=0}^{Q+1} [s_q - (1-\delta)h_n, s_q - \delta h_n]$$

and

$$\grave{D}_{2n,\delta} = D \cap \bigcup_{q=0}^{Q+1} [s_q + \delta h_n, s_q + (1-\delta)h_n].$$

Consider the region $\acute{D}_{2n,\delta}$ and an arbitrarily small number $\epsilon > 0$. Any given point z in $\acute{D}_{2n,\delta}$ satisfies $z = s_q + \tau h_n$ with $\tau \in [-1+\delta, -\delta]$ and s_q is one of $\{s_q\}_{q=0}^{Q+1}$. According to Theorem 2.1, for some $\zeta > 0$ and any $\epsilon > 0$, there exist a positive integer n_1 such that for $n > n_1$,

$$\sup_{z \in \acute{D}_{2n,\delta}} \sqrt{\frac{nh_n}{\ln n}} \left\| \hat{a}_n^{(l)}(z) - a(z) \right\| \leq \zeta$$

with probability larger than $1 - \epsilon$. In the following, we show that for any $z \in \acute{D}_{2n,\delta}$, there exists another positive integer $n_3 > 0$ such that the difference of $\check{a}_n(z)$ and $\hat{a}_n^{(l)}(z)$ is negligible in probability.

By Theorem 2.4, for any $\kappa > 0$ and $\epsilon > 0$, there exists an integer $n_\kappa(\kappa)$ such that for $n > n_\kappa(\kappa)$,

$$\begin{aligned}
&\Psi_n^{(c)}(z) > d_q^\top C_\tau^{(c)} d_q + \sigma^2(s_q) - \kappa, \\
&\Psi_n^{(l)}(z) < \sigma^2(s_q) + \kappa, \\
&\Psi_n^{(r)}(z) > d_q^\top C_\tau^{(r)} d_q + \sigma^2(s_q) - \kappa
\end{aligned}$$

with probability larger than $1-\epsilon$. For $\tau \in [-1+\delta, -\delta]$, matrices $C_\tau^{(c)}$ and $C_\tau^{(r)}$ are positive definite (see the proof of Theorem 2.4). Additionally, the continuity of $C_\tau^{(\iota)}$ in τ follows from the continuity of $\acute{\mu}_{j,\tau}^{(\iota)}$ and $\grave{\mu}_{j,\tau}^{(\iota)}$ as functions of the limits of integration. Given the

continuity of $C_\tau^{(\iota)}$ and thus of $d_q^\top C_\tau^{(\iota)} d_q$, we have for any $d_q \neq 0$,

$$\begin{aligned} a_\tau &= \inf_{\tau\in[-1+\delta,-\delta]} \min\{d_q^\top C_\tau^{(c)} d_q, d_q^\top C_\tau^{(r)} d_q\} \\ &= \min_{\tau\in[-1+\delta,-\delta]} \min\{d_q^\top C_\tau^{(c)} d_q, d_q^\top C_\tau^{(r)} d_q\} > 0. \end{aligned}$$

Set $\kappa = \frac{a_\tau}{4}$. For $n > n_2 = n_\kappa(\frac{a_\tau}{4})$, it follows that

$$\begin{aligned} \Psi_n^{(c)}(z) - \Psi_n^{(l)}(z) &\geq \min\{\Psi_n^{(c)}(z), \Psi_n^{(r)}(z)\} - \Psi_n^{(l)}(z) \\ &> a_\tau - 2\kappa = a_\tau - \frac{a_\tau}{2} = \frac{a_\tau}{2} > 0, \end{aligned}$$

and hence,

$$\begin{aligned} \mathrm{diff}(z) &= \Psi_n^{(c)}(z) - \min\{\Psi_n^{(l)}(z), \Psi_n^{(r)}(z)\} \\ &= \Psi_n^{(c)}(z) - \Psi_n^{(l)}(z) > \frac{a_\tau}{2} > 0 \end{aligned}$$

with probability larger than $1-\epsilon$. Moreover, since $u_n \to 0$, for any $\eta > 0$ there exists $n_\eta(\eta) > 0$ such that, for $n > n_\eta(\eta)$, we have $|u_n| < \eta$. Setting $\eta = a_\tau/4$, it follows for $n > n_3 = \max\{n_\eta(\frac{a_\tau}{4}), n_2\}$,

$$\mathrm{diff}(z) - u_n > \frac{a_\tau}{2} - u_n > \frac{a_\tau}{2} - \frac{a_\tau}{4} = \frac{a_\tau}{4} > 0,$$

which implies that Conditions $A_n(z)$, $C_n(z)$, and $BC_n(z)$ do not hold, i.e., $\mathbf{1}\{A_n(z)\} + \mathbf{1}\{C_n(z)\} + \mathbf{1}\{BC_n(z)\} = 0$ with probability larger than $1-2\epsilon$. Moreover, by equation (2.41), we can claim with an arbitrarily high probability that only Condition $B_n(z)$ is satisfied, which means that $\hat{a}_n^{(l)}(z)$ is chosen for $n > n_3$ with probability larger than $1-2\epsilon$. Hence when $n > n_4 = \max\{n_1, n_3\}$,

$$\sup_{z\in\acute{D}_{2n,\delta}} \sqrt{\frac{nh_n}{\ln n}} \left\|\check{a}_n(z) - a(z)\right\| = \sup_{z\in\acute{D}_{2n,\delta}} \sqrt{\frac{nh_n}{\ln n}} \left\|\hat{a}_n^{(l)}(z) - a(z)\right\| \leq \zeta$$

with probability larger than $1-3\epsilon$, which implies

$$\sup_{z\in\acute{D}_{2n,\delta}} \sqrt{\frac{nh_n}{\ln n}} \left\|\check{a}_n(z) - a(z)\right\| = \mathrm{O_p}(1). \tag{2.42}$$

Similarly, for $z \in \grave{D}_{2n,\delta}$, one can also show that

$$\sup_{z\in\grave{D}_{2n,\delta}} \sqrt{\frac{nh_n}{\ln n}} \left\|\check{a}_n(z) - a(z)\right\| = \sup_{z\in\grave{D}_{2n,\delta}} \sqrt{\frac{nh_n}{\ln n}} \left\|\hat{a}_n^{(r)}(z) - a(z)\right\| + \mathrm{o_p}(1)$$
$$= \mathrm{O_p}(1). \tag{2.43}$$

Combining (2.42) and (2.43) gives

$$\sup_{z\in D_{2n,\delta}} \sqrt{\frac{nh_n}{\ln n}} \left\|\check{a}_n(z) - a(z)\right\| = \mathrm{O_p}(1).$$

Part (iii)

For $z \in D_{2n} \setminus D_{2n,\delta}$, we can show the consistency of $\check{a}_n(z)$ analogously to the proof of Part (ii). Since there is no unique strictly positive lower bound a_τ exists, the result is not uniform with respect to z on $D_{2n} \setminus D_{2n,\delta}$. □

Proof of Theorem 2.6.

We showed in the proof of Theorem 2.5 that the jump-preserving estimator $\check{a}_n(z)$ picks consistently the correct local estimator for $z \in D \setminus \{s_q\}_{q=0}^{Q+1}$. By Theorem 2.2, each local linear estimator is asymptotically normal in the regions, where it is selected. Consequently, $\check{a}_n(z)$ is asymptotically normal for $z \in D \setminus \{s_q\}_{q=0}^{Q+1}$ with distribution given in Theorem 2.6. A detailed argument is given in the proof of Theorem 3.1 of Casas and Gijbels (2012). □

Proof of Theorem 2.7.

Recall that the estimated residual used in Theorems 2.7–2.10 is $\tilde{\varepsilon}_{n,i}^{(\iota)} = Y_i - X_i^\top \hat{a}_n^{(\iota)}(z)$ and the kernel $\tilde{K}$ refers to the uniform kernel. Let us denote

$$\tilde{N}_n^{(\iota)} = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_{n,i}^{(\iota)2} \tilde{K}_h^{(\iota)}(Z_i - z),$$
$$\tilde{T}_n^{(\iota)} = \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{Z_i - z}{h_n}\right) \tilde{\varepsilon}_{n,i}^{(\iota)} \tilde{K}_h^{(\iota)}(Z_i - z),$$

$$\tilde{T}_{n,2}^{(\iota)} = \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{Z_i - z}{h_n}\right) X_i^\top \{a(Z_i) - \hat{a}_n^{(\iota)}(z)\} \tilde{K}_h^{(\iota)}(Z_i - z),$$

$$\tilde{S}_n^{(\iota)} = \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{Z_i - z}{h_n}\right) \rho^\top\left(\frac{Z_i - z}{h_n}\right) \tilde{K}_h^{(\iota)}(Z_i - z), \tag{2.44}$$

$$\tilde{W}_{n,1}^{(\iota)} = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 \tilde{K}_h^{(\iota)}(Z_i - z), \tag{2.45}$$

$$\tilde{W}_{n,2}^{(\iota)} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top \tilde{K}_h^{(\iota)}(Z_i - z), \tag{2.46}$$

$$\tilde{W}_{n,3}^{(\iota)} = \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{Z_i - z}{h_n}\right) X_i^\top \tilde{K}_h^{(\iota)}(Z_i - z), \tag{2.47}$$

$$\tilde{W}_{n,4}^{(\iota)} = \frac{1}{n}\sum_{i=1}^{n} X_i \varepsilon_i \tilde{K}_h^{(\iota)}(Z_i - z), \quad \text{and} \tag{2.48}$$

$$\tilde{W}_{n,5}^{(\iota)} = \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{Z_i - z}{h_n}\right) \varepsilon_i \tilde{K}_h^{(\iota)}(Z_i - z), \tag{2.49}$$

where $\rho(u) = (1, u, \ldots, u^m)^\top$ and $\hat{e}_{n,i}^{(\iota)} = \tilde{\varepsilon}_{n,i}^{(\iota)} - \rho^\top((Z_i - z)/h_n)\hat{\gamma}_n^{(\iota)}$. Further, we define the population counterparts of some of the above kernel weighted averages:

$$\mho(z) = \mathrm{E}[X^\top | Z = z], \quad \tilde{\mu}_0^{(\iota)} = \int_{-1}^{1} \tilde{K}(u)\mathrm{d}u, \quad \tilde{m}^{(\iota)} = \int_{-1}^{1} \rho(u)\tilde{K}(u)\mathrm{d}u, \quad \text{and}$$
$$\tilde{M}^{(\iota)} = \int_{-1}^{1} \rho(u)\rho^\top(u)\tilde{K}(u)\mathrm{d}u. \tag{2.50}$$

With the help of the above notation, we write $\hat{\gamma}_n^{(\iota)}$ in (2.10) as

$$\hat{\gamma}_n^{(\iota)}(z) = {\tilde{S}_n^{(\iota)}}^{-1} \tilde{T}_n^{(\iota)} = {\tilde{S}_n^{(\iota)}}^{-1} (\tilde{W}_{n,5}^{(\iota)} + \tilde{T}_{n,2}^{(\iota)}). \tag{2.51}$$

By Lemma 2.15(vi), $\tilde{W}_{n,5}^{(\iota)} = \mathrm{o_p}(1)$. To show $\tilde{T}_{n,2}^{(\iota)} = \mathrm{o_p}(1)$ by the convergence results for $\hat{a}_n^{(\iota)}(z)$ and $\tilde{W}_{n,3}^{(\iota)}$ in Lemmas 2.11(vi) and 2.15(iv), respectively, the Taylor expansion of $a(Z_i)$ for $Z_i \in [z - h_n, z] : z \in D_{1n}^{(l)}$, $Z_i \in [z, z + h_n] : z \in D_{1n}^{(r)}$, or $Z_i \in [z - h_n, z + h_n] :$

$z \in D_{1n}^{(c)}$ is used along with the boundedness of $a'(\cdot)$ (Assumption 2.A5):

$$\begin{aligned}\tilde{T}_{n,2}^{(\iota)} =& \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{Z_i - z}{h_n}\right) X_i^\top \{a(z) - \hat{a}_n^{(\iota)}(z) + \mathrm{O}(h_n)\} \tilde{K}_h^{(\iota)}(Z_i - z) \\ =& \tilde{W}_{n,3}^{(\iota)}[a(z) - \hat{a}_n^{(\iota)}(z) + \mathrm{O}(h_n)] \\ =& \left\{f_Z(z)\tilde{m}^{(\iota)}\mho(z) + \mathrm{o_p}(1)\right\} \cdot \mathrm{o_p}(1) = \mathrm{o_p}(1). \qquad (h_n \to 0)\end{aligned}$$

As $\tilde{T}_n^{(\iota)} = \tilde{W}_{n,5}^{(\iota)} + \tilde{T}_{n,2}^{(\iota)} = \mathrm{o_p}(1)$ by Lemma 2.15(vi), the convergence result for $\tilde{S}_n^{(\iota)}$ in Lemma 2.15(i) and the invertibility conditions for its population counterparts in Assumptions 2.A2 and 2.D1 imply

$$\hat{\gamma}_n^{(\iota)} = \frac{\tilde{M}^{(\iota)-1}}{f_Z(z)}\mathrm{o_p}(1) = \mathrm{o_p}(1). \tag{2.52}$$

Further, for $Z_i \in [z - h_n, z] : z \in D_{1n}^{(l)}$, $Z_i \in [z, z + h_n] : z \in D_{1n}^{(r)}$, or $Z_i \in [z - h_n, z + h_n] : z \in D_{1n}^{(c)}$, the squared error from the local mth polynomial fitting of $\tilde{\varepsilon}_{n,i}^{(\iota)}$ equals

$$\begin{aligned}\hat{e}_{n,i}^{(\iota)2} =& \varepsilon_i^2 + \left\{\hat{\gamma}_n^{(\iota)\top}\rho\left(\frac{Z_i - z}{h_n}\right)\right\}^2 + \left\{X_i^\top[a(Z_i) - \hat{a}_n^{(\iota)}(z)]\right\}^2 \\ & - 2\hat{\gamma}_n^{(\iota)\top}\rho\left(\frac{Z_i - z}{h_n}\right) X_i^\top[a(Z_i) - \hat{a}_n^{(\iota)}(z)] + 2\varepsilon_i X_i^\top\left\{a(Z_i) - \hat{a}_n^{(\iota)}(z)\right\} \\ & - 2\hat{\gamma}_n^{(\iota)\top}\rho\left(\frac{Z_i - z}{h_n}\right)\varepsilon_i \\ =& \varepsilon_i^2 + \left\{\hat{\gamma}_n^{(\iota)\top}\rho\left(\frac{Z_i - z}{h_n}\right)\right\}^2 + \left\{X_i^\top[a(z) + \mathrm{O}(h_n) - \hat{a}_n^{(\iota)}(z)]\right\}^2 \\ & - 2\hat{\gamma}_n^{(\iota)\top}\rho\left(\frac{Z_i - z}{h_n}\right) X_i^\top[a(z) + \mathrm{O}(h_n) - \hat{a}_n^{(\iota)}(z)] \\ & + 2\varepsilon_i X_i^\top\left\{a(z) + \mathrm{O}(h_n) - \hat{a}_n^{(\iota)}(z)\right\} - 2\hat{\gamma}_n^{(\iota)\top}\rho\left(\frac{Z_i - z}{h_n}\right)\varepsilon_i\end{aligned}$$

uniformly in $i \in \mathbb{N}$ by the Taylor expansion of $a(\cdot)$. To analyze each term of $\tilde{\Psi}_n^{(\iota)}(z)$ in (2.11), let us now look at $\tilde{N}_n^{(\iota)}$. By Lemma 2.15 and (2.52), we have after substitution

for $\hat{e}_{n,i}^{(\iota)2}$

$$\begin{aligned}
\tilde{N}_n^{(\iota)} =& \frac{1}{n}\sum_{i=1}^{n} \hat{e}_{n,i}^{(\iota)2} \tilde{K}_h^{(\iota)}(Z_i - z) \\
=& \tilde{W}_{n,1}^{(\iota)} + \hat{\gamma}_n^{(\iota)\top} \tilde{S}_n^{(\iota)} \hat{\gamma}_n^{(\iota)} + [a(z) - \hat{a}_n^{(\iota)}(z)]^\top \tilde{W}_{n,2}^{(\iota)} [a(z) - \hat{a}_n^{(\iota)}(z)](1 + \mathrm{O}(h_n)) \\
&+ \left(-2\hat{\gamma}_n^{(\iota)\top} \tilde{W}_{n,3}^{(\iota)} + 2\tilde{W}_{n,4}^{(\iota)\top}\right) [a(z) - \hat{a}_n^{(\iota)}(z)](1 + \mathrm{O}(h_n)) - 2\hat{\gamma}_n^{(\iota)\top} \tilde{W}_{n,5} \\
=& f_Z(z)\tilde{\mu}_0^{(\iota)}\sigma^2(z) + \mathrm{o_p}(1) f_Z(z) \tilde{M}^{(\iota)} \mathrm{o_p}(1) + \mathrm{o_p}(1) f_Z(z) \tilde{\mu}_0^{(\iota)} \Omega(z) \mathrm{o_p}(1) \\
&- 2\mathrm{o_p}(1) f_Z(z) \tilde{m}^{(\iota)} \mho(z) \mathrm{o_p}(1) + 2\mathrm{o_p}(1)\mathrm{o_p}(1) - 2\mathrm{o_p}(1)\mathrm{o_p}(1) \\
=& f_Z(z)\tilde{\mu}_0^{(\iota)}\sigma^2(z) + \mathrm{o_p}(1). \qquad (2.53)
\end{aligned}$$

Combining equations (2.52) and (2.53), Lemma 2.15(i), and Assumption 2.D1 finally yields

$$\begin{aligned}
\tilde{\Psi}_n^{(\iota)}(z) =& \mathrm{o_p}(1) \frac{\tilde{M}^{(\iota)-1}}{f_Z(z)} (1 + \mathrm{o_p}(1)) \left(\frac{f_Z(z)\tilde{M}^{(\iota)} + \mathrm{o_p}(1)}{f_Z(z)\tilde{\mu}_0^{(\iota)}\sigma^2(z)} \right) \\
&\cdot \frac{\tilde{M}^{(\iota)-1}}{f_Z(z)} (1 + \mathrm{o_p}(1)) \mathrm{o_p}(1) \\
=& \mathrm{o_p}(1).
\end{aligned}$$

□

Proof of Theorem 2.8.

For any $z = s_q + \tau h_n$ with $\tau \in (-1, 1)$, let

$$\acute{\tilde{S}}_n^{(\iota)} = \frac{1}{n} \sum_{i:Z_i < s_q} \rho\left(\frac{Z_i - z}{h_n}\right) \rho^\top \left(\frac{Z_i - z}{h_n}\right) \tilde{K}_h^{(\iota)}(Z_i - z), \qquad (2.54)$$

$$\grave{\tilde{S}}_n^{(\iota)} = \frac{1}{n} \sum_{i:Z_i \geq s_q} \rho\left(\frac{Z_i - z}{h_n}\right) \rho^\top \left(\frac{Z_i - z}{h_n}\right) \tilde{K}_h^{(\iota)}(Z_i - z), \qquad (2.55)$$

$$\acute{\tilde{W}}_{n,1}^{(\iota)} = \frac{1}{n} \sum_{i:Z_i < s_q} \varepsilon_i^2 \tilde{K}_h^{(\iota)}(Z_i - z), \quad \grave{\tilde{W}}_{n,1}^{(\iota)} = \frac{1}{n} \sum_{i:Z_i \geq s_q} \varepsilon_i^2 \tilde{K}_h^{(\iota)}(Z_i - z), \qquad (2.56)$$

$$\acute{\tilde{W}}_{n,2}^{(\iota)} = \frac{1}{n} \sum_{i:Z_i < s_q} X_i X_i^\top \tilde{K}_h^{(\iota)}(Z_i - z), \quad \grave{\tilde{W}}_{n,2}^{(\iota)} = \frac{1}{n} \sum_{i:Z_i \geq s_q} X_i X_i^\top \tilde{K}_h^{(\iota)}(Z_i - z), \qquad (2.57)$$

$$\acute{\tilde{W}}_{n,3}^{(\iota)} = \frac{1}{n}\sum_{i:Z_i<s_q} \rho\left(\frac{Z_i - z}{h_n}\right) X_i^\top \tilde{K}_h^{(\iota)}(Z_i - z), \tag{2.58}$$

$$\grave{\tilde{W}}_{n,3}^{(\iota)} = \frac{1}{n}\sum_{i:Z_i\geq s_q} \rho\left(\frac{Z_i - z}{h_n}\right) X_i^\top \tilde{K}_h^{(\iota)}(Z_i - z), \tag{2.59}$$

$$\acute{\tilde{W}}_{n,4}^{(\iota)} = \frac{1}{n}\sum_{i:Z_i<s_q} X_i\varepsilon_i \tilde{K}_h^{(\iota)}(Z_i - z), \quad \grave{\tilde{W}}_{n,4}^{(\iota)} = \frac{1}{n}\sum_{i:Z_i\geq s_q} X_i\varepsilon_i \tilde{K}_h^{(\iota)}(Z_i - z), \tag{2.60}$$

$$\acute{\tilde{W}}_{n,5}^{(\iota)} = \frac{1}{n}\sum_{i:Z_i<s_q} \rho\left(\frac{Z_i - z}{h_n}\right) \varepsilon_i \tilde{K}_h^{(\iota)}(Z_i - z), \quad \text{and} \tag{2.61}$$

$$\grave{\tilde{W}}_{n,5}^{(\iota)} = \frac{1}{n}\sum_{i:Z_i\geq s_q} \rho\left(\frac{Z_i - z}{h_n}\right) \varepsilon_i \tilde{K}_h^{(\iota)}(Z_i - z). \tag{2.62}$$

Further, we define the population counterparts of the above kernel weighted averages:

$$\mho_-(s_q) = \lim_{z\uparrow s_q} \mathrm{E}[X^\top|Z = z], \quad \mho_+(s_q) = \lim_{z\downarrow s_q} \mathrm{E}[X^\top|Z = z] \tag{2.63}$$

$$\acute{\tilde{\mu}}_{0,\tau}^{(\iota)} = \int_{-1}^{-\tau} \tilde{K}(u)\mathrm{d}u, \quad \grave{\tilde{\mu}}_{0,\tau}^{(\iota)} = \int_{-\tau}^{1} \tilde{K}(u)\mathrm{d}u, \tag{2.64}$$

$$\acute{m}_\tau^{(\iota)} = \int_{-1}^{-\tau} \rho(u)\tilde{K}(u)\mathrm{d}u, \quad \grave{m}_\tau^{(\iota)} = \int_{-\tau}^{1} \rho(u)\tilde{K}(u)\mathrm{d}u, \tag{2.65}$$

$$\acute{M}_\tau^{(\iota)} = \int_{-1}^{-\tau} \rho(u)\rho^\top(u)\tilde{K}(u)\mathrm{d}u, \quad \text{and} \quad \grave{M}_\tau^{(\iota)} = \int_{-\tau}^{1} \rho(u)\rho^\top(u)\tilde{K}(u)\mathrm{d}u. \tag{2.66}$$

Again, we use decomposition $\tilde{T}_n^{(\iota)} = \tilde{W}_{n,5}^{(\iota)} + \tilde{T}_{n,2}^{(\iota)}$ as in (2.51). By the consistency results for $\acute{\tilde{W}}_{n,5}^{(\iota)}$ and $\grave{\tilde{W}}_{n,5}^{(\iota)}$ in Lemma 2.16(vi),

$$\tilde{W}_{n,5}^{(\iota)} = \acute{\tilde{W}}_{n,5}^{(\iota)} + \grave{\tilde{W}}_{n,5}^{(\iota)} = \mathrm{o_p}(1) + \mathrm{o_p}(1).$$

By (2.33)–(2.35) and the consistency results for $\acute{\tilde{W}}_{n,4}^{(\iota)}$ and $\grave{\tilde{W}}_{n,4}^{(\iota)}$ in Lemma 2.16(v), we obtain

$$\begin{aligned}
\tilde{T}_{n,2}^{(\iota)}(z) =& \frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{Z_i-z}{h_n}\right)X_i^\top\{a(Z_i)-\hat{a}_n^{(\iota)}(z)\}\tilde{K}_h^{(\iota)}(Z_i-z)\\
=& \frac{1}{n}\sum_{i:Z_i<s_q}\rho\left(\frac{Z_i-z}{h_n}\right)X_i^\top[\underbrace{a(Z_i)-a_-(s_q)}_{\mathrm{O}(h_n)}-\Xi_{0,\tau}^{(\iota)}d_q]\tilde{K}_h^{(\iota)}(Z_i-z)\\
&+\frac{1}{n}\sum_{i:Z_i\geq s_q}\rho\left(\frac{Z_i-z}{h_n}\right)X_i^\top[\underbrace{a(Z_i)-a_-(s_q)}_{d_q+\mathrm{O}(h_n)}-\Xi_{0,\tau}^{(\iota)}d_q]\tilde{K}_h^{(\iota)}(Z_i-z)\\
=& -\acute{\tilde{W}}_{n,4}^{(\iota)}\left(\Xi_{0,\tau}^{(\iota)}d_q+\mathrm{O}(h_n)\right)-\grave{\tilde{W}}_{n,4}^{(\iota)}\left((\Xi_{0,\tau}^{(\iota)}-I_p)d_q+\mathrm{O}(h_n)\right)\\
=& -f_Z(s_q)\left(\acute{m}_\tau^{(\iota)}\mho_-(s_q)\Xi_{0,\tau}^{(\iota)}+\grave{m}_\tau^{(\iota)}\mho_+(s_q)(\Xi_{0,\tau}^{(\iota)}-I_p)\right)d_q+\mathrm{o_p}(1).
\end{aligned}$$

Hence, it follows from the consistency results for $\tilde{S}_n^{(\iota)}=\acute{\tilde{S}}_n^{(\iota)}+\grave{\tilde{S}}_n^{(\iota)}$ in Lemma 2.16(i) that

$$\hat{\gamma}_n^{(\iota)}=\gamma_{q,\tau}^{(\iota)}+\mathrm{o_p}(1), \tag{2.67}$$

where

$$\begin{aligned}
\gamma_{q,\tau}^{(\iota)} =& -\left(\acute{M}_\tau^{(\iota)}+\grave{M}_\tau^{(\iota)}\right)^{-1}\left(\acute{m}_\tau^{(\iota)}\mho_-(s_q)\Xi_{0,\tau}^{(\iota)}+\grave{m}_\tau^{(\iota)}\mho_+(s_q)(\Xi_{0,\tau}^{(\iota)}-I_p)\right)d_q\\
=& -\tilde{M}^{(\iota)-1}\left(\acute{m}_\tau^{(\iota)}\mho_-(s_q)\Xi_{0,\tau}^{(\iota)}+\grave{m}_\tau^{(\iota)}\mho_+(s_q)(\Xi_{0,\tau}^{(\iota)}-I_p)\right)d_q.
\end{aligned} \tag{2.68}$$

Next, for $Z_i<s_q$ and $|Z_i-z|\leq h_n$, the squared error $\hat{e}_{n,i}^{(\iota)2}$ equals by the Taylor expansion of $a(\cdot)$ and the boundedness of its derivatives (Assumption 2.A5)

$$\begin{aligned}
\hat{e}_{n,i}^{(\iota)2} =& \varepsilon_i^2+\left\{\hat{\gamma}_n^{(\iota)\top}\rho\left(\frac{Z_i-z}{h_n}\right)\right\}^2+\left\{X_i^\top[a_-(s_q)+\mathrm{O}(h_n)-\hat{a}_n^{(\iota)}(z)]\right\}^2\\
&-2\hat{\gamma}_n^{(\iota)\top}\rho\left(\frac{Z_i-z}{h_n}\right)X_i^\top[a_-(z)+\mathrm{O}(h_n)-\hat{a}_n^{(\iota)}(z)]\\
&+2\varepsilon_i X_i^\top\left\{a_-(s_q)+\mathrm{O}(h_n)-\hat{a}_n^{(\iota)}(z)\right\}-2\hat{\gamma}_n^{(\iota)\top}\rho\left(\frac{Z_i-z}{h_n}\right)\varepsilon_i
\end{aligned}$$

uniformly in $i \in \mathbb{N}$, and by the same argument for $Z_i \geq s_q$ and $|Z_i - z| \leq h_n$,

$$\begin{aligned}\hat{e}_{n,i}^{(\iota)2} =& \varepsilon_i^2 + \left\{\hat{\gamma}_n^{(\iota)\top}\rho\left(\frac{Z_i - z}{h_n}\right)\right\}^2 + \left\{X_i^\top[a_-(s_q) + d_q + \mathrm{O}(h_n) - \hat{a}_n^{(\iota)}(z)]\right\}^2 \\ & - 2\hat{\gamma}_n^{(\iota)\top}\rho\left(\frac{Z_i - z}{h_n}\right)X_i^\top[a_-(s_q) + d_q + \mathrm{O}(h_n) - \hat{a}_n^{(\iota)}(z)] \\ & + 2\varepsilon_i X_i^\top\left\{a_-(s_q) + d_q + \mathrm{O}(h_n) - \hat{a}_n^{(\iota)}(z)\right\} - 2\hat{\gamma}_n^{(\iota)\top}\rho\left(\frac{Z_i - z}{h_n}\right)\varepsilon_i\end{aligned}$$

uniformly in $i \in \mathbb{N}$. For the term $\tilde{N}_n^{(\iota)}$ of $\tilde{\Psi}_n^{(\iota)}(z)$ in (2.11), it now follows after substituting the above expressions for $\hat{e}_{n,i}^{(\iota)2}$ and using equations (2.33)–(2.35) that

$$\begin{aligned}\tilde{N}_n^{(\iota)} =& \frac{1}{n}\sum_{i:Z_i<s_q}\hat{e}_{n,i}^{(\iota)2}\tilde{K}_h^{(\iota)}(Z_i - z) + \frac{1}{n}\sum_{i:Z_i\geq s_q}\hat{e}_{n,i}^{(\iota)2}\tilde{K}_h^{(\iota)}(Z_i - z) \\ =& \acute{\tilde{W}}_{n,1}^{(\iota)} + \grave{\tilde{W}}_{n,1}^{(\iota)} + \Big[\hat{\gamma}_n^{(\iota)\top}\tilde{S}_n^{(\iota)}\hat{\gamma}_n^{(\iota)} + d_q^\top\Xi_{0,\tau}^{(\iota)\top}\acute{\tilde{W}}_{n,2}^{(\iota)}\Xi_{0,\tau}^{(\iota)}d_q \\ & + d_q^\top(\Xi_{0,\tau}^{(\iota)} - I_p)^\top\grave{\tilde{W}}_{n,2}^{(\iota)}(\Xi_{0,\tau}^{(\iota)} - I_p)d_q + 2\hat{\gamma}_n^{(\iota)\top}\acute{\tilde{W}}_{n,3}\Xi_{0,\tau}^{(\iota)}d_q \\ & + 2\hat{\gamma}_n^{(\iota)\top}\grave{\tilde{W}}_{n,3}(\Xi_{0,\tau}^{(\iota)} - I_p)d_q - 2d_q^\top\Xi_{0,\tau}^{(\iota)\top}\acute{\tilde{W}}_{n,4}^{(\iota)} - 2d_q^\top(\Xi_{0,\tau}^{(\iota)} - I_p)^\top\grave{\tilde{W}}_{n,4}^{(\iota)}\Big] \\ & \cdot(1 + \mathrm{O}(h_n)) - 2\hat{\gamma}_n^{(\iota)\top}\acute{\tilde{W}}_{n,5} - 2\hat{\gamma}_n^{(\iota)\top}\grave{\tilde{W}}_{n,5}.\end{aligned}$$

By (2.67) and Lemma 2.16, we thus have

$$\tilde{N}_n^{(\iota)} = f_Z(s_q)\left\{\acute{\tilde{\mu}}_{0,\tau}^{(\iota)}\sigma_-^2(s_q) + \grave{\tilde{\mu}}_{0,\tau}^{(\iota)}\sigma_+^2(s_q) + \sigma_{e,\tau}^{(\iota)2}(s_q)\right\} + \mathrm{o_p}(1), \tag{2.69}$$

where

$$\begin{aligned}\sigma_{e,\tau}^{(\iota)2}(s_q) =& \gamma_{q,\tau}^{(\iota)\top}\tilde{M}^{(\iota)}\gamma_{q,\tau}^{(\iota)} + d_q^\top\Xi_{0,\tau}^{(\iota)\top}\acute{\tilde{\mu}}_{0,\tau}^{(\iota)}\Omega_-(s_q)\Xi_{0,\tau}^{(\iota)}d_q \\ & + d_q^\top(\Xi_{0,\tau}^{(\iota)} - I_p)^\top\grave{\tilde{\mu}}_{0,\tau}^{(\iota)}\Omega_+(s_q)(\Xi_{0,\tau}^{(\iota)} - I_p)d_q \\ & + 2\gamma_{q,\tau}^{(\iota)\top}\acute{m}_\tau^{(\iota)}\mho_-(s_q)\Xi_{0,\tau}d_q + 2\gamma_{q,\tau}^{(\iota)\top}\grave{m}_\tau^{(\iota)}\mho_+(s_q)(\Xi_{0,\tau}^{(\iota)} - I_p)d_q.\end{aligned}$$

Since the term above can be rewritten as

$$\begin{aligned}\sigma_{e,\tau}^{(\iota)2}(s_q) =& \int_{-1}^{-\tau}\int(x^\top\Xi_{0,\tau}d_q + \gamma_{q,\tau}^{(\iota)\top}\rho(u))^2\tilde{K}(u)\frac{f(x,s_q)}{f_Z(s_q)}\mathrm{d}x\mathrm{d}u \\ & + \int_{-\tau}^{1}\int(x^\top[\Xi_{0,\tau} - I_p]d_q + \gamma_{q,\tau}^{(\iota)\top}\rho(u))^2\tilde{K}(u)\frac{f(x,s_q)}{f_Z(s_q)}\mathrm{d}x\mathrm{d}u,\end{aligned} \tag{2.70}$$

it is clearly non-negative.

By equations (2.67)–(2.69) and Lemma 2.16(i), we conclude that

$$\tilde{\Psi}_n^{(\iota)}(z) = \gamma_{q,\tau}^{(\iota)\top} \tilde{C}_\tau^{(\iota)} \gamma_{q,\tau}^{(\iota)} + \mathrm{o_p}(1),$$

where

$$\tilde{C}_\tau^{(\iota)} = \left(\frac{\tilde{M}^{(\iota)}}{\acute{\tilde{\mu}}_{0,\tau}^{(\iota)} \sigma_-^2(s_q) + \grave{\tilde{\mu}}_{0,\tau}^{(\iota)} \sigma_+^2(s_q) + \sigma_{e,\tau}^{(\iota)2}(s_q)} \right). \tag{2.71}$$

By the positive definiteness of $\tilde{M}^{(\iota)}$ (Assumption 2.D1) and non-negative $\sigma_{e,\tau}^{(\iota)2}(s_q)$ from (2.70), we claim that $\tilde{C}_\tau^{(\iota)} \succ 0$ for any $\tau \in (-1, 1)$ and $\iota = c, l, r$. According to Assumption 2.D2, some elements of $\gamma_{q,\tau}^{(\iota)}$, $\iota = c, l$, are non-zero for $\tau \in (0, 1)$. Hence, the limits of $\tilde{\Psi}_n^{(c)}(z)$ and $\tilde{\Psi}_n^{(l)}(z)$ are strictly positive, i.e., $\gamma_{q,\tau}^{(\iota)\top} \tilde{C}_\tau^{(\iota)} \gamma_{q,\tau}^{(\iota)} > 0$ for $\tau \in (0, 1)$ and $\iota = c, l$. For $\tau \in (0, 1)$ and $\iota = r$, we have $\acute{\tilde{\mu}}_{0,\tau}^{(\iota)} = 0$ and $\acute{m}_\tau^{(r)} = 0$. By the expressions of $\gamma_{q,\tau}^{(\iota)}$ in (2.68) and the fact (2.38), $\Xi_{0,\tau}^{(r)} = I_p$ for $\tau \in (0, 1)$, we conclude that $\gamma_{q,\tau}^{(r)} = 0$ and hence $\gamma_{q,\tau}^{(r)\top} \tilde{C}_\tau^{(r)} \gamma_{q,\tau}^{(r)} = 0$. Similarly for $\tau \in (-1, 0)$, we have $\gamma_{q,\tau}^{(c)\top} \tilde{C}_\tau^{(c)} \gamma_{q,\tau}^{(c)} \succ 0$, $\gamma_{q,\tau}^{(r)\top} \tilde{C}_\tau^{(r)} \gamma_{q,\tau}^{(r)} \succ 0$, and $\gamma_{q,\tau}^{(l)\top} \tilde{C}_\tau^{(l)} \gamma_{q,\tau}^{(l)} = 0$ due to equation (2.39), $\Xi_{0,\tau}^{(l)} = 0_p$. □

Proof of Theorem 2.9.

Being based on the results of Theorems 2.7 and 2.8, it follows the same steps as in the proof of Theorem 2.5. □

Proof of Theorem 2.10.

Being based on the results of Theorems 2.7 and 2.8, it follows the same steps as in the proof of Theorem 2.6. □

2.9 Appendix: Some auxiliary lemmas

Lemma 2.11. *Suppose Assumptions 2.A and 2.B hold. For any* $z \in D_{1n}^{(\iota)}$ *and* $\iota = c, l, r$, *it holds as* $n \to +\infty$ *that*

(i) $S_{n,j}^{(\iota)} = \mu_j^{(\iota)} f_Z(z)\Omega(z) + \mathrm{o_p}(1)$ *with* $j = 0, 1, 2, 3$,

(ii) $S_n^{(\iota)-1} = \dfrac{f_Z^{-1}(z)}{\mu_0^{(\iota)}\mu_2^{(\iota)} - \mu_1^{(\iota)2}} \begin{pmatrix} \mu_2^{(\iota)} & -\mu_1^{(\iota)} \\ -\mu_1^{(\iota)} & \mu_0^{(\iota)} \end{pmatrix} \otimes \Omega^{-1}(z)(1 + \mathrm{o_p}(1))$,

(iii) $F_{n,j}^{(\iota)} = \mathrm{o_p}(1)$ *with* $j = 0, 1$,

(iv) $K_n^{(\iota)} = \mu_0^{(\iota)} f_Z(z) + \mathrm{o_p}(1)$,

(v) $N_{n,1}^{(\iota)} = \mu_0^{(\iota)} f_Z(z)\sigma^2(z) + \mathrm{o_p}(1)$,

(vi) $\hat{a}_n^{(\iota)}(z) = a(z) + \mathrm{o_p}(1)$,

(vii) $\hat{b}_n^{(\iota)}(z) = a'(z) + \mathrm{o_p}(h_n^{-1})$,

where the above objects are defined in (2.15)–(2.17), (2.25), *and* (2.26).

Proof. By Assumptions 2.A1–2.A3 and 2.B1–2.B2, the conditions for the weak law of large number for kernel estimators in Hansen (2008) are satisfied. Applying Theorem 1 in Hansen (2008) leads to

$$S_{n,j}^{(\iota)} = \mathrm{E}[S_{n,j}^{(\iota)}] + \mathrm{o_p}(1).$$

After a change of variable ($\dot{z} = z + vh_n$) and the Taylor expansion of the density f in which it partial derivatives with respect to Z are uniformly bounded due to Assumption 2.A2, the expectation of $S_{n,j}^{(\iota)}$ equals

$$\begin{aligned}
\mathrm{E}[S_{n,j}^{(\iota)}] =& \mathrm{E}\left[X_i X_i^\top \left(\frac{Z_i - z}{h_n}\right)^j K_h^{(\iota)}(Z_i - z)\right] \\
=& \frac{1}{h_n} \int\int \dot{x}\dot{x}^\top \left(\frac{\dot{z} - z}{h_n}\right)^j K^{(\iota)}\left(\frac{\dot{z} - z}{h_n}\right) f(\dot{x}, \dot{z}) \mathrm{d}\dot{z}\mathrm{d}\dot{x} \\
=& \int\int \dot{x}\dot{x}^\top v^j K^{(\iota)}(v) f(\dot{x}, z + vh_n) \mathrm{d}\dot{x}\mathrm{d}v \\
=& \int v^j K^{(\iota)}(v)\mathrm{d}v \cdot f_Z(z) \cdot \int \dot{x}\dot{x}^\top \frac{f(\dot{x}, z)}{f_Z(z)} \mathrm{d}\dot{x}\mathrm{d}v + \mathrm{O}(h_n) \\
=& \mu_j^{(\iota)} f_Z(z)\Omega(z) + \mathrm{O}(h_n),
\end{aligned}$$

where $\Omega(z) = \mathrm{E}(XX^\top | Z = z)$. This concludes part (i). Part (ii) the follows trivially by part (i), Lemma 2.17(i): $\mu_0^{(\iota)}\mu_2^{(\iota)} - \mu_1^{(\iota)2} \neq 0$, the full rank conditions for $\Omega(z)$ in

Assumption 2.A4, and $f_Z(z) > 0$ in Assumption 2.A2. Similarly to part (i), one can easily show (iii)–(v). Finally, using (2.21), parts (i)-(iii), and Assumption 2.A5, we have

$$\begin{aligned}
\left\|H_n(\hat{\beta}_n^{(\iota)} - \beta)\right\| \leq & \left\|S_n^{(\iota)-1} F_n^{(\iota)}\right\| + \left\|\frac{h_n^2}{2} S_n^{(\iota)-1} \begin{pmatrix} S_{n,2}^{(\iota)} \\ S_{n,3}^{(\iota)} \end{pmatrix} a''(z)\right\| + \mathrm{o}(h_n^2) \\
= & \left\|\frac{\Omega^{-1}(z)}{f_Z(z)} \left(c_0^{(\iota)} F_{n,0}^{(\iota)} + c_1^{(\iota)} F_{n,1}^{(\iota)}\right)(1 + \mathrm{o_p}(1))\right\| \\
& + \left\|\frac{h_n^2}{2} \left(c_0^{(\iota)} \mu_2^{(\iota)} + c_1^{(\iota)} \mu_3^{(\iota)}\right) a''(z)(1 + \mathrm{o_p}(1))\right\| + \mathrm{o}(h_n^2) \\
\leq & \mathrm{o_p}(1) + \mathrm{O_p}(h_n^2) \left\|a''(z)\right\| + \mathrm{o}(h_n^2) \\
= & \mathrm{o_p}(1),
\end{aligned}$$

where $c_0^{(\iota)}$ and $c_1^{(\iota)}$ are defined in (2.7). This completes the proofs of (vi) and (vii). □

Lemma 2.12. *Under Assumptions 2.A and 2.B, it holds as $n \to +\infty$ that*

(i) $h_n \mathrm{var}(W_1^{(\iota)}) \to f_Z(z)\Theta(z) \left[c_0^{(\iota)2} \nu_0^{(\iota)} + 2c_0^{(\iota)} c_1^{(\iota)} \nu_1^{(\iota)} + c_1^{(\iota)2} \nu_2^{(\iota)}\right]$,

(ii) $h_n \sum_{j=1}^{n-1} |\mathrm{cov}(W_1^{(\iota)}, W_{j+1}^{(\iota)})| = \mathrm{o}(1)$, *and*

(iii) $nh_n \mathrm{var}(U_n^{(\iota)}) \to f_Z(z)\Theta(z) \left[c_0^{(\iota)2} \nu_0^{(\iota)} + 2c_0^{(\iota)} c_1^{(\iota)} \nu_1^{(\iota)} + c_1^{(\iota)2} \nu_2^{(\iota)}\right]$,

where $U_n^{(\iota)}$ and $W_i^{(\iota)}$ are given in (2.23)–(2.24), $\Theta(z) = \mathrm{E}(XX^\top \sigma^2(X,Z)|Z = z)$, and $c_j^{(\iota)}$ and $\nu_j^{(\iota)}$ are defined in equation (2.7).

Under Assumptions 2.A, A6', and 2.B, the limits in points (i) and (iii) are equal to

$$\begin{aligned}
& f_Z(z)\Theta_-(z) \left[c_0^{(\iota)2} \nu_0^{(l)} + 2c_0^{(\iota)} c_1^{(\iota)} \nu_1^{(l)} + c_1^{(\iota)2} \nu_2^{(l)}\right] \mathbf{1}(\iota \in \{c, l\}) \\
& + f_Z(z)\Theta_+(z) \left[c_0^{(\iota)2} \nu_0^{(r)} + 2c_0^{(\iota)} c_1^{(\iota)} \nu_1^{(r)} + c_1^{(\iota)2} \nu_2^{(r)}\right] \mathbf{1}(\iota \in \{c, r\}),
\end{aligned}$$

where $\Theta_-(z) = \mathrm{E}(XX^\top \sigma_-^2(X,Z)|Z = z)$ and $\Theta_+(z) = \mathrm{E}(XX^\top \sigma_+^2(X,Z)|Z = z)$.

Proof. By conditioning on (X_1, Z_1), a change of variables, and the Taylor expansion,

$$\begin{aligned} h_n \mathrm{var}(W_1^{(\iota)}) =& h_n \mathrm{E}\left[X_1 X_1^\top \sigma^2(X_1, Z_1)\left\{c_0^{(\iota)} + c_1^{(\iota)}\left(\frac{Z_1 - z}{h_n}\right)\right\}^2 K_h^{(\iota)2}(Z_1 - z)\right] \\ =& \int\int xx^\top \sigma^2(x, z + h_n u)\left(c_0^{(\iota)2} + 2c_0^{(\iota)}c_1^{(\iota)}u + c_1^{(\iota)2}u^2\right) K^{(\iota)2}(u) \\ & \times f(x, z + h_n u)\mathrm{d}u\mathrm{d}x \\ =& f_Z(z)\Theta(z)\left[c_0^{(\iota)2}\nu_0^{(\iota)} + 2c_0^{(\iota)}c_1^{(\iota)}\nu_1^{(\iota)} + c_1^{(\iota)2}\nu_2^{(\iota)}\right] + \mathrm{O}(h_n) \end{aligned}$$

due to Assumptions 2.A2, 2.A5, and 2.A6. Since part (iii) follows trivially from (i) and (ii) by

$$\begin{aligned} nh_n \mathrm{var}(U_n^{(\iota)}) =& \frac{h_n}{n}\mathrm{var}\left(\sum_{i=1}^n W_i^{(\iota)}\right) \\ =& h_n \mathrm{var}(W_1^{(\iota)}) + 2h_n \sum_{j=1}^{n-1}\left(1 - \frac{j}{n}\right)\mathrm{cov}(W_1^{(\iota)}, W_{j+1}^{(\iota)}), \end{aligned}$$

it remains to prove (ii). To this end, let $c_n \to \infty$ be a sequence of positive integers such that $c_n h_n \to 0$. We write

$$\begin{aligned} h_n \sum_{j=1}^{n-1}\left|\mathrm{cov}(W_1^{(\iota)}, W_{j+1}^{(\iota)})\right| =& h_n \sum_{j=1}^{c_n}\left|\mathrm{cov}(W_1^{(\iota)}, W_{j+1}^{(\iota)})\right| + h_n \sum_{j=c_n}^{n-1}\left|\mathrm{cov}(W_1^{(\iota)}, W_{j+1}^{(\iota)})\right| \\ =& J_{1,n} + J_{2,n}. \end{aligned}$$

We complete the proof by showing that $J_{1,n} = \mathrm{o}(1)$ and $J_{2,n} = \mathrm{o}(1)$.

First, for $j \le c_n$, by conditioning on Z_1 and Z_{j+1} and Assumption 2.A3(iii), we have,

$$\begin{aligned} |\mathrm{cov}(W_1^{(\iota)}, W_{j+1}^{(\iota)})| \le& C_1 \mathrm{E}\left(|X_1 X_{j+1}^\top \varepsilon_1 \varepsilon_{j+1}| K_h^{(\iota)}(Z_1 - z) K_h^{(\iota)}(Z_{j+1} - z)\right) \\ \le& C_2 \mathrm{E}\left(|X_1 X_{j+1}^\top \varepsilon_1 \varepsilon_{j+1}| \big| Z_1 = z, Z_{j+1} = z\right)\left(f_{Z_1 Z_{j+1}}(z, z) + \mathrm{O}(h_n)\right) \\ \le& C_3, \end{aligned}$$

for positive constants C_1, C_2, C_3, which implies that $J_{1,n} \le h_n c_n C = \mathrm{o}(1)$ by the choice of c_n. Next, let $W_{j,m}^{(\iota)}$ be the m-th element of $W_j^{(\iota)}$. Using Davydov's inequality (Fan and

Yao, 2003, Proposition 2.5 with $p = q = \delta$), one has

$$|\mathrm{cov}(W_{1,l}^{(\iota)}, W_{j+1,m}^{(\iota)})| \leq C\alpha^{1-2/\delta}(j) \left(\mathrm{E}|W_{1,l}^{(\iota)}|^{\delta}\right)^{1/\delta} \left(\mathrm{E}|W_{j+1,m}^{(\iota)}|^{\delta}\right)^{1/\delta}. \tag{2.72}$$

By conditioning on Z_1 and Assumptions 2.A2 and 2.A3(ii),

$$\begin{aligned} \mathrm{E}|W_{1,l}^{(\iota)}|^{\delta} &\leq C_1 \mathrm{E}[|X_{1,l}\varepsilon_1|^{\delta} K_h^{(\iota)\delta}(Z_1 - z)] \\ &\leq C_2 h_n^{1-\delta} \{\mathrm{E}[|X_{1,l}\varepsilon_1|^{\delta} \big| Z_1 = z](f_Z(z) + \mathrm{O}(h_n))\} \\ &\leq C_3 h_n^{1-\delta} \end{aligned} \tag{2.73}$$

for some $C_1, C_2, C_3 > 0$. It follows from equations (2.72), (2.73), and Assumption 2.A1 that

$$\begin{aligned} J_{2,n} &= h_n \sum_{j=c_n+1}^{n-1} |\mathrm{cov}(W_1^{(\iota)}, W_{j+1}^{(\iota)})| \\ &\leq C_1 h_n h_n^{2(1-\delta)/\delta} \sum_{j=c_n+1}^{\infty} \alpha^{1-2/\delta}(j) \\ &\leq C_2 h_n^{2/\delta - 1} \sum_{j=c_n+1}^{\infty} j^{-(2-2/\delta)} \\ &\leq C_3 h_n^{2/\delta-1} c_n^{2/\delta-1} = \mathrm{o}(1), \end{aligned}$$

where constants $C_1, C_2, C_3 > 0$ and the last inequality follows from the fact that

$$\sum_{j=k+1}^{\infty} j^{-\tau} \leq \int_k^{\infty} x^{-\tau} \mathrm{d}x = \frac{k^{1-\tau}}{\tau - 1}.$$

Finally, under Assumptions 2.A2, 2.A5, and A6', conditioning on (X_1, Z_1), a change of variables, and the Taylor expansion lead as in the introduction of this proof to

$$\begin{aligned}
h_n \mathrm{var}(W_1^{(\iota)}) =& h_n \mathrm{E}\left[X_1 X_1^\top \sigma^2(X_1, Z_1) \left\{c_0^{(\iota)} + c_1^{(\iota)} \left(\frac{Z_1 - z}{h_n}\right)\right\}^2 K_h^{(\iota)2}(Z_1 - z)\right] \\
=& \int\int xx^\top \sigma^2(x, z + h_n u) \left(c_0^{(\iota)2} + 2c_0^{(\iota)} c_1^{(\iota)} u + c_1^{(\iota)2} u^2\right) K^{(\iota)2}(u) \\
& \times f(x, z + h_n u) \mathrm{d}u \mathrm{d}x \\
=& f_Z(z) \Theta_-(z) \left[c_0^{(\iota)2} \nu_0^{(l)} + 2c_0^{(\iota)} c_1^{(\iota)} \nu_1^{(l)} + c_1^{(\iota)2} \nu_2^{(l)}\right] \mathbf{1}(\iota \in \{c, l\}) + \mathrm{O}(h_n) \\
+& f_Z(z) \Theta_+(z) \left[c_0^{(\iota)2} \nu_0^{(r)} + 2c_0^{(\iota)} c_1^{(\iota)} \nu_1^{(r)} + c_1^{(\iota)2} \nu_2^{(r)}\right] \mathbf{1}(\iota \in \{c, r\}) + \mathrm{O}(h_n).
\end{aligned}$$

□

Lemma 2.13. *Under Assumptions 2.A, 2.B, and 2.C, we have for $n \to +\infty$ and $\iota = c, l, r$,*

$$\begin{aligned}
&\sup_{z \in D_{1n}^{(\iota)}} \left\| S_{n,j}^{(\iota)} - \mu_j^{(\iota)} f_Z(z) \Omega(z) \right\| = \mathrm{O_p}\left(\sqrt{\frac{\ln n}{n h_n}}\right) + \mathrm{O}(h_n) \text{ for } j = 0, 1, 2, 3, \\
&\sup_{z \in D_{1n}^{(\iota)}} \left\| F_{n,j}^{(\iota)} \right\| = \mathrm{O_p}\left(\sqrt{\frac{\ln n}{n h_n}}\right) \text{ for } j = 0, 1,
\end{aligned}$$

and

$$f_Z(z) S_n^{(\iota)-1} = \frac{\begin{pmatrix} \mu_2^{(\iota)} \Omega^{-1}(z) & -\mu_1^{(\iota)} \Omega^{-1}(z) \\ -\mu_1^{(\iota)} \Omega^{-1}(z) & \mu_0^{(\iota)} \Omega^{-1}(z) \end{pmatrix}}{\mu_0^{(\iota)} \mu_2^{(\iota)} - \mu_1^{(\iota)2}} \left\{1 + \mathrm{O_p}\left(\sqrt{\frac{\ln n}{n h_n}}\right) + \mathrm{O}(h_n)\right\}$$

uniformly for $z \in D_{1n}^{(\iota)}$.

Proof. By Assumptions 2.A, 2.B, and 2.C, the conditions for weak uniform convergence result for kernel estimators over expanding sets in Hansen (2008) are satisfied. First, we consider case $\iota = c$, which uses both left and right neighborhoods. For the continuous region $D_{1n}^{(c)} = \bigcup_{q=0}^{Q} (s_q + h_n, s_{q+1} - h_n)$, we apply Theorem 2 in Hansen (2008) on each

subregion $(s_q + h_n, s_{q+1} - h_n)$:

$$\sup_{z\in(s_q+h_n,s_{q+1}-h_n)} \left\| S_{n,j}^{(c)} - \mathrm{E}(S_{n,j}^{(c)}) \right\| = \mathrm{O_p}\left(\sqrt{\frac{\ln n}{nh_n}}\right).$$

Notice the expanding sets considered in Hansen (2008) are allowed to grow to infinity slowly, as $n \to \infty$, while the subregion $(s_q + h_n, s_{q+1} - h_n)$ expands to a bounded set (s_q, s_{q+1}). Taking the maximum over all subregions yields

$$\begin{aligned}\sup_{z\in D_{1n}^{(c)}} \left\| S_{n,j}^{(c)} - \mathrm{E}(S_{n,j}^{(c)}) \right\| \leq & (Q+1)\cdot \max_q \sup_{z\in(s_q+h_n,s_{q+1}-h_n)} \left\| S_{n,j}^{(c)} - \mathrm{E}(S_{n,j}^{(c)}) \right\| \\ = & \mathrm{O_p}\left(\sqrt{\frac{\ln n}{nh_n}}\right).\end{aligned}$$

Since $\mathrm{E}(S_{n,j}^{(c)}) = \mu_j^{(c)} f_Z(z)\Omega(z) + \mathrm{O}(h_n)$, which is shown in the proof in Lemma 2.11, we have

$$\sup_{z\in D_{1n}^{(c)}} \left\| S_{n,j}^{(c)} - \mu_j^{(c)} f_Z(z)\Omega(z) \right\| = \mathrm{O_p}\left(\sqrt{\frac{\ln n}{nh_n}}\right) + \mathrm{O}(h_n).$$

Although Theorem 2 in Hansen (2008) originally excludes the case of one-sided kernel, his theorem is still applicable for one-sided kernel by taking 'one-sided' covering sets A_j, which boosts the size of covering by a constant multiplier 2^p, instead of 'two-sided' A_j in his proof. Then, by similar argument as for $S_{n,j}^{(c)}$, one can prove the uniform consistency results for $S_{n,j}^{(l)}$ and $S_{n,j}^{(r)}$.

Analogously, we can apply Theorem 2 in Hansen (2008) to $F_{n,j}^{(\iota)}$ with $\iota = c, l, r$, where the uniform convergence rates stays equal to $\mathrm{O_p}(\sqrt{\ln n/(nh_n)})$ since $\mathrm{E}(F_{n,j}^{(\iota)}) = 0$. □

Lemma 2.14. *Suppose Assumptions 2.A and 2.B hold. For any $z = s_q + \tau h_n$ with $\tau \in (-1, 1)$ and $\iota = c, l, r$, we have as $n \to +\infty$,*

(i) $\acute{S}_{n,j}^{(\iota)} = f_Z(s_q)\Omega_-(s_q)\acute{\mu}_{j,\tau}^{(\iota)} + \mathrm{o_p}(1)$ *and* $\grave{S}_{n,j}^{(\iota)} = f_Z(s_q)\Omega_+(s_q)\grave{\mu}_{j,\tau}^{(\iota)} + \mathrm{o_p}(1)$ *for* $j = 0, 1, 2$;

(ii) $\acute{F}_{n,j}^{(\iota)} = \grave{F}_{n,j}^{(\iota)} = \mathrm{o_p}(1)$ *for* $j = 0, 1$;

(iii) $K_n^{(\iota)} = f_Z(s_q)\mu_0^{(\iota)} + \mathrm{o_p}(1)$;

(iv) *further, if the derivative of* $\sigma^2(x,z)$ *with respect to* z *is continuous and bounded on the complete* D, $N_{n,1}^{(\iota)} = f_Z(s_q)\mu_0^{(\iota)}\sigma^2(s_q) + \mathrm{o_p}(1)$,

where the above terms are defined in (2.27)–(2.32).

Proof. After a change of variable and the Taylor expansion, we have

$$\begin{aligned}
\mathrm{E}[\acute{S}_{n,j}^{(\iota)}] =& \mathrm{E}\left[X_i X_i^\top \left(\frac{Z_i - z}{h_n}\right)^j K_h^{(\iota)}(Z_i - z)\middle| Z_i < s_q\right] \\
=& \int\int_{-1}^{-\tau} xx^\top u^j K^{(\iota)}(u) f(x, s_q + (\tau+u)h_n)\mathrm{d}u\mathrm{d}x \\
=& \int_{-1}^{-\tau} u^j K^{(\iota)}(u)\mathrm{d}u \cdot \lim_{z\uparrow s_q} f_Z(z) \int xx^\top \frac{f(x,z)}{f_Z(z)}\mathrm{d}x + \mathrm{O}(h_n) \\
=& \acute{\mu}_{j,\tau}^{(\iota)} f_Z(s_q)\Omega_-(s_q) + \mathrm{O}(h_n)
\end{aligned}$$

due to Assumption 2.A2. The convergence of $\acute{S}_{n,j}^{(\iota)}$ to its expectation follows again by applying Theorem 1 of Hansen (2008), which is allowed due to Assumptions 2.A and 2.B. The convergence results for $\grave{S}_{n,j}^{(\iota)}$ and (ii)–(iv) can be proven in a similar manner. □

Lemma 2.15. *Suppose Assumptions 2.A, 2.B, and 2.D1 hold. It holds for* $n \to +\infty$ *and* $\iota = c, l, r$,

(i) $\tilde{S}_n^{(\iota)} = f_Z(z)\tilde{M}^{(\iota)} \otimes \Omega(z) + \mathrm{o_p}(1)$ *and* $\tilde{S}_n^{(\iota)-1} = \dfrac{\tilde{M}^{(\iota)-1}}{f_Z(z)}(1 + \mathrm{o_p}(1))$;

(ii) $\tilde{W}_{n,1}^{(\iota)} = f_Z(z)\tilde{\mu}_0^{(\iota)}\sigma^2(z) + \mathrm{o_p}(1)$;

(iii) $\tilde{W}_{n,2}^{(\iota)} = f_Z(z)\tilde{\mu}_0^{(\iota)}\Omega(z) + \mathrm{o_p}(1)$;

(iv) $\tilde{W}_{n,3}^{(\iota)} = f_Z(z)\tilde{m}^{(\iota)} \otimes \Omega(z) + \mathrm{o_p}(1)$;

(v) $\tilde{W}_{n,4}^{(\iota)} = \mathrm{o_p}(1)$;

(vi) $\tilde{W}_{n,5}^{(\iota)} = \mathrm{o_p}(1)$,

where the above terms are defined in (2.44)–(2.50).

Proof. This lemma is analogous to Lemma 2.11 and the results follow by direct applications of Theorem 1 in Hansen (2008). □

Lemma 2.16. *Suppose Assumptions 2.A, 2.B, and 2.D1 hold. For any* $z = s_q + \tau h_n$ *with* $\tau \in (-1, 1)$ *and* $\iota = c, l, r$, *we have as* $n \to +\infty$,

(i) $\acute{\tilde{S}}_n^{(\iota)} = f_Z(s_q)\acute{M}_\tau^{(\iota)} + \mathrm{o_p}(1)$ *and* $\grave{\tilde{S}}_n^{(\iota)} = f_Z(s_q)\grave{M}_\tau^{(\iota)} + \mathrm{o_p}(1)$;

(ii) $\acute{\tilde{W}}_{n,1}^{(\iota)} = f_Z(s_q)\acute{\tilde{\mu}}_{0,\tau}^{(\iota)}\sigma_-^2(s_q) + \mathrm{o_p}(1)$ *and* $\grave{\tilde{W}}_{n,1}^{(\iota)} = f_Z(s_q)\grave{\tilde{\mu}}_{0,\tau}^{(\iota)}\sigma_+^2(s_q) + \mathrm{o_p}(1)$;

(iii) $\acute{\tilde{W}}_{n,2}^{(\iota)} = f_Z(s_q)\acute{\tilde{\mu}}_{0,\tau}^{(\iota)}\Omega_-(s_q) + \mathrm{o_p}(1)$ *and* $\grave{\tilde{W}}_{n,2}^{(\iota)} = f_Z(s_q)\grave{\tilde{\mu}}_{0,\tau}^{(\iota)}\Omega_+(s_q) + \mathrm{o_p}(1)$;

(iv) $\acute{\tilde{W}}_{n,3}^{(\iota)} = f_Z(s_q)\acute{m}_\tau^{(\iota)}\mho_-(s_q) + \mathrm{o_p}(1)$ *and* $\grave{\tilde{W}}_{n,3}^{(\iota)} = f_Z(s_q)\grave{m}_\tau^{(\iota)}\mho_+(s_q) + \mathrm{o_p}(1)$;

(v) $\acute{\tilde{W}}_{n,4}^{(\iota)} = \grave{\tilde{W}}_{n,4}^{(\iota)} = \mathrm{o_p}(1)$;

(vi) $\acute{\tilde{W}}_{n,5}^{(\iota)} = \grave{\tilde{W}}_{n,5} = \mathrm{o_p}(1)$,

where the above terms are defined in (2.54)–(2.66).

Proof. This lemma is similar to Lemma 2.14. The results follow mainly by applying Theorem 1 in Hansen (2008). □

Lemma 2.17. *Under Assumption 2.B1, we have*

(i) $\mu_0^{(\iota)}\mu_2^{(\iota)} - \mu_1^{(\iota)2} > 0, \quad \iota = c, l, r$;

(ii)
$$\acute{\mu}_{0,\tau}^{(\iota)}\acute{\mu}_{2,\tau}^{(\iota)} - \acute{\mu}_{1,\tau}^{(\iota)2} \begin{cases} > 0, & \text{if } \iota = c \text{ and } \tau \in (-1, 1), \\ > 0, & \text{if } \iota = l \text{ and } \tau \in (-1, 1), \\ > 0, & \text{if } \iota = r \text{ and } \tau \in (-1, 0), \\ = 0, & \text{if } \iota = r \text{ and } \tau \in [0, 1); \end{cases}$$

(iii)
$$\grave{\mu}_{0,\tau}^{(\iota)}\grave{\mu}_{2,\tau}^{(\iota)} - \grave{\mu}_{1,\tau}^{(\iota)2} \begin{cases} > 0, & \textit{if } \iota = c \textit{ and } \tau \in (-1,1), \\ = 0, & \textit{if } \iota = l \textit{ and } \tau \in (-1,0], \\ > 0, & \textit{if } \iota = l \textit{ and } \tau \in (0,1), \\ > 0, & \textit{if } \iota = r \textit{ and } \tau \in (-1,1), \end{cases}$$

Proof. Here, we prove part (ii) only and (i) and (iii) can be shown analogically. Suppose that U has a density $K^{(\iota)}(\cdot)$. We have

$$\begin{aligned} \mathrm{var}(U|U<-\tau) =& \mathrm{E}[\{U - \mathrm{E}(U|U<-\tau)\}^2|U<-\tau] \\ =& \mathrm{E}(U^2|U<-\tau) - \{\mathrm{E}(U|U<-\tau)\}^2 \\ =& \int_{-1}^{-\tau} u^2 \frac{K^{(\iota)}(u)}{\int_{-1}^{-\tau} K^{(\iota)}(u)\mathrm{d}u}\mathrm{d}u - \left\{\int_{-1}^{-\tau} u \frac{K^{(\iota)}(u)}{\int_{-1}^{-\tau} K^{(\iota)}(u)\mathrm{d}u}\mathrm{d}u\right\}^2 \\ =& \frac{\int_{-1}^{-\tau} u^2 K^{(\iota)}(u)\mathrm{d}u}{\int_{-1}^{-\tau} K^{(\iota)}(u)\mathrm{d}u} - \left\{\frac{\int_{-1}^{-\tau} uK^{(\iota)}(u)\mathrm{d}u}{\int_{-1}^{-\tau} K^{(\iota)}(u)\mathrm{d}u}\right\}^2 \\ =& \frac{\acute{\mu}_{2,\tau}^{(\iota)}}{\acute{\mu}_{0,\tau}^{(\iota)}} - \frac{\acute{\mu}_{1,\tau}^{(\iota)2}}{\acute{\mu}_{0,\tau}^{(\iota)2}}. \end{aligned}$$

By Assumption 2.B1 and definitions of $K^{(r)}(\cdot)$ and $K^{(l)}(\cdot)$ in (2.2),

$$\acute{\mu}_{0,\tau}^{(\iota)}\acute{\mu}_{2,\tau}^{(\iota)} - \acute{\mu}_{1,\tau}^{(\iota)2} = \acute{\mu}_{0,\tau}^{(\iota)2}\mathrm{var}(U|U<-\tau) \begin{cases} > 0, & \text{if } \iota = c \text{ and } \tau \in (-1,1), \\ > 0, & \text{if } \iota = l \text{ and } \tau \in (-1,1), \\ > 0, & \text{if } \iota = r \text{ and } \tau \in (-1,0), \\ = 0, & \text{if } \iota = r \text{ and } \tau \in [0,1). \end{cases}$$

□

Lemma 2.18. *Let X be a symmetric matrix given by*

$$X = \begin{pmatrix} A & B^\top \\ B & C \end{pmatrix}.$$

Then

(i) *X is positive definite if and only if A and the Schur complement of A, $C - BA^{-1}B^\top$, are both positive definite.*

(ii)
$$X^{-1}\begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} I_p \\ 0_p \end{pmatrix},$$

where I_p is the $p \times p$ identity matrix and 0_p is the null matrix of size $p \times p$, if A, B and C are $p \times p$ matrices.

Proof. Part (i) is one of the fundamental results of Schur complement, where the proof can be found in Zhang (2005, Theorem 1.12). For part (ii), since $X^{-1}X = I_{2p}$, we have

$$\begin{aligned}
&X^{-1}X\begin{pmatrix} I_p \\ 0_p \end{pmatrix} = I_{2p}\begin{pmatrix} I_p \\ 0_p \end{pmatrix} \\
&\Leftrightarrow X^{-1}\begin{pmatrix} A & B^\top \\ B & C \end{pmatrix}\begin{pmatrix} I_p \\ 0_p \end{pmatrix} = \begin{pmatrix} I_p & 0_p \\ 0_p & I_p \end{pmatrix}\begin{pmatrix} I_p \\ 0_p \end{pmatrix} \\
&\Leftrightarrow X^{-1}\begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} I_p \\ 0_p \end{pmatrix}.
\end{aligned}$$

□

Lemma 2.19. *Under Assumptions 2.B1 and 2.A4,*

(i) *the variance matrix*
$$\acute{\Omega}^{(\iota)}_{-,\tau}(s_q) = \begin{pmatrix} \acute{\mu}^{(\iota)}_{0,\tau}\Omega_-(s_q) & \acute{\mu}^{(\iota)}_{1,\tau}\Omega_-(s_q) \\ \acute{\mu}^{(\iota)}_{1,\tau}\Omega_-(s_q) & \acute{\mu}^{(\iota)}_{2,\tau}\Omega_-(s_q) \end{pmatrix}$$
is
$$\begin{cases} \text{positive definite,} & \text{if } \iota = c \text{ and } \tau \in (-1,1), \\ \text{positive definite,} & \text{if } \iota = l \text{ and } \tau \in (-1,1), \\ \text{positive definite,} & \text{if } \iota = r \text{ and } \tau \in (-1,0), \\ \text{a null matrix,} & \text{if } \iota = r \text{ and } \tau \in [0,1); \end{cases}$$

(ii) *the variance matrix*

$$\grave{\Omega}^{(\iota)}_{+,\tau}(s_q) = \begin{pmatrix} \grave{\mu}^{(\iota)}_{0,\tau}\Omega_+(s_q) & \grave{\mu}^{(\iota)}_{1,\tau}\Omega_+(s_q) \\ \grave{\mu}^{(\iota)}_{1,\tau}\Omega_+(s_q) & \grave{\mu}^{(\iota)}_{2,\tau}\Omega_+(s_q) \end{pmatrix}$$

is

$$\begin{cases} \textit{positive definite,} & \textit{if } \iota = c \textit{ and } \tau \in (-1,1), \\ \textit{a null matrix,} & \textit{if } \iota = l \textit{ and } \tau \in (-1,0], \\ \textit{positive definite,} & \textit{if } \iota = l \textit{ and } \tau \in (0,1), \\ \textit{positive definite,} & \textit{if } \iota = r \textit{ and } \tau \in (-1,1); \end{cases}$$

(iii) *for* $\tau \in (-1,1)$ *and* $\iota = c, l, r$, *the variance matrix* $\acute{\Omega}^{(\iota)}_{-,\tau}(s_q) + \grave{\Omega}^{(\iota)}_{+,\tau}(s_q)$ *is positive definite.*

Proof. By Assumptions 2.B1 and 2.A4, $\acute{\mu}^{(\iota)}_{0,\tau}\Omega_-(s_q)$ is positive definite except for $\iota = r$ and $\tau \in [0,1)$ when it equals the null matrix. Also, the Schur complement of $\acute{\mu}^{(\iota)}_{0,\tau}\Omega_-(s_q)$ is

$$\acute{\mu}^{(\iota)}_{2,\tau}\Omega_-(s_q) - \acute{\mu}^{(\iota)}_{1,\tau}\Omega_-(s_q)\acute{\mu}^{(\iota)-1}_{0,\tau}\Omega^{-1}_-(s_q)\acute{\mu}^{(\iota)}_{1,\tau}\Omega_-(s_q) = \left(\acute{\mu}^{(\iota)}_{2,\tau} - \frac{\acute{\mu}^{(\iota)2}_{1,\tau}}{\acute{\mu}^{(\iota)}_{0,\tau}}\right)\Omega_-(s_q),$$

which is also positive definite by Lemma 2.17 and Assumption 2.A4 except for the case of $\iota = r$ and $\tau \in [0,1)$ when it equals the null matrix. After applying Lemma 2.18(i), the proof of part (i) is complete. Similarly, one can prove (ii). The claim (iii) then follows immediately from (i) and (ii). □

Lemma 2.20. *Under Assumptions 2.B1 and 2.A4, for* $\iota = l, r, c$,

(i) $\operatorname{rank}\left(\Xi^{(\iota)}_{0,\tau}\right) = p$, *if* $\grave{\mu}^{(\iota)}_{0,\tau} > 0$;

(ii) $\operatorname{rank}\left(\Xi^{(\iota)}_{0,\tau} - I_p\right) = p$, *if* $\acute{\mu}^{(\iota)}_{0,\tau} > 0$,

where the matrix $\Xi^{(\iota)}_{0,\tau}$ *is defined in* (2.36).

Proof. Using Lemma 2.19(iii) and the properties of a positive definite matrix, matrix $\acute{\Omega}^{(\iota)}_{-,\tau}(s_q) + \grave{\Omega}^{(\iota)}_{+,\tau}(s_q)$ is non-singular and its inverse is also positive definite. By Lemma 2.19(ii)

and the fact that $AB \succ 0$ if $A \succ 0$ and $B \succ 0$, the matrix

$$\Xi_{\tau}^{(\iota)} = \left[\acute{\Omega}_{-,\tau}^{(\iota)}(s_q) + \grave{\Omega}_{+,\tau}^{(\iota)}(s_q)\right]^{-1} \grave{\Omega}_{+,\tau}^{(\iota)}(s_q) \succ 0$$

if $\grave{\mu}_{0,\tau}^{(\iota)} > 0$. Since

$$\begin{aligned}
\Xi_{0,\tau}^{(\iota)} =& [I_p \;\; 0_p] \begin{pmatrix} \Xi_{0,\tau}^{(\iota)} \\ \Xi_{1,\tau}^{(\iota)} \end{pmatrix} \\
=& [I_p \;\; 0_p] \left[\acute{\Omega}_{-,\tau}^{(\iota)}(s_q) + \grave{\Omega}_{+,\tau}^{(\iota)}(s_q)\right]^{-1} \begin{pmatrix} \grave{\mu}_{0,\tau}^{(\iota)}\Omega_+(s_q) \\ \grave{\mu}_{1,\tau}^{(\iota)}\Omega_+(s_q) \end{pmatrix} \\
=& [I_p \;\; 0_p] \left[\acute{\Omega}_{-,\tau}^{(\iota)}(s_q) + \grave{\Omega}_{+,\tau}^{(\iota)}(s_q)\right]^{-1} \grave{\Omega}_{+,\tau}^{(\iota)}(s_q) \begin{pmatrix} I_p \\ 0_p \end{pmatrix}
\end{aligned}$$

and the property of positive definite matrix that $A^\top BA \succ 0$ if $B \succ 0$ and A has full column rank, we conclude that $\Xi_{0,\tau}^{(\iota)} \succ 0$. Hence $\Xi_{0,\tau}^{(\iota)}$ has full rank, i.e., rank $\left(\Xi_{0,\tau}^{(\iota)}\right) = p$, which completes the proof of (i).

To show (ii), we write

$$\begin{aligned}
I_p - \Xi_{0,\tau}^{(\iota)} =& [I_p \;\; 0_p] \left\{ I_{2p} - \left[\acute{\Omega}_{-,\tau}^{(\iota)}(s_q) + \grave{\Omega}_{+,\tau}^{(\iota)}(s_q)\right]^{-1} \grave{\Omega}_{+,\tau}^{(\iota)}(s_q) \right\} \begin{pmatrix} I_p \\ 0_p \end{pmatrix} \\
=& [I_p \;\; 0_p] \left[\acute{\Omega}_{-,\tau}^{(\iota)}(s_q) + \grave{\Omega}_{+,\tau}^{(\iota)}(s_q)\right]^{-1} \left\{ \acute{\Omega}_{-,\tau}^{(\iota)}(s_q) + \grave{\Omega}_{+,\tau}^{(\iota)}(s_q) - \grave{\Omega}_{+,\tau}^{(\iota)}(s_q) \right\} \begin{pmatrix} I_p \\ 0_p \end{pmatrix} \\
=& [I_p \;\; 0_p] \left[\acute{\Omega}_{-,\tau}^{(\iota)}(s_q) + \grave{\Omega}_{+,\tau}^{(\iota)}(s_q)\right]^{-1} \acute{\Omega}_{-,\tau}^{(\iota)}(s_q) \begin{pmatrix} I_p \\ 0_p \end{pmatrix}.
\end{aligned}$$

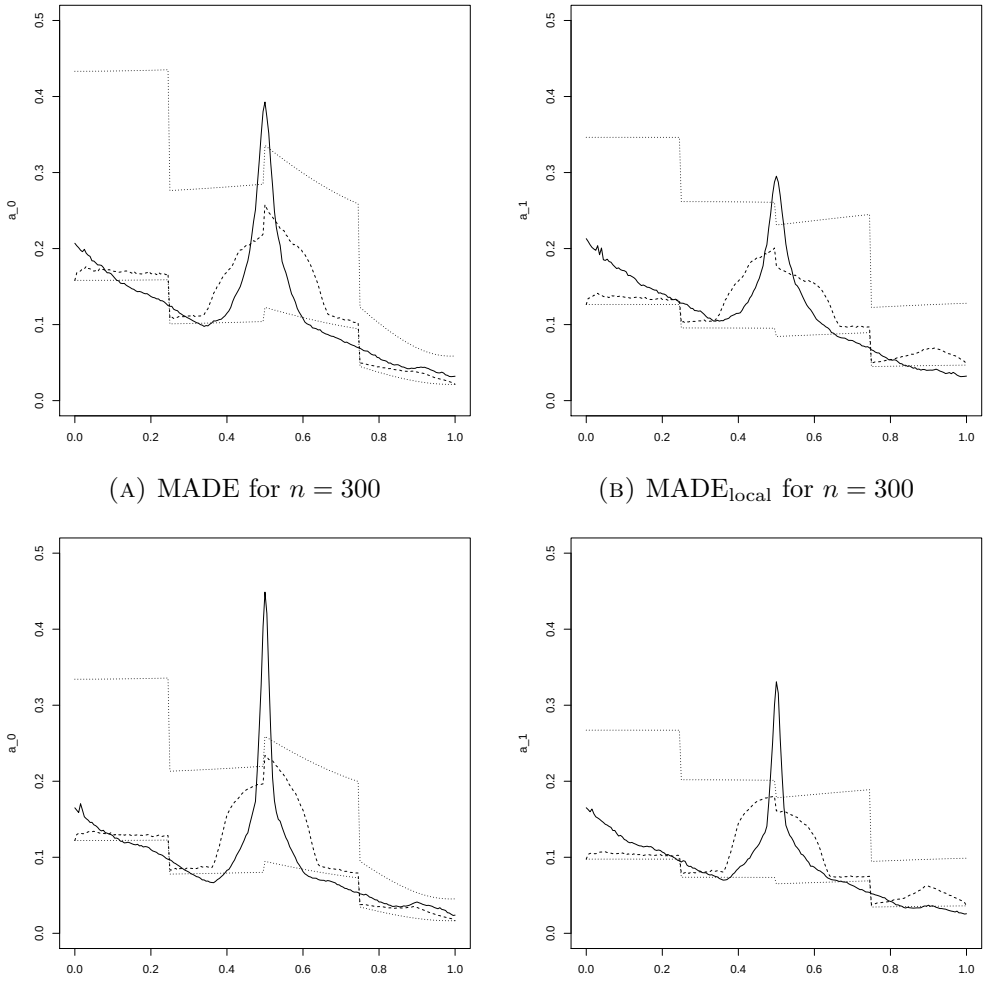
By similar arguments as in part (i) and Lemmas 2.19(i) and 2.19(iii), it follows that $I_p - \Xi_{0,\tau}^{(\iota)} \succ 0$. As a result, $I_p - \Xi_{0,\tau}^{(\iota)}$ has the full rank just as matrix $\Xi_{0,\tau}^{(\iota)} - I_p$. □

2.10 Appendix: Experiment 2: discontinuous conditional variance function with multiple jumps

Here we consider the same time-varying AR(1) process as in (2.13), but with a discontinuous conditional variance function:

$$\sigma^2(Z_t) = (1.4 - 0.6 \cdot \mathbf{1}\{Z_t \geq 0.25\} - 0.6 \cdot \mathbf{1}\{Z_t \geq 0.75\})^2 \,. \tag{2.74}$$

The evaluation is performed in the same way as in Section 2.5, see Figures 2.12–2.15. In addition to that, we also compare the standard errors obtained from this simulation experiment (with fixed bandwidths set again to $0.54n^{-1/5}$) and the standard errors implied by Theorem 2.10 for the estimator proposed in Section 2.4, see Figure 2.11. There is a relatively close correspondence between the simulated and asymptotic standard errors once we take into account that the asymptotic distribution obtained in Theorem 2.10 does not apply at the jump points of the coefficient functions and that the simulation uses a positive bandwidth around 0.15 (contrary to the asymptotic results obtained for the limiting bandwidth being zero).



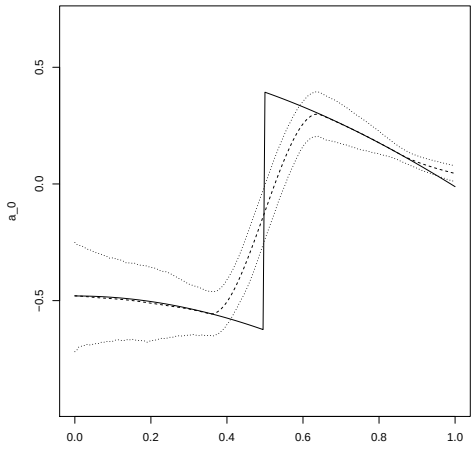
(A) MADE for $n = 300$

(B) $\text{MADE}_{\text{local}}$ for $n = 300$

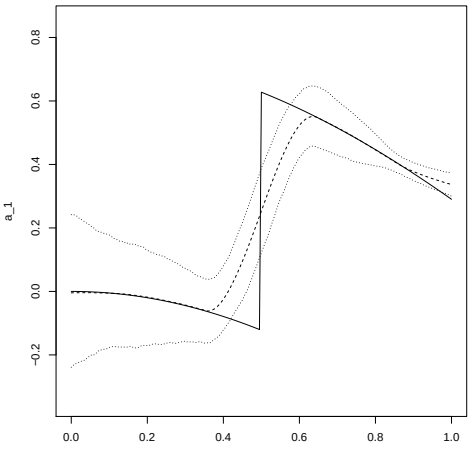
(C) MADE for $n = 600$

(D) $\text{MADE}_{\text{local}}$ for $n = 600$

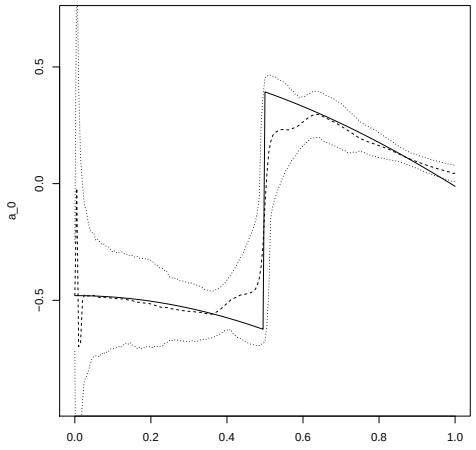
FIGURE 2.11: Heteroskedastic model with the fixed bandwidth: the standard errors of the varying-coefficient estimates based on the simulation (solid line) and on the asymptotic distribution (dashed line). Additionally, the asymptotic standard errors of the centered and left/right kernel estimators are displayed (dotted lines).



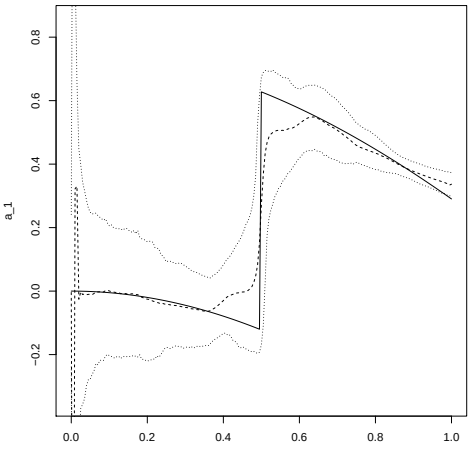
(A) Conventional $\hat{a}_{n,0}$



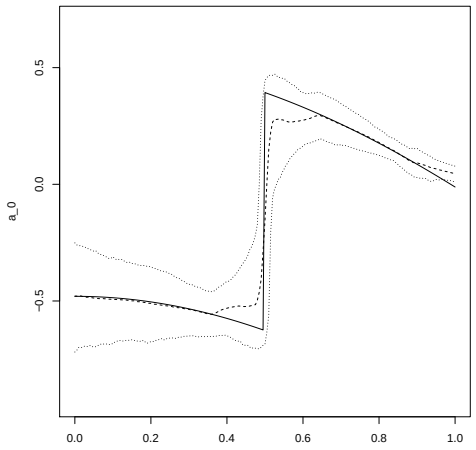
(B) Conventional $\hat{a}_{n,1}$



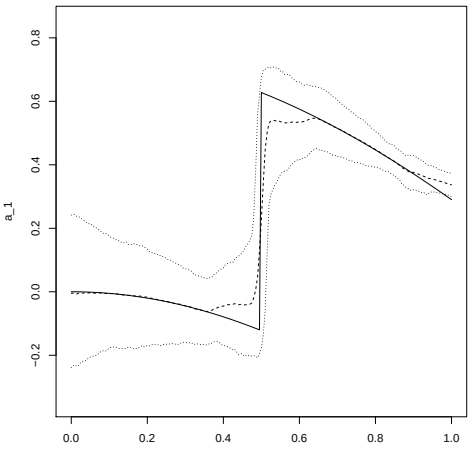
(C) Jump-preserving $\check{a}_{n,0}$ (WRMSE)



(D) Jump-preserving $\check{a}_{n,1}$ (WRMSE)

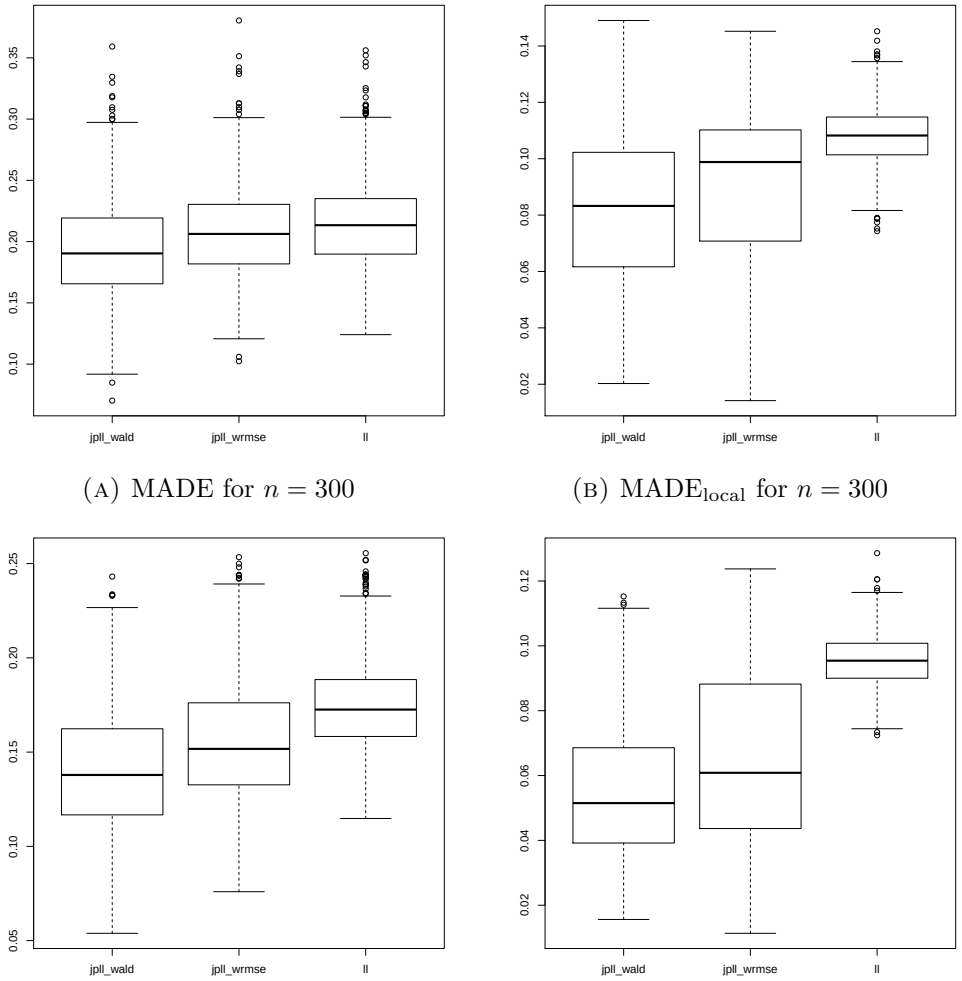


(E) Jump-preserving $\tilde{a}_{n,0}$ (Wald)



(F) Jump-preserving $\tilde{a}_{n,1}$ (Wald)

FIGURE 2.12: Heteroskedastic model with the fixed bandwidth: the solid lines represent the true coefficient functions, the dashed lines are the average varying coefficient estimates, and the dotted lines are the 95% confidence bands.



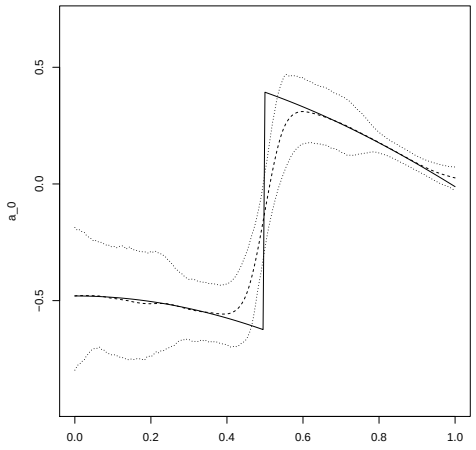
(A) MADE for $n = 300$

(B) $\text{MADE}_{\text{local}}$ for $n = 300$

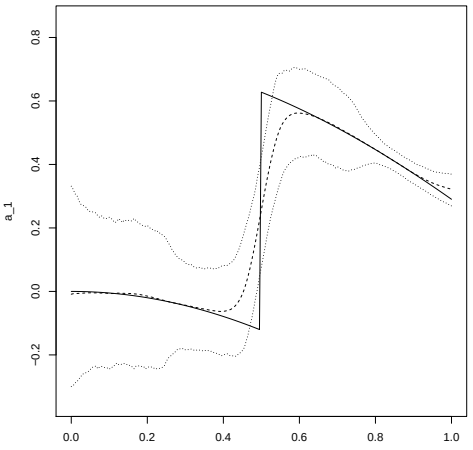
(C) MADE for $n = 600$

(D) $\text{MADE}_{\text{local}}$ for $n = 600$

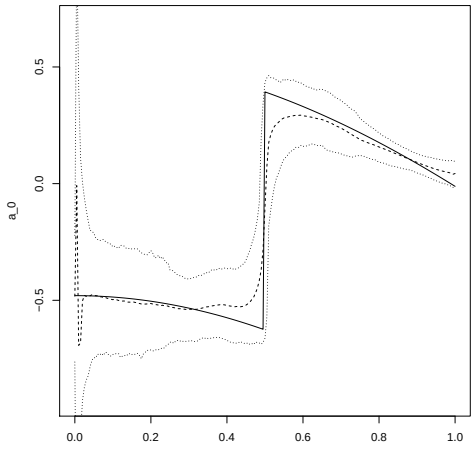
FIGURE 2.13: Heteroskedastic model with the fixed bandwidth: global and local mean absolute deviations of the estimates. Each plot contains boxplots for (from left to right) the jump-preserving estimator based on the Wald statistics, the jump-preserving estimator based on WRMSE, and the conventional estimator.



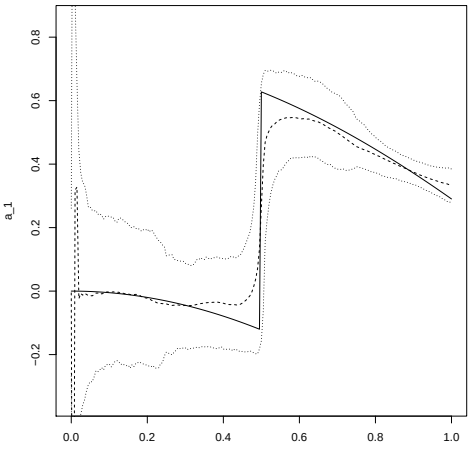
(A) Conventional $\hat{a}_{n,0}$



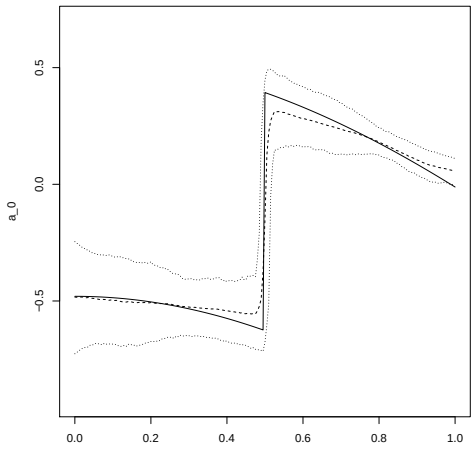
(B) Conventional $\hat{a}_{n,1}$



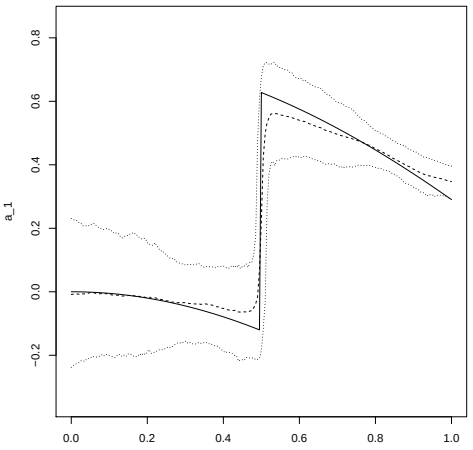
(C) Jump-preserving $\check{a}_{n,0}$ (WRMSE)



(D) Jump-preserving $\check{a}_{n,1}$ (WRMSE)

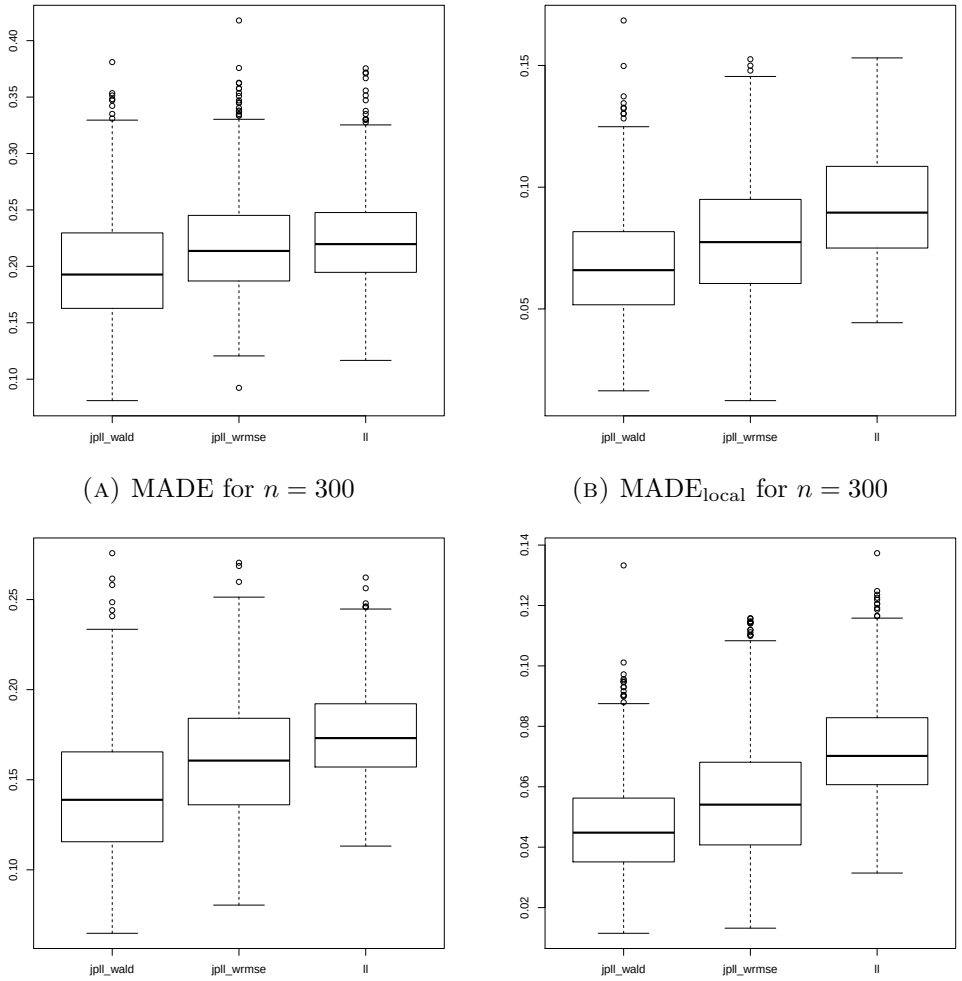


(E) Jump-preserving $\tilde{a}_{n,0}$ (Wald)



(F) Jump-preserving $\tilde{a}_{n,1}$ (Wald)

FIGURE 2.14: Heteroskedastic model with the cross-validated bandwidth: the solid lines represent the true coefficient functions, the dashed lines are the average varying coefficient estimates, and the dotted lines are the 95% confidence bands.



(A) MADE for $n = 300$

(B) $\text{MADE}_{\text{local}}$ for $n = 300$

(C) MADE for $n = 600$

(D) $\text{MADE}_{\text{local}}$ for $n = 600$

FIGURE 2.15: Heteroskedastic model with the cross-validated bandwidth: global and local mean absolute deviations of the estimates. Each plot contains boxplots for (from left to right) the jump-preserving estimator based on the Wald statistics, the jump-preserving estimator based on WRMSE, and the conventional estimator.

Chapter 3

Semiparametric Transition Models*

3.1 Introduction

One class of nonlinear time series models that has been widely applied, for example, in macroeconomics and finance, is the regime-switching model. Among the regime-switching models, the threshold autoregressive (TAR) model of Tong (1983) is a classical one: it has been widely studied (see Hansen, 2011, for an overview) and applied (e.g., Potter, 1995; Rothman, 1998). The TAR model however describes only data generating processes that follow purely one of the two regimes – no gradual transition between the regimes is allowed – and can be difficult to estimate due to discontinuous regression function.

To overcome these limitations, the smooth transition autoregressive (STAR) model was first introduced by Chan and Tong (1986) and further developed by Teräsvirta (1994); see van Dijk et al. (2002) for a survey. The two-regime STAR model is given by

$$y_t = x_t^\top \beta_1 \{1 - w(z_t;\theta)\} + x_t^\top \beta_2 w(z_t;\theta) + \varepsilon_t, \qquad t = 1,\dots,T, \tag{3.1}$$

where x_t contains lagged values of the response variable y_t, z_t is an observable continuously distributed transition variable, and $w(\cdot;\theta) : \mathbb{R} \mapsto \mathbb{R}$ is a smooth transition function

*This chapter is based on Čížek and Koo (2017b), Semiparametric transition models. Unpublished manuscript, Tilburg University.

known up to a finite-dimensional vector θ of parameters. The TAR model would correspond to $w(z;\theta) = \mathbf{1}(z > \theta)$ ($\mathbf{1}(\cdot)$ denotes the indicator function). Among smooth transition functions, a popular choice of $w(\cdot;\theta)$ is a logistic distribution function $\Lambda(z;\mu,s) = \{1+\exp[-s(z-\mu)]\}^{-1}$ with $\theta = (\mu,s)^\top$, which increases smoothly and monotonically with z. The corresponding logistic STAR (LSTAR) model has been used to model business cycle asymmetry, for instance, where the regimes correspond to expansions and recessions (Teräsvirta and Anderson, 1992; Skalin and Teräsvirta, 2002). Another practically applied transition function is an exponential function $G(z;\mu,s) = 1-\exp[-s(z-\mu)^2]$, in which the regimes are associated with large and small absolute deviations of z from μ. This so-called exponential STAR (ESTAR) model has been applied, for example, to real exchange rate data (Taylor et al., 2001; Sarantis, 1999). Recent extensions of the two-regime STAR models (3.1) include the multiple-regime STAR models (van Dijk and Franses, 1999), flexible-coefficient STAR models (Medeiros and Veiga, 2003, 2005), time-varying STAR models (Lundbergh et al., 2003), STAR models with multivariate z_t (Taylor et al., 2000), vector STAR models (Hubrich and Teräsvirta, 2013), and transition models with endogenous explanatory variables (Areosa et al., 2011).

In the STAR model (3.1), the transition function $w(\cdot;\theta)$ characterized by the parameter θ is assumed to be a known continuously differentiable function; typically, it is also bounded between 0 and 1. The assumption that the transition function is smooth and has a certain parametric form is however hardly justified. Although a misspecified transition function can lead to inconsistent estimates and wrong inference, it can sometimes serve as an approximation in practice (Chan and Tong, 1986). The original TAR avoids this problem by focusing purely on the two regimes, but contrary to STAR, it cannot adapt to situations with intermediate states that are combinations of the two regimes. Therefore, we introduce a flexible transition model in which its transition function is of an unknown form, possibly with a finite set of discontinuities: the semiparametric transition (SETR) model. The SETR model extends TAR similarly to STAR, but has three main advantages over the STAR model. First, the risk of model misspecification is substantially reduced as the transition function is only assumed to be smooth (up a finite set of discontinuities). Next, even though the estimators of regression coefficients β_1 and β_2 do not rely on any parametric form of the transition function w, their rate of convergence is proved to be the same as in the (S)TAR model. Finally, the estimates of the transition function in

the SETR model can be used to study important features of the transition between the regimes (e.g., the size and location of a jump or overshooting behavior in the transition process). Contrary to the parametric STAR models, the identification of the general SETR model with an unknown form of the transition function requires that the transition process reaches each regime with a positive probability just like in the TAR models, and since the transition function is estimated nonparametrically, the transition function can be estimated only in the range of observed values of transition variable z_t.

Although the SETR model nests the TAR and STAR models if the transition functions in the STAR models reach 0 and 1 with a positive probability,[†] the SETR model is a special case of varying-coefficient models studied by Chen and Tsay (1993) and Hastie and Tibshirani (1993), for instance. The varying-coefficient model has the following form:

$$\mathrm{E}[y_t|x_t, z_t] = x_t^\top m(z_t), \qquad t = 1, \ldots, T, \tag{3.2}$$

where $m(\cdot) : \mathbb{R} \to \mathbb{R}$ is an unknown vector function and z_t is a scalar index. Recent works on model (3.2) include Hoover et al. (1998), Wu et al. (1998), and Fan and Zhang (2000) on longitudinal data analysis and Chen and Tsay (1993), Cai et al. (2000), and Huang and Shen (2004) on nonlinear time series. Moreover, Zhang et al. (2002), Fan and Huang (2005), and Ahmad et al. (2005) studied the partial linearly varying-coefficient model in which some elements of $m(\cdot)$ are constant. Recently, Chen and Hong (2012) designed a test of the STAR model (3.1) versus the varying-coefficient model (3.2).

In the varying-coefficient models, the vector of coefficient functions $m(\cdot)$ is of interest and is estimated nonparametrically. Consequently, its estimates cannot reach the rate of convergence typical for estimates of parametric models such as (S)TAR and require thus larger data sets for sufficiently precise inference. On the contrary, as the SETR model applies nonparametric estimation only to the transition function, the estimators of regression coefficients β_1 and β_2, which are fixed in each regime, are proved to converge to their corresponding true values at the same rate as the slope estimates of the parametric (S)TAR model (3.1).

[†]In principle, the LSTAR and ESTAR models are excluded from the class of the SETR models as their transition functions converge to but never reach 0 or 1. See Section 3.2 for details.

This chapter is organized as follows. In Section 3.2, the model and identification conditions are presented. In Section 3.3, an estimation method of the semiparametric transition model is proposed. The consistency and asymptotic distribution of the proposed estimators are discussed in Section 3.4. Finally, a simulation study and real-data application of the SETR estimators are presented in Sections 3.5 and 3.6. All proofs are collected in Sections 3.8 and 3.9.

Throughout this chapter, the following notation is used. Let $\|x\| = (x^\top x)^{1/2}$ for any vector $x \in \mathbb{R}^p$ and $\|X\| = \mathrm{tr}(X^\top X)^{1/2}$ for any matrix X. For a transition function $w(z_t)$ of the random variable z_t with density f_z and a given $\epsilon > 0$, the following (semi)norm is used: $\|w\|_{\infty,\epsilon,f_z} = \sup_{f_z(z)>\epsilon} |w(z)|$. In addition, let $\mathbf{1}(\cdot)$ denote the indicator function, $\xrightarrow{P}$ convergence in probability, and $\xrightarrow{d}$ convergence in distribution.

3.2 The semiparametric transition model

Consider the following two-regime semiparametric transition model (noting that two regimes are considered only for simplicity and the proposed model and estimation procedure trivially extend to more regimes):

$$y_t = x_t^\top \beta_1^0 \cdot \{1 - w^0(z_t)\} + x_t^\top \beta_2^0 \cdot w^0(z_t) + \varepsilon_t, \qquad t = 1, \ldots, T, \tag{3.3}$$

where y_t is the dependent variable, $x_t \in \mathbb{R}^p$ is a vector of covariates, $z_t \in \mathbb{R}$ is a scalar transition variable, and ε_t denotes the error term. The parameters of interest – slopes β_1^0 and β_2^0 – are the vectors of regression coefficients corresponding to the first and second regimes, respectively, and $w^0(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ is an unknown piecewise-smooth transition function. When lagged dependent variables are included in the explanatory variables x_t, that is, $x_t = (1, y_{t-1}, y_{t-2}, \ldots, y_{t-p+1})^\top$, model (3.3) can be referred to as the semiparametric transition autoregressive model. Here, the transition variable z_t can be both exogenous or endogenous in the sense that it contains lagged dependent variables analogously to STAR by Teräsvirta (1994). The proposed estimation procedure also extends to a deterministic transition variable z_t such as the linear time trend t/T in Lin and Teräsvirta (1994). The assumptions and asymptotic analysis presented in this chapter are however designed only

for a stationary continuous random variable z_t and the case $z_t = t/T$ is thus excluded from the asymptotic analysis, although it is empirically tested in simulations.

The structural-break, threshold, and smooth transition models can be thus viewed as special cases of the SETR model (3.3). For $z_t = t/T$ being a time fraction and the transition function w^0 equal to the indicator function $\mathbf{1}(z_t \geq t_B/T)$ for an unknown break point t_B, the SETR model reduces to the structural-break model. Similarly, when $w^0(z_t; z_B) = \mathbf{1}(z_t \geq z_B)$ for a random variable z_t and an unknown threshold z_B, model (3.3) becomes the threshold model. Finally, imposing that transition function $w^0(\cdot)$ has a smooth parametric form $w^0(\cdot;\theta)$ characterized by the parameter θ yields the smooth transition model (3.1).

In the SETR model, we however consider more general functions $w^0(\cdot)$ that belong to some space $\mathcal{W}$ of functions satisfying the following definition.

Definition 3.1. Let $\mathcal{W}$ represents the space of functions $w : \mathbb{R} \to \mathbb{R}$ that are continuous up to a finite number J of points $s_1, \ldots, s_J \in \mathbb{R}$, uniformly bounded by $M > 0$ on $\mathbb{R}$, differentiable (from the right at points $s_1, \ldots, s_J$) with derivatives uniformly bounded by M too, and that are equal to 0 and 1 on given intervals (a_1, b_1) and (a_2, b_2), respectively.

The parameters given in the general model (3.3) cannot however be identified unless additional restrictions are imposed on the explanatory variables, slope parameters, and the transition function.

Assumption 3.A. Let $\{x_t, z_t, \varepsilon_t\}_{t=1}^{\infty}$ be a sequence of strictly stationary random vectors with the marginal distributions of z_t and ε_t being absolutely continuous such that

3.A1. $\mathrm{E}[\varepsilon_t|\mathcal{I}_t] = 0$ with $\mathcal{I}_t = \{x_{t-j}, z_{t-j}\}_{j=0}^{\infty}$;

3.A2. the true slope parameters $\beta^0 = (\beta_1^{0\top}, \beta_2^{0\top})^\top \in \mathcal{B}$, where $\mathcal{B}$ is a compact subset of $\mathbb{R}^{2p}$, are such that the k-th elements of the parameter vectors in each regime satisfy $\beta_{1k}^0 < \beta_{2k}^0$, where k represents the smallest integer such that $\beta_{1k}^0 \neq \beta_{2k}^0$;

3.A3. for any $\delta > 0$, the infimum of eigenvalues of $\mathrm{E}[x_t x_t^\top | z_t \in I_z]$ taken across all intervals $I_z \subseteq \mathbb{R}$ with $P(z_t \in I_z) \geq \delta$ is strictly positive and $\mathrm{E}[x_t x_t^\top | z_t \in I_z]$ is continuous with respect to the bounds of I_z.

3.A4. Furthermore, the true transition function $w^0 \in \mathcal{W}$ for a function space $\mathcal{W}$ that satisfies Definition 3.1 with intervals (a_1, b_1) and (a_2, b_2) such that $P(z_t \in (a_1, b_1)) > 0$ and $P(z_t \in (a_2, b_2)) > 0$.

The first Assumption 3.A is stated for strictly stationary random vectors since model (3.3) contains general nonlinear functions of the data. In Assumption 3.A1, the error term ε_t is conditionally mean independent of the σ-field $\mathcal{I}_t$ generated by the current and past values of x_t and z_t so that the conditional mean of response y_t is correctly represented by the regression function in model (3.3). Assumption 3.A2 requires the slope coefficients to be different in the two regimes: otherwise, it is not possible to distinguish the regimes and to identify the transition function. The imposed inequality between β_1^0 and β_2^0 prevents relabeling of the symbols $(\beta_1^0, \beta_2^0, w^0)$ to $(\beta_2^0, \beta_1^0, 1 - w^0)$. Further, the full-rank condition 3.A3 is similar to usual assumptions in the threshold and structural-break models for identification (e.g., Assumption A2 in Bai and Perron, 1998) and it can be seen as a weaker form of the standard assumption $\mathrm{E}(x_t x_t^\top | z_t = z) > 0$, see for example Assumption 1.7 in Hansen (2000), which is sufficient for the presented results and reduces to $\mathrm{E}(x_t x_t^\top) > 0$ if x_t is independent of z_t. The full-rank condition is imposed for any interval I_z with a non-zero probability of $z_t \in I_z$ to identify the transition function $w^0(\cdot)$ almost everywhere. If the aim is to identify only the slopes β_1 and β_2, a substantially weaker assumption has to hold: two matrices $\mathrm{E}[x_t x_t^\top | z_t \in (a_1, b_1)]$ and $\mathrm{E}[x_t x_t^\top | z_t \in (a_2, b_2)]$ have to be non-singular, where the intervals (a_1, b_1) and (a_2, b_2) are given in Assumption 3.A4.

Next, Assumption 3.A4 defines the space of functions $\mathcal{W}$ in which the transition function is searched for. Although we assume differentiability of the functions, which will be necessary later to derive the asymptotic distribution, assuming that functions w are Lipschitz continuous (within the intervals of continuity) uniformly on $\mathcal{W}$ would be sufficient for identification. Moreover, note that – without the right continuity (or differentiability) of functions at the points of discontinuity – the identification of w^0 would not be possible at those points.

Finally, Assumption 3.A4 ensures that the system described by model (3.3) is with a positive probability in the first regime described by β_1^0 (when $z_t \in (a_1, b_1)$) and in the second regime defined by β_2^0 (when $z_t \in (a_2, b_2)$). This is essential because the slope

parameters β_1 and β_2 are not identifiable by using other values of z_t alone due to further unspecified $w(z_t)$. This assumption is satisfied in the TAR and some STAR models, but although practical difference is likely negligible, it excludes the LSTAR and ESTAR models as their corresponding transition functions reach 0 or 1 only asymptotically. The SETR analog of LSTAR would be based on the assumption that $w(z_t) = 0$ if $z_t < b_1$, $P(z_t \in (-\infty, b_1)) > 0$, and $w(z_t) = 1$ if $z_t > a_2$, $P(z_t \in (a_2, +\infty)) > 0$. Analogously to common practice in the structural-break estimation, one could thus set that z_t below its α-th quantile and above its $(1-\alpha)$-th quantile correspond to the first and second regime, respectively. Similarly, the SETR analog of ESTAR would hinge on the assumption that $w(z_t) = 0$ if $|z_t| < b_1$, $P(z_t \in (-b_1, b_1)) > 0$, and $w(z_t) = 1$ if $|z_t| > a_2$, $P(z_t \in (-\infty, -a_2) \cup (a_2, +\infty)) > 0$. As in these two examples, $w(z)$ can be assumed to differ from 0 or 1, $w(z) \notin \{0, 1\}$, only on a compact subset of $\mathbb{R}$ in most applications.

The main identification result for model (3.3) is stated in the following theorem. Note that the transition function is identified only up to a set with $f_z(z) = 0$, where f_z represents the density function of z_t.

Theorem 3.2. *If the process $\{y_t, x_t, z_t\}$ follows model (3.3) and Assumption 3.A is satisfied, then (β^0, w^0) are uniquely identified in $\mathcal{B} \times \mathcal{W}$ (up to a set with zero density in the case of w^0): it holds for any $\delta > 0$ and $\epsilon > 0$ that*

$$\inf_{\|\beta-\beta^0\|>\delta \text{ or } \|w-w^0\|_{\infty,\epsilon,f_z}>\delta} \mathrm{E}[y_t - x_t^\top \beta_1 - x_t^\top(\beta_2 - \beta_1)w(z_t)]^2 > \mathrm{E}[y_t - x_t^\top \beta_1^0 - x_t^\top(\beta_2^0 - \beta_1^0)w^0(z_t)]^2, \tag{3.4}$$

where $\beta \in \mathcal{B}$ and $w \in \mathcal{W}$.

Theorem 3.2 establishes that the slopes and transition function can be found by minimizing the nonlinear least squares criterion, where the population version (3.4) has to be replaced by its finite-sample equivalent. The joint minimization of this criterion with respect to $\beta = (\beta_1^\top, \beta_2^\top)^\top$ and w is however computationally cumbersome (see Section 3.3 for details). We therefore design an algorithm that requires only linear least squares (LS) estimation in each step. Let us now introduce the basic notation and concepts for this algorithm. The key point is to identify the transition function w for a given value of β and to identify β given some transition function w.

First, given some parameter values $\beta \in \mathcal{B} \subset \mathbb{R}^{2p}$, the expected squared error (3.4) can be minimized with respect to $w(z_t)$ at any $z_t = z$ in the support of z_t. The first-order condition of (3.4) with respect to $w(z)$ at a fixed point z equals

$$\mathrm{E}[(\beta_2 - \beta_1)^\top x_t \{y_t - x_t^\top \beta_1 - x_t^\top (\beta_2 - \beta_1) w(z_t)\} | z_t = z] = 0.$$

If $\mathrm{E}(x_t x_t^\top | z_t = z) > 0$ holds, it is possible to directly solve for $w(z_t) = w(z)$ at a given β:

$$w(z, \beta) = \frac{\mathrm{E}[(\beta_2 - \beta_1)^\top x_t (y_t - x_t^\top \beta_1) | z_t = z]}{\mathrm{E}[(\beta_2 - \beta_1)^\top x_t x_t^\top (\beta_2 - \beta_1) | z_t = z]}. \tag{3.5}$$

However, Assumption 3.A3 only guarantees $\mathrm{E}(x_t x_t^\top | z_t \in I_z) > 0$ for any interval I_z, $z \in I_z$, with length $|I_z| > 0$. The first-order condition will be thus used conditionally on $z_t \in I_z$,

$$\mathrm{E}[(\beta_2 - \beta_1)^\top x_t \{y_t - x_t^\top \beta_1 - x_t^\top (\beta_2 - \beta_1) w(z_t)\} | z_t \in I_z] = 0, \tag{3.6}$$

to solve for $w(z)$. Unless the transition function $w(z_t)$ is constant on I_z, $w(z_t) = w(z)$, the limit $|I_z| \to 0$ has to be taken though to obtain the value of $w(z)$ (under the assumption that intervals I_z are chosen so that $w(z)$ is continuous on them; see Section 3.3.2). The limit of the solution $w(z)$ of (3.6) for a given fixed β at point z can be thus denoted and expressed as[‡]

$$w(z, \beta) = \lim_{|Iz| \to 0} \frac{\mathrm{E}[(\beta_2 - \beta_1)^\top x_t (y_t - x_t^\top \beta_1) | z_t \in I_z]}{\mathrm{E}[(\beta_2 - \beta_1)^\top x_t x_t^\top (\beta_2 - \beta_1) | z_t \in I_z]}. \tag{3.7}$$

Unless $\beta = \beta^0$, function $w(z, \beta)$ does not have to coincide with $w^0(z)$ at z with $f_z(z) > 0$.

On the other hand, given some transition function $w \in \mathcal{W}$, the slope parameters β can be estimated by minimizing the least squares criterion (3.4) with respect to β only. Considering a fixed function $w(\cdot)$ and using the abbreviated notation $\omega_t = [1 - w(z_t), w(z_t)]^\top$, the expected squared error (3.4) can written as $\mathrm{E}[y_t - (\omega_t \otimes x_t)^\top \beta]^2$ due to the expression $x_t^\top \beta_1 \{1 - w(z_t)\} - x_t^\top \beta_2 w(z_t) = (\omega_t \otimes x_t)^\top \beta$. The minimizer of $\mathrm{E}[y_t - (\omega_t \otimes x_t)^\top \beta]^2$, that is, of (3.4) for a given transition function w, can be therefore denoted and expressed as

$$\beta(w) = \{\mathrm{E}[(\omega_t \otimes x_t)(\omega_t \otimes x_t)^\top]\}^{-1} \mathrm{E}[(\omega_t \otimes x_t) y_t]. \tag{3.8}$$

[‡]Note that the finite-sample equivalent of this expression corresponds to a nonparametric local LS estimator with the support of z_t localized to I_z; see Section 3.3.2.

Because of the uniqueness of both partial solutions (3.7) and (3.8), it holds according to Theorem 3.2 that $\beta^0 = \beta(w^0)$ and $\|w^0(z) - w(z, \beta^0)\|_{\infty,\epsilon,f_z} = 0$ for any $\epsilon > 0$. For $w \neq W^0$, $\beta(w)$ generally differs from β^0.

3.3 Estimation

Before discussing the estimation procedure, let $\hat{\beta}_T$ and $\hat{w}_T(\cdot)$ denote the unconditional estimators of β^0 and $w^0(\cdot)$ that directly minimize the sum of squared residuals ($\beta = (\beta_1^\top, \beta_2^\top)^\top$) that correspond to the quadratic loss in Theorem 3.2:

$$\min_{\beta,w} \sum_{t=1}^{T} \left\{ y_t - x_t^\top \beta_1 - x_t^\top (\beta_2 - \beta_1) w(z_t) \right\}^2 . \tag{3.9}$$

Similarly, let $\hat{\beta}_T(w)$ and $\hat{w}_T(\cdot, \beta)$ be the conditional estimators of $\beta(w)$ in (3.8) and $w(\cdot, \beta)$ in (3.7) that minimize the sum of squared residuals given a fixed w and a fixed β, respectively.

Estimating the slope coefficients β and transition function $w(\cdot)$ through direct minimization in (3.9) is intractable in practice. One common strategy in regime-switching models is concentration (e.g., see Hansen, 2000, for the TAR model and Leybourne et al., 1998, for the STAR model). Given some fixed β, the SETR model (3.3) can be viewed as a varying-coefficient model. Applying a nonparametric estimation method from the varying-coefficient literature (see Fan and Zhang, 2008, for a review) yields the conditional estimators $\hat{w}_T(z_1, \beta), \ldots, \hat{w}_T(z_T, \beta)$. The $2p$ regression coefficients are then estimated by minimizing the following concentrated sum of squared residuals:

$$\hat{\beta}_T = \arg \min_{\beta \in \mathbb{R}^{2p}} \sum_{t=1}^{T} \left\{ y_t - x_t^\top \beta_1 - x_t^\top (\beta_2 - \beta_1) \hat{w}_T(z_t, \beta) \right\}^2 .$$

This minimization is however computationally rather demanding.

Instead of the concentration approach, we propose the following estimation algorithm. First, an initial consistent slope estimator $\hat{\beta}_T^{(0)}$ can be constructed by using the data that are purely from the first and second regimes, provided that (sub)intervals of (a_1, b_1) and (a_2, b_2) from Assumption 3.A4 are known. Then the sum of squared residuals with

a fixed $\beta = \hat{\beta}_T^{(0)}$ can be minimized locally (in neighborhoods of points $z_1, \ldots, z_T$) to obtain an initial estimator $\hat{w}_T^{(0)} = \hat{w}_T(\cdot, \hat{\beta}_T^{(0)})$ of the transition function outside of intervals (a_1, b_1) and (a_2, b_2), where w is assumed to be 0 and 1, respectively; see Section 3.3.2 for details. The first step of the algorithm, which is described in Section 3.3.1, is thus to find (sub)intervals of (a_1, b_1) and (a_2, b_2) and to obtain the corresponding estimates $\hat{\beta}_T^{(0)}$ and $\hat{w}_T^{(0)}$.

Next, the initial estimate $\hat{\beta}_T^{(0)}$ uses only subsets of data corresponding purely to the first or the second regime and it is thus inefficient. Given $\hat{\beta}_T^{(0)}$ and $\hat{w}_T^{(0)}$, the slope estimates can however be updated to $\hat{\beta}_T^{(1)} = \hat{\beta}_T(\hat{w}_T^{(0)})$ by minimizing the sum of all squared residuals given the initial estimates $\hat{w}_T^{(0)}(z_t)$, $t = 1, \ldots, T$. Similarly, using the slope estimate $\hat{\beta}_T^{(1)}$, the estimates of the transition function can now be renewed to $\hat{w}_T^{(1)} = \hat{w}_T(\cdot, \hat{\beta}_T^{(1)})$ at points outside of intervals (a_1, b_1) and (a_2, b_2). This procedure can be iterated by estimating $\hat{\beta}_T^{(k)} = \hat{\beta}_T(\hat{w}_T^{(k-1)})$ and $\hat{w}_T^{(k)} = \hat{w}_T(\cdot, \hat{\beta}_T^{(k)})$ for $k = 2, 3, \ldots, K$. In practice, $K = 1$ or $K = 2$ steps are sufficient: in Section 3.4, the asymptotic distribution of $\hat{\beta}_T^{(k)}$ is shown to be asymptotically independent of $k \geq 1$.

In the rest of this section, the choice of the initial intervals and slope estimator $\hat{\beta}_T^{(0)}$ are described in Section 3.3.1, the local nonparametric estimation of $\hat{w}_T(\cdot, \beta)$ is discussed in Section 3.3.2, and finally, the updated LS estimator $\hat{\beta}_T(w)$ is introduced in Section 3.3.3. The full algorithm is summarized in Section 3.3.4.

3.3.1 Initial estimator of β

By Assumption 3.A4, there are regions within the support of transition variable z_t such that the process (3.3) follows only the first or second regime as $w(z_t) = 0$ or $w(z_t) = 1$, respectively. If these regions are assumed to be known, consistent initial estimators $\hat{\beta}_{1,T}^{(0)}$ and $\hat{\beta}_{2,T}^{(0)}$ can be obtained by employing the ordinary LS method for data with values z_t within the regions corresponding to the first and second regimes, respectively. For example, a researcher can assume the observations with $z_t < q_z(\alpha)$ and $z_t > q_z(1-\alpha)$ follow purely the first and second regimes, respectively, where $q_z(\alpha)$ denotes the αth quantile of the z_t distribution. As such an assumption can be usually be made only for a rather small α to avoid misspecification, only small fraction of data can be used to obtain

the initial estimators and they would be very imprecise. To alleviate this problem, we suggest the following interval-selection scheme, which will be later generalized to the case without prior knowledge about regions corresponding to each regime.

To improve the quality of the initial estimators $\hat{\beta}_{1,T}^{(0)}$ and $\hat{\beta}_{2,T}^{(0)}$, consider sufficiently short intervals $(a_1^{(0)}, b_1^{(0)}) \subset (a_1, b_1)$ and $(a_2^{(0)}, b_2^{(0)}) \subset (a_2, b_2)$ and construct increasing sequences of intervals $(a_j^{(0)}, b_j^{(0)}) \subset (a_j^{(1)}, b_j^{(1)}) \subset \ldots \subset (a_j^{(\kappa)}, b_j^{(\kappa)})$ for $j = 1, 2$. For each pair of intervals $(a_1^{(k)}, b_1^{(k)})$ and $(a_2^{(k)}, b_2^{(k)})$, $k = 1, \ldots, \kappa$, estimate $\hat{\beta}_{1,T}^{(0,k)}$ and $\hat{\beta}_{2,T}^{(0,k)}$ forming estimate $\hat{\beta}_T^{(0,k)}$, compute the estimated transition function $\hat{w}_T^{(0,k)} = \hat{w}_T(\cdot, \hat{\beta}_T^{(0,k)})$ outside of intervals $(a_1^{(k)}, b_1^{(k)})$ and $(a_2^{(k)}, b_2^{(k)})$ (see Section 3.3.2), and evaluate the sum $S_{(k)}^2$ of squared residuals (3.9) at $\hat{\beta}_T^{(0,k)}$, $\hat{w}_T^{(0,k)}$. If we do not wish to iterate further (see Section 3.3.4), we define the final initial estimates by $\hat{\beta}_{1,T}^{(0)} = \hat{\beta}_{1,T}^{(0,\hat{k})}$ and $\hat{\beta}_{2,T}^{(0)} = \hat{\beta}_{2,T}^{(0,\hat{k})}$ for $\hat{k} = \arg\min_{k=1,\ldots,\kappa} S_{(k)}^2$, that is, the estimates minimizing the unconditional LS criterion. This procedure is proved to be consistent later in Theorem 3.3 in Section 3.4. Its main practical benefit is that it makes estimation insensitive to the choice of initial intervals $(a_1^{(0)}, b_1^{(0)})$ and $(a_2^{(0)}, b_2^{(0)})$.

If a researcher assumes that α-fractions of the observations are in the first and second regimes, respectively, but the locations of the first and second regimes are not known in advance, one can find those regions by comparing the sums of squared residuals similar to the method described in the previous paragraph. First, we divide the support of z_t into $\lceil 2/\alpha \rceil$ partitions, where each interval contains around the $(\alpha/2)$-fraction of observations. We form all feasible combinations from these intervals by setting each pairs of intervals as candidates of the first and second regimes: $\{(a_1^{(k)}, b_1^{(k)}), (a_2^{(l)}, b_2^{(l)})\}$ for $k, l = 1, \ldots, \lceil 2/\alpha \rceil$ and $k < l$. For each pair of intervals, we construct a potential initial slope estimate $\hat{\beta}_T^{(0,k)}$, estimate the transition function $\hat{w}_T^{(0,k)} = \hat{w}_T(\cdot, \hat{\beta}_T^{(0,k)})$, and evaluate the sum of squared residuals $S_{(k)}^2$. The estimate with the lowest sum of squared residuals is defined as the initial slope estimate, that is, $\hat{\beta}_T^{(0)} = \hat{\beta}_T^{(0,\hat{k})}$ for $\hat{k} = \arg\min_{k=1} S_{(k)}^2$, and the corresponding pair of intervals is recognized as the first and second regimes.

3.3.2 Local linear estimator of $w(\cdot, \beta)$

Given $\beta = (\beta_1^\top, \beta_2^\top)^\top$ with $\beta_1 \neq \beta_2$, the SETR model (3.3) can be reformulated as a varying-coefficient model with a single covariate and no intercept term:

$$\tilde{y}_t = \tilde{x}_t m(z_t) + \varepsilon_t, \tag{3.10}$$

where $\tilde{y}_t = y_t - x_t^\top \beta_1$, $\tilde{x}_t = x_t^\top(\beta_2 - \beta_1)$, $m(\cdot) = w(\cdot, \beta)$, and $x_t^\top \beta_1$ was subtracted from both sides of equation (3.3). Model (3.10) can be now used to estimate $m(\cdot) = w(\cdot, \beta)$.

If the function $m(\cdot)$ is twice continuously differentiable, a number of estimation methods in the existing varying-coefficient literature can be used. There are three main approaches to estimate smooth function $m(\cdot)$: kernel local polynomial smoothing (e.g., Wu et al., 1998; Fan and Zhang, 1999), polynomial splines (e.g., Huang et al., 2002, 2004), or spline smoothing (e.g., Hoover et al., 1998). In this chapter, we opt for the local constant smoothing. The local constant estimator $\hat{m}_T(z)$ of $m(z)$ is the minimizer of

$$\min_{a \in \mathbb{R}} \sum_{t=1}^{T} [\tilde{y}_t - \tilde{x}_t \cdot a]^2 K_h(z_t - z),$$

where $K_h(v) = K(v/h_T)/h_T$, $K(v)$ is a symmetric kernel function, and $h_T > 0$ is the bandwidth that $h_T \to 0$ as $T \to +\infty$. Solving the first-order conditions leads to

$$\hat{m}_T(z) = \left\{ \frac{1}{T} \sum_{t=1}^{T} \tilde{x}_t \tilde{x}_t^\top K_h(z_t - z) \right\}^{-1} \frac{1}{T} \sum_{t=1}^{T} \tilde{x}_t \tilde{y}_t K_h(z_t - z). \tag{3.11}$$

The local linear estimator can be defined similarly.

Although the local constant or local linear smoothers are sufficient for consistent estimation of the slope parameter β^0 even if the transition function contains a finite number of discontinuities (see Section 3.4), the estimation of transition function $w^0(\cdot)$ will possibly suffer. Unfortunately, there is a rather limited research on the nonparametric estimation of functions with discontinuities in the context of varying-coefficient models. If the transition function and thus function $m(\cdot)$ possibly contain discontinuities, we suggest to use the estimation procedure of Čížek and Koo (2017a). Its short description follows.

Let the conventional kernel function be $K^{(c)}(v) = K(v)$, where $K(v)$ is a symmetric kernel with compact support $[-1, 1]$, and the left-sided and right-sided kernels be $K^{(l)}(v) = K(v) \cdot \mathbf{1}(v \in (-1, 0))$ and $K^{(r)}(v) = K(v) \cdot \mathbf{1}(v \in [0, 1))$, respectively. Using these three kernels, three local constant estimates of $m(z)$ can be constructed:

$$\hat{a}^{(j)}(z) = \arg\min_{a \in \mathbb{R}} \sum_{t=1}^{T} [\tilde{y}_t - \tilde{x}_t \cdot a]^2 K_h^{(j)}(z_t - z), \quad j = l, r, c,$$

where superscripts l, r, and c indicate whether the left, right, and two-sided neighborhood of z is used, respectively. The goodness of fit of the three estimates can be measured, for example, by the weighted residual mean squared error (WRMSE) defined by

$$\text{WRMSE}^{(j)}(z) = \frac{\sum_{t=1}^{T} [\tilde{y}_t - \tilde{x}_t \cdot \hat{a}^{(j)}(z)]^2 K_h^{(j)}(z_t - z)}{\sum_{t=1}^{T} K_h^{(j)}(z_t - z)}, \quad j = l, r, c.$$

If $m(z_t)$ is continuous around z, all three WRMSEs are consistent estimates of $\mathrm{E}[\varepsilon_t^2 | z_t = z]$, while $\text{WRMSE}^{(l)}(z)$ and $\text{WRMSE}^{(r)}(z)$ are the only consistent estimates of $\mathrm{E}[\varepsilon_t^2 | z_t = z]$ for z in the left and right h_T-neighborhoods of a point of discontinuity, respectively (cf. Proposition 2.2 in Gijbels et al., 2007, and Theorem 4 in Čížek and Koo, 2017a). Since $\mathrm{E}[\varepsilon_t^2 | z_t = z]$ represents asymptotically the smallest value of WRMSE given model (3.10), the jump-preserving estimator of function $m(\cdot)$ can be defined by

$$\hat{m}_T(z) = \begin{cases} \hat{a}^{(c)}(z), & \text{if } \text{diff}(z) \le u_T, \\ \hat{a}^{(l)}(z), & \text{if } \text{diff}(z) > u_T \text{ and } \text{WRMSE}^{(l)}(z) < \text{WRMSE}^{(r)}(z), \\ \hat{a}^{(r)}(z), & \text{if } \text{diff}(z) > u_T \text{ and } \text{WRMSE}^{(l)}(z) > \text{WRMSE}^{(r)}(z), \\ \dfrac{\hat{a}^{(l)}(z) + \hat{a}^{(r)}(z)}{2}, & \text{if } \text{diff}(z) > u_T \text{ and } \text{WRMSE}^{(l)}(z) = \text{WRMSE}^{(r)}(z), \end{cases} \tag{3.12}$$

where $\text{diff}(z) = \text{WRMSE}^{(c)}(z) - \min\{\text{WRMSE}^{(l)}(z), \text{WRMSE}^{(r)}(z)\}$ and the threshold parameter $u_T > 0$ is such that $u_T \to 0$ as $T \to +\infty$ (u_T can be determined along with h_T by the least-squares cross-validation, see Čížek and Koo, 2017a, Section 5, for details).

3.3.3 Least squares estimator of $\beta(w)$

Given a transition function w, the SETR model (3.3) is linear in the slope β. Hence, the ordinary LS estimation can be directly applied. Recall that $\omega_t = [1 - w(z_t), w(z_t)]^\top$. Similarly to (3.8), the sum of squared residuals $T^{-1}\sum_{t=1}^{T}\{y_t - (\omega_t \otimes x_t)'\beta\}^2$ is minimized with respect to β while w is fixed, which yields the conditional LS estimator

$$\hat{\beta}_T(w) = \left\{\frac{1}{T}\sum_{t=1}^{T}(\omega_t \otimes x_t)(\omega_t \otimes x_t)^\top\right\}^{-1}\frac{1}{T}\sum_{t=1}^{T}(\omega_t \otimes x_t)y_t. \tag{3.13}$$

3.3.4 The proposed algorithm

The estimation algorithm can be summarized as follows. Given the sets of 2κ intervals $(a_1^{(k)}, b_1^{(k)})$ and $(a_2^{(k)}, b_2^{(k)})$, $k = 1, \ldots, \kappa$, from the interval sequences defined in Section 3.3.1,

1. estimate $\hat{\beta}_{1,T}^{(0,k)}$ and $\hat{\beta}_{2,T}^{(0,k)}$ by LS using data points with $z_t \in (a_1^{(k)}, b_1^{(k)})$ and $z_t \in (a_2^{(k)}, b_2^{(k)})$, respectively, for all $k = 1, \ldots, \kappa$
2. given $\hat{\beta}_T^{(0,k)} = (\hat{\beta}_{1,T}^{(0,k)\top}, \hat{\beta}_{2,T}^{(0,k)\top})^\top$, estimate $\hat{w}_T^{(0,k)}(z_t, \hat{\beta}_T^{(0,k)})$ by (3.12) at $z_t \notin (a_1^{(k)}, b_1^{(k)}) \cup (a_2^{(k)}, b_2^{(k)})$ and set $\hat{w}_T^{(0,k)}(z_t, \hat{\beta}_T^{(0,k)}) = 0$ or 1 for $z_t \in (a_1^{(k)}, b_1^{(k)})$ or $z_t \in (a_2^{(k)}, b_2^{(k)})$, respectively; $t = 1, \ldots, T$, $k = 1, \ldots, \kappa$
3. evaluate sums $S_{(k)}^2$ of squared residuals (3.9) at $\hat{\beta}_T^{(0,k)}$, $\hat{w}_T^{(0,k)}$ for all $k = 1, \ldots, \kappa$ and set $\hat{k} = \arg\min_{k=1,\ldots,\kappa} S_{(k)}^2$, $\hat{\beta}_T^{(0)} = \hat{\beta}_T^{(0,\hat{k})}$, and $\hat{w}_T^{(0)} = \hat{w}_T^{(0,\hat{k})}$
4. given $\hat{w}_T^{(0,k)}$, estimate $\hat{\beta}_{1,T}^{(1,k)}(\hat{w}_T^{(0,k)})$ and $\hat{\beta}_{2,T}^{(1,k)}(\hat{w}_T^{(0,k)})$ by LS (3.13) for $k = 1, \ldots, \kappa$
5. given $\hat{\beta}_T^{(1,k)} = (\hat{\beta}_{1,T}^{(1,k)\top}, \hat{\beta}_{2,T}^{(1,k)\top})^\top$, estimate $\hat{w}_T^{(1,k)}(z_t, \hat{\beta}_T^{(1,k)})$ by (3.12) at $z_t \notin (a_1^{(k)}, b_1^{(k)}) \cup (a_2^{(k)}, b_2^{(k)})$ and set $\hat{w}_T^{(1,k)}(z_t, \hat{\beta}_T^{(1,k)}) = 0$ or 1 for $z_t \in (a_1^{(k)}, b_1^{(k)})$ or $z_t \in (a_2^{(k)}, b_2^{(k)})$, respectively; $t = 1, \ldots, T$, $k = 1, \ldots, \kappa$
6. evaluate sums $S_{(k)}'^2$ of squared residuals (3.9) at $\hat{\beta}_T^{(1,k)}$, $\hat{w}_T^{(1,k)}$ for all $k = 1, \ldots, \kappa$ and set $\hat{k}' = \arg\min_{k=1,\ldots,\kappa} S_{(k)}'^2$
7. define the final estimates by $\hat{\beta}_T^{(1)} = \hat{\beta}_T^{(1,\hat{k}')}$ and $\hat{w}_T^{(1)} = \hat{w}_T^{(1,\hat{k}')}$

Note that step 3 just introduces notation for the estimates after the initial steps 1 and 2 and can thus be omitted and that steps 4 and 5 can be iterated as discussed in the introduction of Section 3.3.

3.4 Asymptotic properties

In the asymptotic analysis, we consider absolutely regular time series and transition functions from $\mathcal{W}$ constrained to piecewise smooth functions.

First, the definition of an absolutely regular (or β-mixing) process is provided. Consider a strictly stationary process $\{x_t\}_{t=1}^{\infty}$ and let $\mathcal{F}_k^l$ be the σ-algebra generated by $\{x_t\}_{t=k}^{l}$. The β-mixing coefficients are defined by

$$\beta(m) = \sup_{t\in\mathbb{N}} \mathrm{E}[\sup_{A\in\mathcal{F}_{t+m}^{\infty}} \left|P(A|\mathcal{F}_1^t) - P(A)\right|].$$

If $\lim_{m\to\infty} \beta(m) = 0$, the process $\{x_t\}_{t=1}^{\infty}$ is called β-mixing or absolutely regular.

Next, let us define the class of smooth functions $C_M^{\gamma}(\mathcal{X})$ on a bounded set $\mathcal{X} \subset \mathbb{R}^d$ (e.g., $[s_{J-1}, s_J]$ in Assumption 3.A) following van der Vaart and Wellner (1996, p. 154); see also Ichimura and Lee (2010). Let $\underline{\gamma}$ be the largest integer smaller than γ, and for any vector $k = (k_1, \ldots, k_d) \in \mathbb{N}^d$, let the differential operator $D^k = \frac{\partial^{|k|}}{\partial x_1^{k_1} \ldots \partial x_d^{k_d}}$ for $|k| = \sum_{i=1}^d k_i$. Additionally, define the function norm

$$\|f\|_{\gamma} = \max_{|k|\leq\underline{\gamma}} \sup_x \left|D^k f(x)\right| + \max_{|k|=\underline{\gamma}} \sup_{x\neq x'} \frac{\left|D^k f(x) - D^k f(x')\right|}{\|x - x'\|^{\gamma-\underline{\gamma}}},$$

where the suprema are taken over all x and x' in the interior of $\mathcal{X}$. Then $C_M^{\gamma}(\mathcal{X})$ is the set of all continuous functions $f : \mathcal{X} \mapsto \mathbb{R}$ with $\|f\|_{\gamma} \leq M$.

Using the above notation, the following assumptions are introduced to prove the consistency of the estimators proposed in Section 3.3.

Assumption 3.B. Let the random variables x_t, z_t, and ε_t, and random vector $v_t = (z_t, v_{2t}, v_{3t})^{\top}$ with v_{2t} and v_{3t} representing any element of vectors x_t and $(x_t^{\top}, \varepsilon_t)^{\top}$, respectively, satisfy the following conditions.

3.B1. The process $\{x_t, z_t, \varepsilon_t\}_{t=1}^T$ is strictly stationary and absolutely regular with β-mixing coefficients $\beta(m)$ such that $\beta(m) = \mathrm{o}(m^{-(2+\xi)/\xi})$ as $m \to +\infty$ for some $\xi > 0$.

3.B2. The following moments are finite: $\mathrm{E}\left\|x_t x_t^\top\right\|^{2+\xi} < \infty$, $\mathrm{E}\left\|\varepsilon_t x_t\right\|^{2+\xi} < \infty$, $\mathrm{E}\left|z_t\right|^{2+\xi} < \infty$, and $\mathrm{E}\left|\varepsilon_t\right|^{2+\xi} < \infty$, where ξ is given in Assumption 3.B1.

3.B3. Assuming that the support $\mathcal{Z}$ of the variable z_t is partitioned into bounded and convex sets I_j with nonempty interiors, $\mathcal{Z} = \bigcup_{j=1}^\infty I_j$, the space $\mathcal{W}$ of transition functions contains only piecewise continuous functions such that, after restricting them to I_j, $\mathcal{W}_{|I_j}$ belongs to $C_M^\gamma(I_j)$ for some $\gamma > 3$ and $j \in \mathbb{N}$.

3.B4. Finally, let $\sum_{j=1,k,l=-\infty}^{\infty,\infty} \max\{\lambda(I_{jkl}^3), 1\} \cdot \max I_{jkl}^3 \cdot Q^{[(1+\delta)(3+\xi)]^{-1}}(I_{jkl}^3)$ be finite for some $\delta > 0$, where the partition of $\mathbb{R}^3 = \bigcup_{j=1,k,l=-\infty}^{\infty,\infty} I_{jkl}^3$ is defined by $I_{jkl}^3 = I_j \times [k, k+1) \times [l, l+1)$, $\lambda(I_{jkl}^3)$ denotes the Lebesque measure of I_{jkl}^3, $Q(I_{jkl}^3) = P(v_t \in I_{jkl}^3)$, and $\max I_{jkl}^3 = \sup_{v=(v_1,v_2,v_3)^\top \in I_{jkl}^3} \max\{|v_1|, |v_2|, |v_3|\}$.

If $\{x_t, z_t, \varepsilon_t\}_{t=1}^T$ is a series of independent random vectors, Assumption 3.B1 is automatically fulfilled. Under dependence, the stationarity condition in Assumption 3.B1 excludes time trends and integrated processes. Additionally, the mixing condition in Assumption 3.B1 controls the degree of dependence in the process $\{x_t, z_t, \varepsilon_t\}_{t=1}^T$ and is a standard assumption to guarantee the validity of the stochastic limit theorems. Sufficient conditions such that the nonlinear autoregressive models (which contain the TAR, STAR, and the semiparametric transition model for measurable transition functions w) are geometrically ergodic and thus β-mixing under Assumption 3.B1 can be found in Chen and Tsay (1993) and Meitz and Saikkonen (2010). Specifically, if the support of continuously distributed innovations ε_t spans $\mathbb{R}$, $m_w = \inf_{z\in\mathbb{R}} w(z)$, $M_w = \sup_{z\in\mathbb{R}} w(z)$, and $M_j = \max\{|\beta_{1j}^0(1 - M_w) + \beta_{2j}^0 M_w|, |\beta_{1j}^0(1 - m_w) + \beta_{2j}^0 m_w|\}$, $j = 1, \ldots, p$, Theorem 1.1 of Chen and Tsay (1993) applied to the autoregressive model (3.3) with $x_t = (y_{t-1}, \ldots, y_{t-p})^\top$ and $z_t = y_{t-d}$ for some $p, d \in \mathbb{N}$ states the sufficient condition for the geometric ergodicity: all roots of the characteristic equation $z^p - M_1 z^{p-1} - \ldots - M_p z^0 = 0$ have to lie inside the unit circle. In the most typical case of a transition function w restricted to $[0, 1]$, $M_j = \max\{|\beta_{1j}^0|, |\beta_{2j}^0|\}$. In the simplest case of $p = 1$ and $w : \mathbb{R} \to [0, 1]$, the sufficient condition is then $\max\{|\beta_{11}^0|, |\beta_{21}^0|\} < 1$.

Furthermore, Assumption 3.B2 requires that a sufficient number of moments exists. Assumption 3.B2 together with Assumption 3.B1 are essential to guarantee the validity of the law of large numbers (LLN) and the central limit theorem (CLT) for dependent sequences (e.g., Arcones and Yu, 1994, and Davidson, 1994, Section 24.4). Assumption 3.B3 defines a class of functions such that LLN can be applied uniformly to this class of functions (cf. van der Vaart and Wellner, 1996, Sections 2.7 and 2.8). The transition functions have to be piecewise smooth and at least three times differentiable in the continuity regions. Finally, Assumption 3.B4 is a technical assumption used again for the uniform LLN and it is also sufficient if it holds with another partitioning than the given one. It does not restrict variables with a bounded support, which are commonly used or imposed by means of trimming in semiparametric literature. For variables with an infinite support, it requires that the probability of observing large values are small. To facilitate an easier understanding, consider the univariate equivalent of Assumption 3.B4: $\sum_{j=1}^{\infty} \max\{\lambda(I_j), 1\} \cdot \max I_j \cdot Q^{[(1+\delta)(3+\xi)]^{-1}}(I_j)$. As intervals I_j can be chosen of the maximum length 1 without loss of generality, the sum is bounded by $\sum_{j=1}^{\infty} |j+1| \cdot \{Q^{[(1+\delta)(3+\xi)]^{-1}}([j, +\infty)) + Q^{[(1+\delta)(3+\xi)]^{-1}}([-\infty, -j])\}$. Considering case of small $\xi > 0$ so that $(1+\delta)(3+\xi) < 3.5$, this bound is finite if the distribution of random variable v_t has tails decreasing to zero proportionally to or faster than $1/j^7$, for instance. This assumption can be further weakened (along with the order of differentiability) if the error term ε_t is independent of the transition variable z_t.

The following theorem establishes the consistency of the unconditional estimators $\hat{\beta}_T$ and $\hat{w}_T$. This guarantees that minimizing the LS criterion (3.9) with respect to both $\beta \in \mathcal{B}$ and $w \in \mathcal{W}$ leads to consistent estimates.

Theorem 3.3. *Under Assumptions 3.A and 3.B, it holds as $T \to +\infty$ that*

$$\hat{\beta}_T \xrightarrow{P} \beta^0, \quad \|\hat{w}_T - w^0\|_{\infty,\epsilon,f_z} \xrightarrow{P} 0 \text{ for any } \epsilon > 0, \quad \text{and} \quad \mathrm{E}\{\hat{w}_T(z_t) - w^0(z_t)\}^2 \to 0.$$

Before deriving the asymptotic properties of the conditional estimators $\hat{w}_T(z, \check{\beta}_T)$ and $\hat{\beta}_T(\check{w}_T)$ and of the proposed algorithm, it is necessary to impose some conditions on the nonparametric estimator of $w(z, \beta)$ in the varying-coefficient model (3.10).

Assumption 3.C. Let $\zeta_T > 0$ such that $\zeta_T \to 0$ as $T \to +\infty$, $\mathcal{Z}_T^c$ be a subset of the support $\mathcal{Z}$ of the transition variable z_t excluding all ζ_T-neighborhoods of discontinuities $\{s_j\}_{j=1}^J$, $\mathcal{Z}_T^c = \mathcal{Z} \setminus \bigcup_{j=1}^J (s_j - \zeta_T, s_j + \zeta_T)$, and $U(\beta^0, \delta) = \{\beta \in \mathcal{B} : \|\beta - \beta^0\| < \delta\}$. It is assumed that there exist some sequence ζ_T and $\delta > 0$ such that, for any $\beta \in U(\beta^0, \delta)$ and $0 < \tilde{\delta} < \delta$,

3.C1. $P\{\hat{w}_T(z, \beta) \in \mathcal{W}\} \to 1$ as $T \to +\infty$;

3.C2. estimator $\hat{w}_T(z, \beta)$ is uniformly bounded on $\mathcal{Z} \times \mathcal{B}$ and uniformly consistent on $\mathcal{Z}_T^c$: $\sup_{z \in \mathcal{Z}_T^c} |\hat{w}_T(z, \beta) - w_T(z, \beta)| \xrightarrow{P} 0$ as $T \to +\infty$ for any $\beta \in U(\beta^0, \delta)$;

3.C3. estimator $\hat{w}_T(z, \beta)$ is stochastically equicontinuous at β^0 on $\mathcal{Z}_T^c$: $\sup_{z \in \mathcal{Z}_T^c} \sup_{\beta \in U(\beta^0, \delta)} \sup_{\tilde{\beta} \in U(\beta, \tilde{\delta})} \left| \hat{w}_T(z, \beta) - \hat{w}_T(z, \tilde{\beta}) \right| \xrightarrow{P} 0$ as $T \to +\infty$ and $\tilde{\delta} \to 0$;

3.C4. function $w(z, \beta)$ has a uniformly bounded derivative with respect to $\beta \in U(\beta^0, \delta)$ on $\mathcal{Z}_T^c$: $\sup_{z \in \mathcal{Z}_T^c} \sup_{\beta \in U(\beta^0, \delta)} \|\partial w(z, \beta)/\partial \beta\| < \infty$;

3.C5. the density of z_t is bounded on $\mathcal{Z}$.

While Assumptions 3.C4 and 3.C5 are additional regularity conditions, Assumptions 3.C1–3.C3 are relevant to the properties of the conditional estimator $\hat{w}_T(z, \beta)$ of the transition function, and for the jump-preserving varying-coefficient estimator suggested in Section 3.3.2, are therefore verified in Section 3.9. In their general form, Assumptions 3.C1–3.C3 provide conditions for other nonparametric estimators $\hat{w}_T(z, \beta)$ that can be applied to the univariate varying-coefficient model (3.10), where the response variable $\tilde{y}_t = y_t - x_t^\top \beta_1$ and explanatory variable $\tilde{x}_t = x_t^\top (\beta_2 - \beta_1)$ for some fixed slopes β_1 and β_2. First, the estimate $\hat{w}_T(z, \beta)$ is supposed to converge to a function from the function space $\mathcal{W}$ in Assumption 3.C1 as is common in semiparametric literature (e.g., Ichimura and Lee, 2010). Next, Assumption 3.C2 requires the nonparametric estimator to be uniformly consistent. This condition is typically satisfied on compact subsets of $\mathbb{R}$, but can be extended to $\mathbb{R}$ for bounded functions. In the examples discussed in Sections 3.2 and 3.3.1, transition functions are always estimated on a compact set since they are assumed to be 0 or 1 outside of a sufficiently large compact set and are thus not estimated there.

Finally, the nonparametric estimator $\hat{w}_T(\cdot, \beta)$ is required to be stochastically equicontinuous by Assumption 3.C3 as in Ichimura and Lee (2010), who argue that this restriction holds for estimators $\hat{w}_T(z, \beta)$ continuously differentiable in $\beta \in U(\beta^0, \delta)$.

In the following theorems, the consistency and asymptotic distribution of the estimators proposed in Section 3.3 is derived. The estimation starts with the initial estimate $\hat{\beta}_T^{(0)}$, which is shown to be consistent if at least one considered pair of intervals satisfies Assumption 3.A4. Based on a consistent estimator $\check{\beta}_T$ such as $\hat{\beta}_T^{(0)}$ or any subsequent iterations $\hat{\beta}_T^{(k)}$, the transition function is estimated by $\hat{w}_T(z, \check{\beta}_T)$, which is proved to be asymptotically equivalent to the infeasible estimate $\hat{w}_T(z, \beta^0)$ based on the true slope β_0, and subsequently, to be a consistent estimator of transition function $w(z) = w(z, \beta^0)$.

Theorem 3.4. *Suppose that the algorithm defined in Section 3.3.4 employs intervals* (a_1^k, b_1^k) *and* (a_2^k, b_2^k), $k = 1, \dots, \kappa$. *If Assumptions 3.A–3.C are satisfied and there is at least one* $k^* \in \{1, \dots, \kappa\}$ *and a pair of intervals* $(a_1^{k^*}, b_1^{k^*}) \subseteq (a_1, b_1)$ *and* $(a_2^{k^*}, b_2^{k^*}) \subseteq (a_2, b_2)$ *satisfying Assumption 3.A4, then it holds as* $T \to +\infty$

1. *the LS estimator* $\hat{\beta}_T^{(0,k^*)}$ *based on the* k^**th pair of intervals is consistent,* $\hat{\beta}_T^{(0,k^*)} \xrightarrow{P} \beta^0$;
2. *for any* $\check{\beta}_T \xrightarrow{P} \beta^0$ *such as* $\hat{\beta}_T^{(0,k^*)}$, $\sup_{z \in \mathcal{Z}_T^c} \left|\hat{w}_T(z, \check{\beta}_T) - \hat{w}_T(z, \beta^0)\right| \xrightarrow{P} 0$, $\sup_{z \in \mathcal{Z}_T^c} \left|\hat{w}_T(z, \check{\beta}_T) - w(z, \beta^0)\right| \xrightarrow{P} 0$, *and* $\mathrm{E}[\hat{w}_T(z, \check{\beta}_T) - w(z, \beta^0)]^2 \to 0$;
3. *the initial estimator is consistent,* $\hat{\beta}_T^{(0)} = \hat{\beta}_T^{(0,\hat{k})} \xrightarrow{P} \beta^0$, *and for* $\hat{w}_T^{(0)}(z_t) = \hat{w}_T(z_t, \hat{\beta}_T^{(0)})$, $\sup_{z \in \mathcal{Z}_T^c} \left|\hat{w}_T^{(0)}(z_t) - w(z, \beta^0)\right| \xrightarrow{P} 0$ *and* $\mathrm{E}[\hat{w}_T^{(0)}(z_t) - w(z, \beta^0)]^2 \to 0$.

Theorem 3.4 establishes the consistency of the initial estimators $\hat{\beta}_T^{(0)}$ and $\hat{w}_T^{(0)}$. The next step of the estimation procedure is based on a consistent estimate $\check{w}_T$ of the transition function such as $\check{w}_T = \hat{w}_T(\cdot, \hat{\beta}_T^{(0)})$ or later iterations $\check{w}_T = \hat{w}_T(\cdot, \hat{\beta}_T^{(k)})$: based on a transition function $\check{w}_T$, the slope parameters can be re-estimated by $\hat{\beta}_T(\check{w}_T)$. To derive their consistency and limiting distribution, the matrices entering the asymptotic variance of the slope estimator are introduced.

Assumption 3.D. Let $\omega_t^0 = [1 - w^0(z_t), w^0(z_t)]^\top$ and the matrices

$$Q^0 = \mathrm{E}[(\omega_t^0 \otimes x_t)(\omega_t^0 \otimes x_t)^\top] \quad \text{and} \quad V^0 = \mathrm{E}[\varepsilon_t^2 (\omega_t^0 \otimes x_t)(\omega_t^0 \otimes x_t)^\top]$$

be finite and positive definite.

Assumption 3.D corresponds to the usual full-rank condition in the least squares theory. With Assumptions 3.A–3.D, we first claim that the difference between the feasible slope estimator $\hat{\beta}_T(\check{w}_T)$ and the infeasible estimator $\hat{\beta}_T(w^0)$, based on the true transition function w^0, converges to zero in probability at a rate faster than $T^{-1/2}$. The consistency of the iterated estimators $\hat{\beta}_T^{(1)}$ and $\hat{w}_T^{(1)}$ (see Section 3.3.4) then immediately follows.

Theorem 3.5. *If Assumptions 3.A–3.D hold and estimator* $\check{w}_T$ *satisfies* $\mathrm{E}[\check{w}_T(z_t) - w^0(z_t)]^2 \to 0$ *as* $T \to +\infty$*, then it holds for* $T \to +\infty$ *that*

$$\sqrt{T}(\hat{\beta}_T(\check{w}_T) - \hat{\beta}_T(w^0)) \xrightarrow{P} 0,$$

and consequently under the assumptions of Theorem 3.4, $\hat{\beta}_T^{(1)} = \hat{\beta}_T^{(1,\hat{k}')} \xrightarrow{P} \beta^0$*, and for* $\hat{w}_T^{(1)}(z_t) = \hat{w}_T(z_t, \hat{\beta}_T^{(1)})$*,* $\sup_{z \in \mathcal{Z}_T^c} \left|\hat{w}_T^{(1)}(z_t) - w(z, \beta^0)\right| \xrightarrow{P} 0$ *and* $\mathrm{E}[\hat{w}_T^{(1)}(z_t) - w(z, \beta^0)]^2 \to 0$.

Finally, the limiting distribution of the infeasible estimator $\hat{\beta}_T(w^0)$ (assuming known w^0) is derived in Theorem 3.6, and by Theorem 3.5, this distribution describes asymptotically also the feasible estimator $\hat{\beta}_T(\check{w}_T)$.

Theorem 3.6. *Under Assumptions 3.A–3.D,* $\sqrt{T}\{\hat{\beta}_T(w^0) - \beta^0\} \xrightarrow{d} N(0, Q^{0^{-1}} V^0 Q^{0^{-1}})$ *as* $T \to +\infty$*. Additionally, if an estimator* $\check{w}_T$ *satisfies* $\mathrm{E}[\check{w}_T(z_t) - w^0(z_t)]^2 \to 0$*, then it holds for* $T \to +\infty$

$$\sqrt{T}(\hat{\beta}_T(\check{w}_T) - \beta^0) \xrightarrow{d} N(0, Q^{0^{-1}} V^0 Q^{0^{-1}}).$$

The asymptotic variance of the feasible slope estimator corresponds to the variance of the linear least-squares estimator of model (3.3) with a known transition w^0. In practice, the asymptotic variance in Theorem 3.6 can be estimated directly by taking the finite sample equivalents of Q^0 and V^0 since a consistent estimate of w^0 is obtained as a part of the estimation procedure. In particular, if the estimation stops after K steps, one can define $\hat{\omega}_t = [1 - \hat{w}_T(z_t, \hat{\beta}_T^{(K)}), \hat{w}_T(z_t, \hat{\beta}_T^{(K)})]^\top$ and $\hat{\varepsilon}_t = y_t - (\hat{\omega}_t \otimes x_t)^\top \hat{\beta}_T^{(K)}$ and estimate Q^0 and V^0 by $\hat{Q}_T = \frac{1}{T}\sum_{t=1}^T (\hat{\omega}_t \otimes x_t)(\hat{\omega}_t \otimes x_t)^\top$ and $\hat{V}_T = \frac{1}{T}\sum_{t=1}^T \hat{\varepsilon}_t^2 (\check{\omega}_t \otimes x_t)(\check{\omega}_t \otimes x_t)^\top$. Finally, if inference on the slope estimates needs to be complemented by inference concerning the transition function estimate $\hat{w}_T$, the results in Čížek and Koo (2017a) can be used.

3.5 Simulation study

In this section, the performance of the proposed estimator of the semiparametric transition model is evaluated by Monte Carlo simulations. These simulations provide a comparison with the least squares (LS) estimators of the parametric TAR and LSTAR models.

Four different data generating processes (DGPs) are considered in this section. All these DGPs are based on the semiparametric transition autoregressive model of order 2:

$$y_t = [\beta_{1;0} + \beta_{1;1} y_{t-1} + \beta_{1;2} y_{t-2}]\{1 - w(z_t)\} + [\beta_{2;0} + \beta_{2;1} y_{t-1} + \beta_{2;2} y_{t-2}] w(z_t) + \varepsilon_t,$$

where the error term $\varepsilon_t \sim N(0,1)$ is independent and identically distributed and the values of regression coefficients used in the simulation are given by $(\beta_{1;0}, \beta_{1;1}, \beta_{1;2}) = (-0.25, 0.4, -0.6)$ and $(\beta_{2;0}, \beta_{2;1}, \beta_{2;2}) = (0.25, -0.8, 0.2)$. The functional forms of the transition function $w(z_t)$ and their arguments are listed below ($U(0,1)$ denotes the uniform distribution on interval $[0,1]$):

DGP1a $w(z) = \mathbf{1}(z > \tau)$ with $\tau = 0.4$ and $z_t = y_{t-2}$;

DGP1b $w(z) = \mathbf{1}(z > \tau)$ with $\tau = 0.4$ and $z_t = t/T$, where $t = 1, \ldots, T$;

DGP2 $w(z) = [1 + \exp\{-\nu(z - \tau)\}]^{-1}$ with $\nu = 2$, $\tau = 0.4$, and $z_t = y_{t-2}$;

DGP3 $w(z) = 0.5[1 - \cos\{4\pi(z - 0.1)\}]\mathbf{1}(z \in [0.1, 0.85]) + \mathbf{1}(z > 0.85)$ and $z_t \sim U(0,1)$ is independent and identically distributed;

DGP4 $w(z) = (z^{-1/2} - 1)\mathbf{1}(z \in [0.2, 0.7]) + \mathbf{1}(z > 0.7)$ and $z_t \sim U(0,1)$ is independent and identically distributed.

The DGP1a is a TAR model, where its transition function is piecewise constant with a discontinuity at 0.4. Although the case of deterministic transition variable z_t is not in the focus of this chapter, DGP1b replicates DGP1a for the case of z_t being a time fraction, which violates Assumption 3.B1. The DGP2 corresponds to the standard LSTAR model, where the shape parameter $\nu = 2$ so that the logistic function is flat enough to be distinguished from the indicator function of DGP1. While DGP1a and DGP2 use the lagged dependent variable y_{t-2} in the role of the transition variable, the last two DGP3

and DGP4 rely on a uniformly distributed transition variable independent of ε_s and y_{s-1}, $s \leq t$, and moreover, they are not nested in neither the TAR nor LSTAR models. The transition function in DGP3 is continuous and reaches both regimes two times (see Figure 3.3), whereas the transition function in DGP4 is discontinuous with two jumps (see Figure 3.4). In all cases, the order of the baseline autoregressive process is 2 and is assumed to be known.

For each data-generating process, 1000 samples of sizes $T = 200$, 400, and 800 are generated and estimated by the TAR, LSTAR, and the semiparametric transition (SETR) procedure, where the transition function is estimated by the local-constant estimator (3.11) of varying-coefficient model (3.10) assuming continuity of w (SETR/C) or by the jump-preserving local-constant estimator (3.12) described in Section 3.3.2 for piecewise smooth functions w with jumps (SETR/J). In both cases, the quartic kernel is used and the bandwidth h_T and procedure parameter u_T in (3.12) are determined by the least squares leave-one-out cross-validation. The proposed SETR estimation uses four pairs of initial estimators (for each of the two regimes), which are based on the data below the αth quantile and above the $(1-\alpha)$th quantile of the transition variable z_t for $\alpha = 0.05, 0.10, 0.20,$ and 0.40. Furthermore, the estimation involves two iterations: (i) based on the initial estimates $\hat{\beta}_T^{(0)}$, the transition function $\hat{w}_T^{(0)}(z) = \hat{w}_T(z, \hat{\beta}_T^{(0)})$ is estimated; (ii) using $\hat{w}_T^{(0)}$, the LS estimate $\hat{\beta}_T^{(1)} = \hat{\beta}_T(\hat{w}_T^{(0)})$ is obtained and then $\hat{w}_T^{(1)}(z) = \hat{w}_T(z, \hat{\beta}_T^{(1)})$ is computed given $\hat{\beta}_T^{(1)}$; as the initial estimators $\hat{\beta}_T^{(0)}$ are typically rather imprecise, the procedure is repeated again so that (iii) based on the estimates $\hat{\beta}_T^{(1)}$ and $\hat{w}_T^{(1)}$, the corresponding estimates of the slope and transition, $\hat{\beta}_T^{(2)} = \hat{\beta}_T(\hat{w}_T^{(1)})$ and $\hat{w}_T^{(2)}(z) = \hat{w}_T(z, \hat{\beta}_T^{(2)})$, are estimated and reported (see Section 3.3 for details). Regarding the TAR and LSTAR models, their transition location and shape parameters τ and ν are determined by a grid search to minimize their corresponding sums of squared residuals. All estimates are summarized by means of their biases and mean squared errors (MSE).

3.5.1 TAR results

The estimation results for the TAR model are summarized in Tables 3.1 and 3.2 for DGP1a and DGP1b, respectively; sample sizes cover $T = 200, 400,$ and 800. The TAR

TABLE 3.1: Biases and MSEs of all estimator for DGP1a and $T = 200, 400$, and 800.

		TAR		LSTAR		SETR/C		SETR/J	
T		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
200	$\hat{\beta}_{1,0}$	0.001	0.142	-0.008	0.149	-0.007	0.257	0.016	0.201
	$\hat{\beta}_{1,1}$	-0.009	0.078	-0.003	0.079	0.041	0.142	0.012	0.133
	$\hat{\beta}_{1,2}$	-0.004	0.133	-0.006	0.137	0.038	0.189	0.029	0.165
	$\hat{\beta}_{2,0}$	0.002	0.215	0.018	0.227	0.124	0.399	0.046	0.338
	$\hat{\beta}_{2,1}$	0.005	0.072	0.000	0.073	-0.010	0.123	0.005	0.128
	$\hat{\beta}_{2,2}$	-0.004	0.124	-0.010	0.127	-0.052	0.168	-0.023	0.148
400	$\hat{\beta}_{1,0}$	-0.004	0.093	-0.009	0.095	-0.024	0.162	0.010	0.115
	$\hat{\beta}_{1,1}$	-0.005	0.055	-0.002	0.055	0.049	0.110	0.007	0.088
	$\hat{\beta}_{1,2}$	-0.004	0.091	-0.005	0.091	0.025	0.125	0.016	0.103
	$\hat{\beta}_{2,0}$	0.008	0.149	0.014	0.150	0.136	0.287	0.029	0.214
	$\hat{\beta}_{2,1}$	0.005	0.052	0.002	0.052	-0.017	0.090	0.003	0.083
	$\hat{\beta}_{2,2}$	-0.004	0.083	-0.007	0.084	-0.048	0.118	-0.011	0.097
800	$\hat{\beta}_{1,0}$	-0.001	0.066	-0.003	0.066	-0.027	0.110	0.012	0.075
	$\hat{\beta}_{1,1}$	-0.001	0.038	0.000	0.038	0.045	0.090	-0.001	0.064
	$\hat{\beta}_{1,2}$	-0.002	0.063	-0.003	0.063	0.017	0.084	0.010	0.068
	$\hat{\beta}_{2,0}$	-0.003	0.102	-0.000	0.103	0.123	0.224	0.005	0.149
	$\hat{\beta}_{2,1}$	0.002	0.035	0.001	0.034	-0.020	0.068	0.002	0.058
	$\hat{\beta}_{2,2}$	-0.001	0.058	-0.001	0.058	-0.042	0.084	-0.002	0.066

and LSTAR estimates provide best and precise estimates as both correspond to the specified DGP: the grid for the transition shape parameter ν for LSTAR was reaching up to $\nu = 1000$ and the logistic transition function can thus became numerically identical to the discontinuous transition of TAR. Regarding the SETR estimation, both SETR/C and SETR/J provide consistent estimates in the sense that the biases and mean squared errors (MSE) decrease with an increasing sample size; the MSEs even support the $\sqrt{T}$ convergence rate of the semiparametric estimators in that the MSEs at $T = 800$ are approximately half of the MSEs at $T = 200$. It is however noticeable that the SETR/J, which accounts for the discontinuity of the transition function, exhibits much smaller biases than the SETR/C. The source of the SETR/C bias is visible on Figure 3.1, where the average of estimated transition functions is presented along with the corresponding 90% confidence bands. Whereas SETR/C estimates are significantly biased, SETR/J exhibits

TABLE 3.2: Biases and MSE of all estimator for DGP1b and $T = 200, 400$, and 800.

		TAR		LSTAR		SETR/C		SETR/J	
T		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
200	$\hat{\beta}_{1,0}$	-0.001	0.116	-0.004	0.117	-0.002	0.129	0.007	0.127
	$\hat{\beta}_{1,1}$	-0.005	0.077	0.001	0.079	0.000	0.136	-0.026	0.135
	$\hat{\beta}_{1,2}$	-0.002	0.070	-0.006	0.071	-0.010	0.101	0.006	0.096
	$\hat{\beta}_{2,0}$	0.003	0.100	0.005	0.101	-0.024	0.103	-0.007	0.102
	$\hat{\beta}_{2,1}$	-0.009	0.089	-0.012	0.090	0.052	0.127	0.022	0.115
	$\hat{\beta}_{2,2}$	-0.024	0.091	-0.027	0.092	0.012	0.105	-0.008	0.101
400	$\hat{\beta}_{1,0}$	0.000	0.082	-0.002	0.082	-0.006	0.094	0.001	0.092
	$\hat{\beta}_{1,1}$	-0.004	0.055	-0.001	0.056	0.007	0.100	-0.010	0.095
	$\hat{\beta}_{1,2}$	0.002	0.050	0.000	0.050	-0.007	0.073	0.003	0.069
	$\hat{\beta}_{2,0}$	0.007	0.071	0.008	0.071	-0.014	0.070	0.002	0.071
	$\hat{\beta}_{2,1}$	-0.004	0.065	-0.005	0.065	0.041	0.088	0.013	0.075
	$\hat{\beta}_{2,2}$	-0.012	0.066	-0.014	0.066	0.015	0.072	-0.004	0.069
800	$\hat{\beta}_{1,0}$	-0.001	0.055	-0.002	0.056	-0.006	0.064	-0.001	0.063
	$\hat{\beta}_{1,1}$	-0.001	0.040	0.001	0.040	0.009	0.072	-0.006	0.072
	$\hat{\beta}_{1,2}$	0.000	0.036	-0.001	0.036	-0.008	0.053	0.001	0.052
	$\hat{\beta}_{2,0}$	0.000	0.046	0.001	0.046	-0.017	0.048	-0.002	0.046
	$\hat{\beta}_{2,1}$	-0.002	0.046	-0.003	0.046	0.031	0.061	0.007	0.050
	$\hat{\beta}_{2,2}$	-0.007	0.046	-0.007	0.046	0.015	0.051	-0.002	0.048

much smaller bias and its confidence band includes the true transition function.

Contrary to the experiments DGP2–DGP4 presented later, the parametric estimates are more precise when comparing SETR/J to the parametric TAR and LSTAR estimates: the overall MSE of SETR (across the full vector of parameters) is approximately 10%–30% higher depending on the model and sample size. The main difference of the TAR model to other DGPs is that the TAR threshold parameter estimate converges to its true value at a rate faster than the regression parameters and thus practically does not influence their precision even in finite samples; this is not the case of SETR/J with a nonparametrically estimated transition function. On the other hand, the threshold and transition function estimates in all other models converge at most at $n^{-1/2}$ rate and the difference between parametric and semiparametric estimators will become negligible.

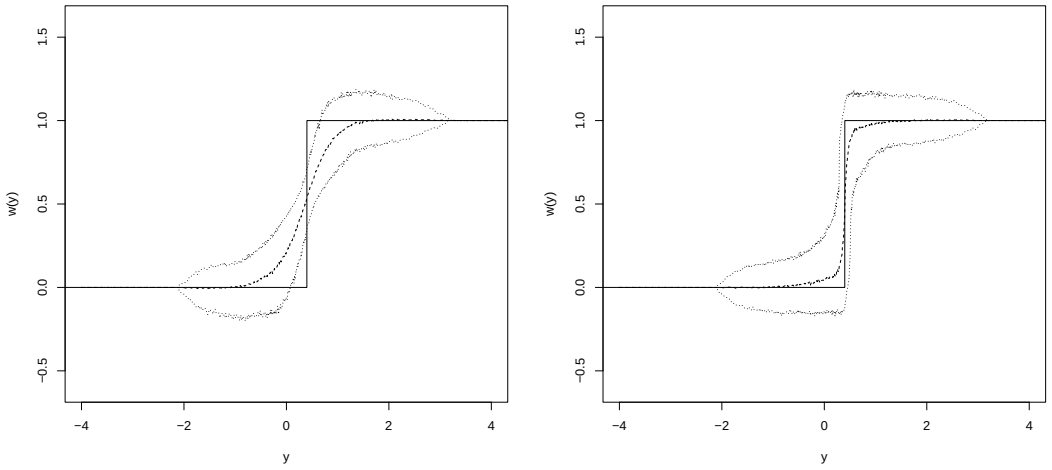


FIGURE 3.1: The mean estimates (dashed line) and 5% and 95% quantiles (dotted lines) of the transition function in DGP1a with $T = 400$; the solid line depicts the true transition function. The left and right panels correspond to SETR/C and SETR/J estimates, respectively.

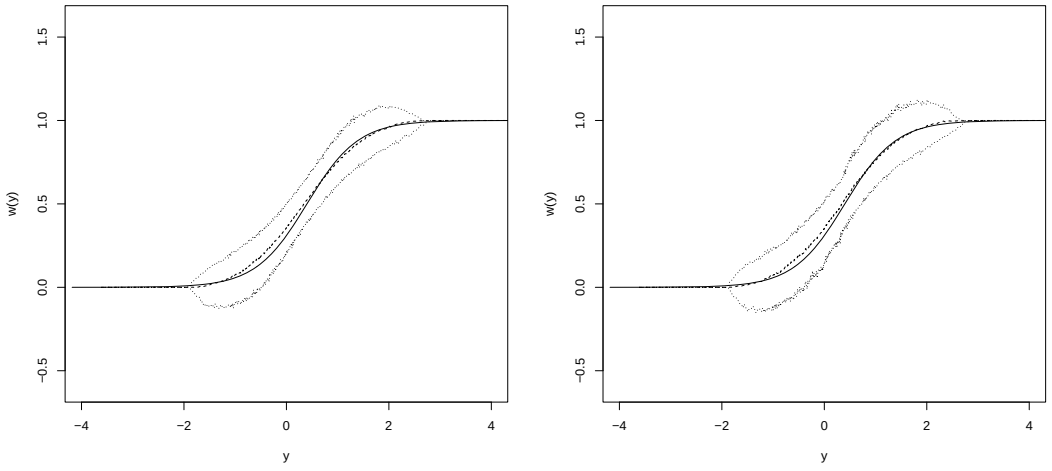


FIGURE 3.2: The mean estimates (dashed line) and 5% and 95% quantiles (dotted lines) of the transition function in DGP2 with $T = 400$; the solid line depicts the true transition function. The left and right panels correspond to SETR/C and SETR/J estimates, respectively.

Finally, note that the estimates are overall more precise in the case of DGP1b with the deterministic transition variable than in the case of DGP1a with the lagged dependent variable acting as the transition variable.

TABLE 3.3: Biases and MSE of all estimator for DGP2 and $T = 400$.

	TAR		LSTAR		SETR/C		SETR/J	
	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
$\hat{\beta}_{1,0}$	0.081	0.153	0.027	0.262	0.044	0.265	0.062	0.287
$\hat{\beta}_{1,1}$	-0.158	0.179	-0.004	0.118	0.024	0.113	0.021	0.115
$\hat{\beta}_{1,2}$	0.013	0.121	0.013	0.171	0.043	0.182	0.053	0.193
$\hat{\beta}_{2,0}$	-0.356	0.395	-0.004	0.451	-0.027	0.390	-0.057	0.423
$\hat{\beta}_{2,1}$	0.177	0.203	0.009	0.117	0.003	0.105	0.013	0.108
$\hat{\beta}_{2,2}$	0.102	0.143	0.007	0.171	0.011	0.162	0.020	0.174

3.5.2 LSTAR results

The estimation results for the LSTAR model are summarized in Tables 3.3, from now on only for $T = 400$. The LSTAR model and estimator provide now correct parametric specification and thus best results in terms of very small bias and MSE. On the other hand, TAR is misspecified, which manifests itself by relatively large bias of some parameter estimates. Further, both SETR/C and SETR/J provide consistent estimates with relatively small biases and MSEs, which are surprisingly close to those of LSTAR: the precision of the parametric and semiparametric estimation is on the same level. Since the transition function is now smooth, SETR/C is more precise than SETR/J, which accounts for the possible discontinuities of the transition function and provides thus slightly more noisy estimates of the transition function. The difference is not very large though as can be seen from the transition function estimates on Figure 3.2.

3.5.3 Cosinus function

Another example of a model with a continuous transition function is DGP3 with the corresponding estimation results in Tables 3.4 and the transition function estimates on Figure 3.3 (again for $T = 400$). In this case, both parametric models – TAR and LSTAR – are misspecified, which leads to substantial biases in both cases. On the other hand, both SETR/C and SETR/J provide consistent estimates with relatively small biases and

TABLE 3.4: Biases and MSE of all estimator for DGP3 and $T = 400$.

	TAR		LSTAR		SETR/C		SETR/J	
	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
$\hat{\beta}_{1,0}$	0.130	0.178	0.122	0.177	-0.003	0.095	-0.002	0.096
$\hat{\beta}_{1,1}$	-0.307	0.372	-0.282	0.364	0.033	0.124	0.031	0.126
$\hat{\beta}_{1,2}$	0.201	0.250	0.185	0.245	-0.023	0.096	-0.022	0.096
$\hat{\beta}_{2,0}$	-0.060	0.136	-0.053	0.137	0.008	0.091	0.007	0.092
$\hat{\beta}_{2,1}$	0.153	0.247	0.130	0.246	-0.036	0.122	-0.033	0.125
$\hat{\beta}_{2,2}$	-0.104	0.175	-0.089	0.176	0.022	0.096	0.020	0.097

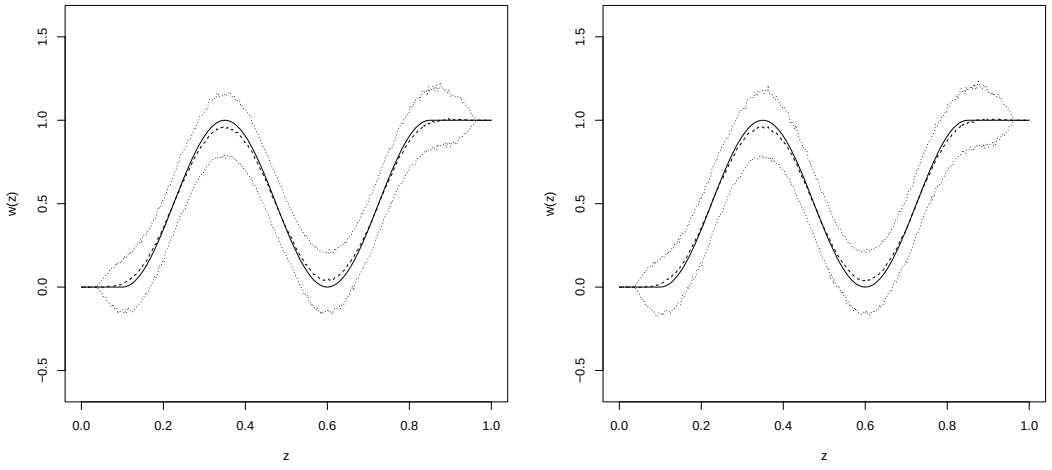


FIGURE 3.3: The mean estimates (dashed line) and 5% and 95% quantiles (dotted lines) of the transition function in DGP3 with $T = 400$; the solid line depicts the true transition function. The left and right panels correspond to SETR/C and SETR/J estimates, respectively.

the smallest MSEs. Since the transition function is again smooth, SETR/C should be more precise than SETR/J, but the difference between the two methods seems negligible.

3.5.4 Two-jump function

Finally, we present the results for DGP4, which includes two jumps with a smooth transition between them, see Figure 3.4. Also in this case, both parametric models, TAR and LSTAR, are misspecified, which lead to substantial biases – see Table 3.5 for the simulation results ($T = 400$). The semiparametric transition methods SETR/C and SETR/J

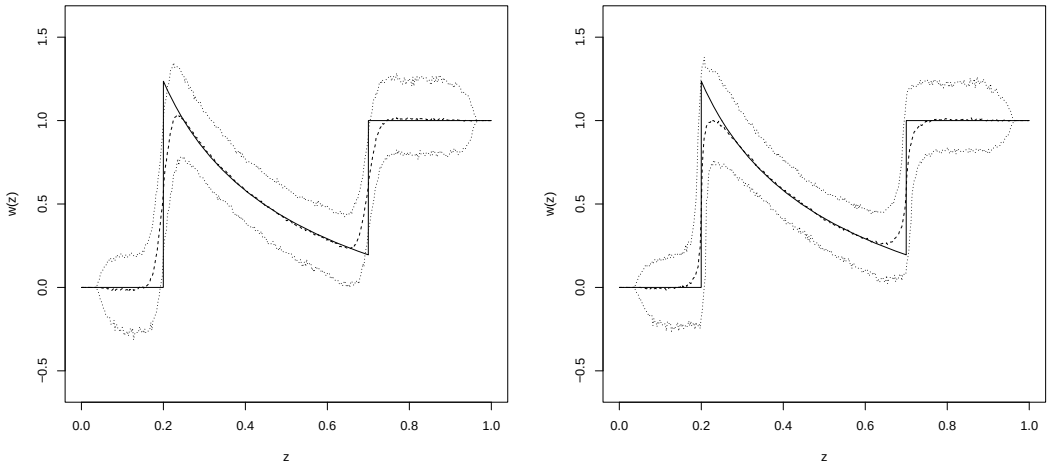


FIGURE 3.4: The mean estimates (dashed line) and 5% and 95% quantiles (dotted lines) of the transition function in DGP4 with $T = 400$; the solid line depicts the true transition function. The left and right panels correspond to SETR/C and SETR/J estimates, respectively.

TABLE 3.5: Biases and MSE of all estimator for DGP4 and $T = 400$.

	TAR		LSTAR		SETR/C		SETR/J	
	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
$\hat{\beta}_{1,0}$	0.064	0.158	0.062	0.166	0.005	0.102	0.004	0.105
$\hat{\beta}_{1,1}$	-0.162	0.291	-0.158	0.314	0.021	0.132	0.010	0.125
$\hat{\beta}_{1,2}$	0.106	0.198	0.103	0.215	-0.013	0.102	-0.006	0.098
$\hat{\beta}_{2,0}$	-0.085	0.134	-0.082	0.138	-0.005	0.087	-0.002	0.086
$\hat{\beta}_{2,1}$	0.202	0.268	0.196	0.277	-0.017	0.122	-0.012	0.118
$\hat{\beta}_{2,2}$	-0.133	0.185	-0.128	0.190	0.010	0.090	0.007	0.088

provide consistent estimates with relatively small biases and the smallest MSEs. Due to discontinuities of the transition function, SETR/J performs slightly better than SETR/C. The difference is not very large though as the biases of the transition function estimates are similar in both cases (see Figure 3.4). The reason behind this seemingly surprising results, especially in comparison to DGP1a and DGP1b, is the bandwidth choice: the cross-validation selects for SETR/C a smaller bandwidth in the presence of two breaks than in the case of a constant function with one break only, which leads to a reasonable approximation of the discontinuous transition function.

To sum up, the estimation of the semi-parametric transition model performs well in

all cases. Obviously, the MSEs of the estimates from the semiparametric estimation are larger than those from the parametric estimations, when the DGPs are correctly specified in the case of TAR or LSTAR. But the gap is relatively small in the case of TAR and practically negligible in the case of LSTAR and the semiparametric procedure offers extra flexibility in modeling the transition function.

3.6 Application to GDP

To demonstrate the use of the proposed semiparametric transition model, we analyze the quarterly GDP of the USA in years 1948–2007. The GDP (and GNP) series have been analyzed in the context of threshold autoregression or multiple regime models by many authors, for example, by Potter (1995) or Tiao and Tsay (1994); see Hansen (2011) for an overview of this line of research. In particular, we consider the logarithm of the growth of quaterly GDP in two time periods (similarly to Clements and Krolzig, 1998) corresponding to the first and last 2/3 of the sample:[§] from 1948–1987 and from 1967–2007 as some authors suspect that the post-war behavior was characterized by a different dynamic behavior than later at the end of the 20th century.

A suitable autoregressive model was chosen by minimizing the bias-corrected Akaike information criterion AIC_C of Hurvich et al. (1998) for autoregressive orders p and lags d of the transition variable from 1 to 5. In the second period 1967–2007, the selected model is the same as in other works such as Potter (1995), that is, the employed model is AR(5) without the third and fourth autoregressive terms (although their omission does not affect results much) and the transition variable z_t is chosen as the second lag of the dependent variable. The results are however different in the first period 1948–1987, where the AIC_C criterion leads to the AR(2) model, and more importantly, the transition between regimes are best characterized by the fourth lag of the dependent variable (i.e., the GDP growth one year ago).

The estimation was performed by the algorithm described in Section 3.3, where we assume that observations with the values of the transition variable below its 5% quantile or above

[§]The results are not sensitive to the selection of the interval end-points as results are rather similar for time intervals 1948–198x and 196x–2007.

TABLE 3.6: Coefficient estimates for the semiparametric transition model of US GDP in 1948–1987 and 1967–2007 based on AR(2) and AR(5) without the 3rd and 4th autoregressive terms, respectively. The standard errors are in brackets.

		1948–1987		1967–2007	
		TAR	SETR/J	TAR	SETR/J
Regime 1	AR(1)	-1.139	-0.901	0.736	0.756
		(0.441)	(0.417)	(0.365)	(0.380)
	AR(2)	0.784	0.530	-2.231	-1.971
		(0.506)	(0.472)	(0.669)	(0.757)
	AR(5)	—	—	1.166	1.249
		—	—	(0.518)	(0.549)
Regime 2	AR(1)	0.326	0.525	0.238	0.240
		(0.078)	(0.105)	(0.080)	(0.083)
	AR(2)	0.047	0.010	0.138	0.164
		(0.077)	(0.106)	(0.090)	(0.092)
	AR(5)	—	—	-0.140	-0.149
		—	—	(0.075)	(0.078)
	Threshold	-1.243	—	-0.692	—

its 95% quantile are completely in regime 1 or regime 2. Recall that this constraint is also imposed on the estimates of the transition function $w(z_t)$. The estimation was performed by the jump-preserving local-constant estimator of Čížek and Koo (2017a), see Section 3.3. The bandwidth h and the discontinuity-cutoff value u_T were chosen by the leave-one-out cross-validation. Estimation employs the quartic kernel.

The coefficient estimates are reported in Table 3.6 along with the TAR estimates traditionally used for this kind of analysis and the estimates of the transition function $w(z_t)$ for both periods are in Figure 3.5. Considering first the period 1967–2007, the estimated transition function approximately attains only values 0 and 1 with one jump between them at $z_t \approx -0.75$ and the SETR model thus reduces to the TAR model. This likely explains the model selection equivalent to the models found previously in the literature as well as the coefficients of both TAR and SETR estimates being close and exhibiting commonly known patterns: similarly to Potter (1995), for instance, the AR(1) coefficients are positive in all regimes, but the AR(2) coefficients are negative in regime 1, which corresponds to small values of z_t (below threshold in TAR), that is, to recession. In regime

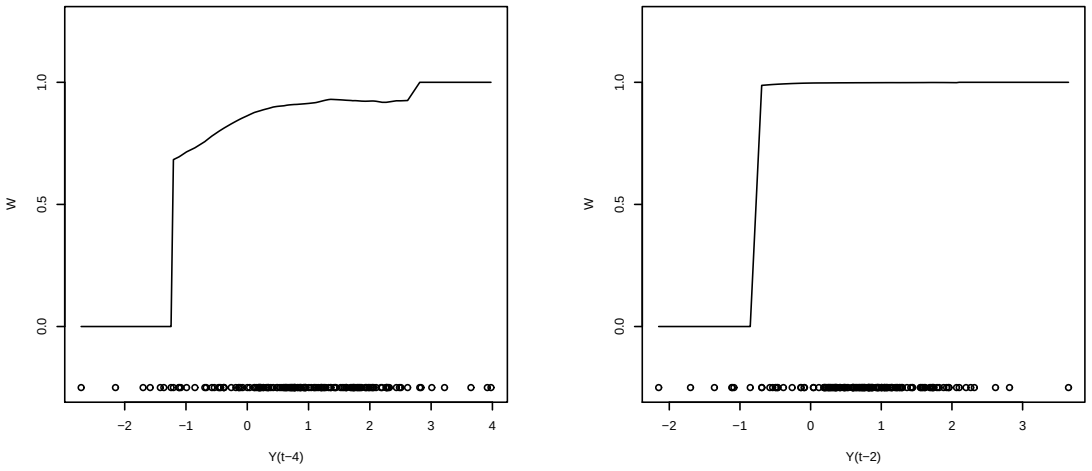


FIGURE 3.5: Transition function estimates for the semiparametric transition model of US GDP based on AR(5): 1948–1987 in the left panel and 1967–2007 in the right panel. The circles indicate the values of the transition variable observed in the data set.

2, which corresponds to large values of z_t (above threshold in TAR), the AR(2) coefficients are typically close to zero or positive depending on the time period used, which is the case also in Table 3.6. Apart from the estimates, the standard errors of TAR and SETR are close to each other as well, where the latter are slightly larger. All standard errors are large in regime 1 since there are only 8 observations available in that regime for both estimators. This lack of precision could also led to the substantially negative AR(2) coefficient in regime 1 for data 1960–2007.

Next, looking at the earlier period 1948–1987, the estimated transition function still exhibits one jump approximately at $z_t \approx -1.25$, but above the jump, the transition function monotonically increases from 0.7 to 1. This indicates the continuum of models characterizing the GDP growth depending on its past level, where the actual model always corresponds to a convex combination of the two regimes in Table 3.6. Given this departure from the standard TAR model, it is not surpising that the model selection resulted in a different autoregressive order, a different choice transition variable, and different coefficients than in the previously discussed period. The signs of coefficients now differ in the case of the AR(1) coefficient, which is negative in recession (regime 1) and positive otherwise (regime 2). The SETR estimates have a lower magnitude in regime 1 and a larger magnitude of the AR(1) coefficient in regime 2, and given the estimated transition function, the effective AR(1) coefficient implied by the SETR model ranges

from 0.08 to 0.52 for the lagged GDP growth $z_t \in (-1.2, 3.0)$ in contrast to the TAR model implying the same AR(1) coefficient 0.36 for any $z_t > -1.2$. The standard errors of both estimators are similar and large in regime 1 again, while being relatively small in regime 2.

Altogether, these results provide some evidence in favor of the semiparametric transition model by demonstrating that, for example, TAR might be too restrictive in some situations, even though a formal rejection of TAR would have to be based on confidence bands, and due to their likely width, a larger sample size.

3.7 Conclusion

The traditional TAR and STAR models both rely on the parametric form of the transition function. When the transition function differs from the presumed one, the estimation results often become biased and inconsistent. As a remedy, we develop the semiparametric transition (SETR) model that generalizes the two-regime (smooth) transition model by assuming an unknown transition function. We propose an iterative estimation procedure for the SETR model which is based on the straightforward application of (local) least squares. Practically most consistent estimators discussed in the varying-coefficient literature can be used to estimate the conditional transition function as long as they are stochastically equicontinuous in its dependent variable and regressors. For the slope estimator, the consistency and asymptotic normality are derived in the chapter, while the nonparametric estimates of the transition function are only shown to be consistent.

The simulation study using different types of transition functions indicates that the slope estimators from the LS estimation of the parametric TAR and STAR models are sensitive to the choice of the transition functions. On the other hand, the proposed estimator of the SETR model performs similarly to the parametric procedures (with a correctly specified transition function). Hence, the SETR model is a practically applicable alternative in the parametric settings.

Although there is only a single transition variable and a two-regime case considered, the SETR model can be extended to a linear combination of several transition variables and

to multiple regimes similarly to the STAR model. Moreover, the asymptotic properties of the transition function estimator could be further investigated in order to develop tests for studying various features of the transition function (e.g., overshooting behavior).

3.8 Appendix: Proofs of the main theorems

The proofs of all theorems and related lemmas are collected in this section. Before discussing the proofs, let $d_t = (y_t, x_t^\top, z_t)^\top$ represent all observables at time t and $g(d_t, \beta, w) = \left\{y_t - x_t^\top \beta_1 - x_t^\top(\beta_2 - \beta_1)w(z_t)\right\}^2$ denote the squared residual. Additionally, notation $\omega_t = [1 - w(z_t), w(z_t)]^\top$ is used to facilitate a shorter notation, where necessary (e.g., the squared residual $g(d_t, \beta, w)$ can be written as $\left\{y_t - (\omega_t \otimes x_t)^\top \beta\right\}^2$).

Proof of Theorem 3.2.

Let the smallest eigenvalue of $\mathrm{E}[x_t x_t^\top | z_t \in (a, b)]$ be $\lambda_{\inf}(a, b)$ and

$$\begin{aligned} r(d_t, \beta, w) &= [y_t - x_t^\top \beta_1^0 - x_t^\top(\beta_2^0 - \beta_1^0)w(z_t)] - [y_t - x_t^\top \beta_1 - x_t^\top(\beta_2 - \beta_1)w(z_t)] \\ &= x_t^\top(\beta_1 - \beta_1^0) + x_t^\top((\beta_2 - \beta_1)w(z_t) - (\beta_2^0 - \beta_1^0)w^0(z_t)). \end{aligned}$$

By $\mathrm{E}[\varepsilon_t | \mathcal{I}_t] = 0$ in Assumption 3.A1, (3.3) implies that the expected squared error

$$\mathrm{E}g(d_t, \beta, w) = \mathrm{E}\left[\varepsilon_t - r(d_t, \beta, w)\right]^2 = \mathrm{E}g(d_t, \beta^0, w^0) + \mathrm{E}r^2(d_t, \beta, w).$$

To prove that (β^0, w^0) is the unique minimum of $\mathrm{E}g(d_t, \beta, w)$ in the sense specified in the theorem, we show for any $\delta > 0$ and $\epsilon > 0$

$$\inf_{\|\beta - \beta^0\| > \delta, w \in \mathcal{W}} \mathrm{E}r^2(d_t, \beta, w) > 0 \tag{3.14}$$

and

$$\inf_{\beta \in \mathcal{B}, \|w - w^0\|_{\infty, \epsilon, f_z} > \delta} \mathrm{E}r^2(d_t, \beta, w) > 0. \tag{3.15}$$

To verify these claims, it is enough to find a set $\mathcal{Z}_*$, $P(z_t \in \mathcal{Z}_*) > 0$, independent of β and w such that both inequalities (3.14) and (3.15) hold conditionally upon $z_t \in \mathcal{Z}_*$.

First, the identification of parameters β_1^0 and β_2^0 is discussed. By Assumption 3.A4, there exist intervals (a_1, b_1) and (a_2, b_2), $P(z_t \in (a_j, b_j)) > 0$ for $j = 1, 2$, such that $w(z_t) = j-1$ for $z_t \in (a_j, b_j)$ and $j = 1, 2$ for any considered function w (i.e., also for w^0). Due to Assumption 3.A3, it follows that

$$\begin{aligned}\mathrm{E}[r^2(d_t, \beta, w)|z_t \in (a_j, b_j)] &= (\beta_j - \beta_j^0)^\top \mathrm{E}[x_t x_t^\top | z_t \in (a_j, b_j)](\beta_j - \beta_j^0) \\ &\geq \lambda_{\inf}(a_j, b_j) \left\|\beta_j - \beta_j^0\right\|^2 \geq 0,\end{aligned}$$

where the last inequality becomes equality if and only if $\beta_j = \beta_j^0$ for $j = 1, 2$. If $\|\beta - \beta^0\| > \delta$, inequality (3.14) follows.

Next, as (3.14) implies that the claim of the theorem holds for any w if $\|\beta - \beta^0\| > \delta$ for any arbitrarily small $\delta > 0$, we just have to verify (3.15) for $\|\beta - \beta^0\| \leq \delta$, assuming without loss of generality that $\delta \ll 1$. Let us consider a continuous function $w \in \mathcal{W}$ such that $\|w - w^0\|_{\infty,\epsilon,f_z} > \delta$. The continuity of the z_t distribution and the uniformly bounded derivatives of function w (within $\mathcal{W}$) then imply that there is an interval $I_w = (a, b]$ or $[a, b)$, $b - a > \epsilon_1 > 0$, such that (i) $|w(z) - w^0(z)| > \delta_1 > 0$ for any $z \in I_w$, (ii) $w(z)$ and $w^0(z)$ are continuous on I_w, and (iii) $P(z_t \in I_w) > \epsilon_2 > 0$. (Note that while the intervals I_w differ across functions $w \in \mathcal{W}$ such that $\|w - w^0\|_{\infty,\epsilon,f_z} > \delta$, ϵ_1, ϵ_2, and δ_1 can be chosen common to all such functions since ε_t is absolutely continuous and thus of bounded variation). Consider now a partition of $\mathbb{R} = \bigcup_{k=1}^{\infty} I_k$ consisting of intervals I_k such that $P(z_t \in I_k) = \epsilon_2/2$. Then for each w, there exists some $k(w) \in \mathbb{N}$ such that

$I_{k(w)} \subset I_w$ and it follows that (for β such that $\|\beta - \beta^0\| \leq \delta$)

$$\begin{aligned}
&\mathrm{E}r^2(d_t, \beta, w) = \sum_{k=1}^{\infty} \mathrm{E}[r^2(d_t, \beta, w) | z_t \in I_k] \cdot P(z_t \in I_k) \\
&\geq \sum_{k=1}^{\infty} \mathrm{E}\left[\{w(z_t) - w^0(z_t)\} \cdot (\beta_2^0 - \beta_1^0)^\top x_t x_t^\top (\beta_2^0 - \beta_1^0) \cdot \{w(z_t) - w^0(z_t)\}\right. \\
&\quad \left. \cdot (1 - \mathrm{O}(\delta)) \, | z_t \in I_k\right] \cdot P(z_t \in I_k) \\
&\geq \delta_1^2 \cdot (\beta_2^0 - \beta_1^0)^\top \mathrm{E}[x_t x_t^\top | z_t \in I_{k(w)}] (\beta_2^0 - \beta_1^0) \cdot (1 - \mathrm{O}(\delta)) \cdot P(z_t \in I_{k(w)}) \\
&\geq \delta_1^2 \epsilon_2 / 2 \cdot \inf_{k \in \mathbb{N}} \lambda_{\inf}(I_k) \cdot \left\|\beta_2^0 - \beta_1^0\right\|^2 \cdot (1 - \mathrm{O}(\delta)) > 0
\end{aligned}$$

by Assumption 3.A2 and 3.A3. Hence, (3.15) follows and the least-squares criterion (3.4) is minimized at (β^0, w) only if $\|w - w^0\|_{\infty, \epsilon, f_z} = 0$. □

Lemma 3.7. *Under Assumptions 3.A and 3.B, it holds for $T \to +\infty$ that*

$$\sup_{(\beta, w) \in \mathcal{B} \times \mathcal{W}} \left| \frac{1}{T} \sum_{t=1}^{T} g(d_t; \beta, w) - \mathrm{E} g(d_t; \beta, w) \right| \xrightarrow{P} 0.$$

Proof. Note first that $g(d_t, \beta, w) = (y_t - x_t^\top \beta_1 - x_t^\top (\beta_2 - \beta_1) w(z_t))^2 = \{\varepsilon_i - x_t^\top (\beta_1 - \beta_1^0) - x_t^\top [(\beta_2 - \beta_1) w(z_t) - (\beta_2^0 - \beta_1^0) w^0(z_t)]\}^2$ and thus

$$\begin{aligned}
g(d_t, \beta, w) =& \{\varepsilon_t - x_t^\top (\beta_1 - \beta_1^0)\}^2 - 2\varepsilon_t x_t^\top [(\beta_2 - \beta_1) w(z_t) - (\beta_2^0 - \beta_1^0) w^0(z_t)] \\
&+ \left\{[w(z_t)(\beta_2 - \beta_1) - w^0(z_t)(\beta_2^0 - \beta_1^0)]^\top x_t x_t^\top \right. \\
&\left. \times [(\beta_2 - \beta_1) w(z_t) - (\beta_2^0 - \beta_1^0) w^0(z_t)]\right\}. \qquad (3.16)
\end{aligned}$$

Using the triangle inequality and additivity of expectation, we have to show the uniform law of large numbers holds for each term on the right-hand side of (3.16). Obviously, the pointwise law of large numbers (for a given β_1) holds for

$$\frac{1}{T} \sum_{t=1}^{T} \{\varepsilon_t - x_t^\top (\beta_1 - \beta_1^0)\}^2 = \frac{1}{T} \sum_{t=1}^{T} \{\varepsilon_t^2 - 2\varepsilon_t x_t^\top (\beta_1 - \beta_1^0) + (\beta_1 - \beta_1^0)^\top x_t x_t^\top (\beta_1 - \beta_1^0)\}$$

due to Assumptions 3.B1 and 3.B2 and Example 16.3 and Theorem 20.15 of Davidson (1994), for instance. It applies also uniformly since terms of $\{\varepsilon_t - x_t^\top (\beta_1 - \beta_1^0)\}^2$ are linear

and quadratic forms of β_1, which is independent of t and belongs to a compact subset of $\mathbb{R}^p$ by Assumption 3.A2 (e.g., $T^{-1}\sum_{t=1}^{T} 2\varepsilon_t x_t^\top(\beta_1 - \beta_1^0) = 2(\beta_1 - \beta_1^0)\cdot T^{-1}\sum_{t=1}^{T}\varepsilon_t x_t^\top$).

Next, we have to prove that the uniform law of large numbers applies to the remaining terms in (3.16). All these terms contain the functions w and w^0 at z_t and can be written as sums of elements with a common form $\tilde{x}_t^1\tilde{x}_t^2\tilde{\beta}\tilde{w}(z_t)$, where $\tilde{x}_t^1$ and $\tilde{x}_t^2$ represent some individual elements of x_t or ε_t, $\tilde{\beta}$ represents $\beta_1, \beta_2, \beta_1^0, \beta_2^0$ or their product, and $\tilde{w}(z_t)$ stands for $w(z_t)$, $w^0(z_t)$, or their product. Consequently, we can define a function $f(z_t, \tilde{x}_t^1, \tilde{x}_t^2) = \tilde{x}_t^1\tilde{x}_t^2\tilde{\beta}\tilde{w}(z_t)$ of three random variables and a class $\mathcal{F} = \{f(z_t, \tilde{x}_t^1, \tilde{x}_t^2) = \tilde{x}_t^1\tilde{x}_t^2\tilde{\beta}\tilde{w}(z_t) : \beta_1 \in \mathcal{B}, \beta_2 \in \mathcal{B}, w \in \mathcal{W}\}$. By Assumption 3.B3, these functions are piecewise continuous on the partition of $\mathbb{R}^3 = \bigcup_{j=1,k,l=-\infty}^{\infty,\ \infty} I_j \times [k, k+1) \times [l, l+1)$. The boundedness of $\tilde{\beta}$ and $\tilde{w}(z_t)$ (Assumptions 3.A2 and 3.B3) guarantees that $f(z_t, \tilde{x}_t^1, \tilde{x}_t^2) = \tilde{x}_t^1\tilde{x}_t^2\tilde{\beta}\tilde{w}(z_t)$ have finite $(2+\xi)$-th moments uniformly bounded in $\mathcal{F}$. Furthermore, as the functions $f \in \mathcal{F}$ belong on each partition of $\mathbb{R}^3$ to $C_M^\gamma\left(I_j \times [k, k+1) \times [l, l+1)\right)$ for $\gamma > 3$, Assumption 3.B4 and van der Vaart and Wellner (1996, Corollary 2.7.4), applied with $V \in [3/\gamma, 1)$ and $r = 2+\xi$, imply that Theorem 5.2 of Dedecker and Louhichi (2002, see also page 146, point 1) holds for $\mathcal{F}$ and $\mathcal{F}$ forms thus a Donsker class of functions. By van der Vaart and Wellner (1996, Corollary 2.3.12), the class $\mathcal{F}$ is thus totally bounded and satisfies the stochastic equicontinuity condition.

Finally, as the pointwise law of large numbers applies to $T^{-1}\sum_{t=1}^{T} f(z_t, \tilde{x}_t^1, \tilde{x}_t^2)$ by Davidson (1994, Theorem 20.15) (see Assumptions 3.B1 and 3.B2, noting that w are measurable functions), uniform convergence on $\mathcal{F}$ follows from Andrews (1992, Theorem 1), which concludes the proof. □

Proof of Theorem 3.3.

Suppose Assumptions 3.A and 3.B hold. We now show that estimators $(\hat{\beta}_T, \hat{w}_T)$ are consistent in the sense that $\hat{\beta}_T \xrightarrow{P} \beta^0$ and $\|\hat{w}_T - w^0\|_{\infty,\epsilon,f_z} \to 0$. Consider some $\delta > 0$ and a set $U(\delta) = \{(\beta, w) \in \mathcal{B}\times\mathcal{W} : \|\beta - \beta^0\| > \delta \text{ or } \|w - w^0\|_{\infty,\epsilon,f_z} > \delta\}$. Further, let $\eta = \inf_{(\beta,w)\in U(\delta)} \mathrm{E}[g(d_t, \beta, w)] - \mathrm{E}[g(d_t, \beta^0, w^0)] > 0$ due to Theorem 3.2. By Lemma 3.7,

there exist for any $\epsilon' > 0$ some finite $T_1 > 0$ and $T_2 > 0$ such that for all $T > T_1$

$$\inf_{(\beta,w)\in U(\delta)} \frac{1}{T}\sum_{t=1}^{T} g(d_t,\beta,w) > \inf_{(\beta,w)\in U(\delta)} \mathrm{E}[g(d_t,\beta,w)] - \eta/2 = \mathrm{E}[g(d_t,\beta^0,w^0)] + \eta/2,$$

and for all $T > T_2$,

$$\frac{1}{T}\sum_{t=1}^{T} g(d_t,\beta^0,w^0) < \mathrm{E}[g(d_t,\beta^0,w^0)] + \eta/2$$

with an arbitrarily high probability $1-\epsilon'$. Thereby, $T > \max\{T_1, T_2\}$ implies $P\{(\hat{\beta}_T, \hat{w}_T) \notin U(\delta)\} \geq 1-\epsilon'$ and letting $\delta \to 0$ completes the proof of $\hat{\beta}_T \xrightarrow{P} \beta^0$ and $\|\hat{w}_T - w^0\|_{\infty,\epsilon,f_z} \to 0$ as $T \to \infty$. Since $P\{f_z(z_t) \leq \epsilon\} \to 0$ as $\epsilon \to 0$, it follows that $\{\hat{w}_T(z_t) - w^0(z_t)\}^2 \xrightarrow{P} 0$, and due to uniform boundedness of functions $\hat{w}_T$ and w^0 (Assumption 3.B3), $\mathrm{E}\{\hat{w}_T(z_t) - w^0(z_t)\}^2 \to 0$ as $T \to \infty$. □

Proof of Theorem 3.4.

Part 1. By Assumption 3.A4, $w(z_t) = 0$ and thus $y_t = x_t^\top \beta_1^0 + \varepsilon_t$ for $z_t \in (a_1^{k^*}, b_1^{k^*})$. Denoting $\mathcal{T}_{1T} = \{t \leq T : z_t \in (a_1^{k^*}, b_1^{k^*})\}$ and $|\mathcal{T}_{1T}|$ its cardinality, the LS estimator $\hat{\beta}_{1,T}^{(0,k^*)}$ using data point $z_t \in (a_1^{k^*}, b_1^{k^*})$ thus equals

$$\hat{\beta}^1 = \left(\frac{1}{T}\sum_{t\in\mathcal{T}_{1T}} x_t x_t^\top\right)^{-1}\left(\frac{1}{T}\sum_{t\in\mathcal{T}_{1T}} x_t y_t\right) = \beta_1^0 + \left(\frac{1}{T}\sum_{t\in\mathcal{T}_{1T}} x_t x_t^\top\right)^{-1}\left(\frac{1}{T}\sum_{t\in\mathcal{T}_{1T}} x_t \varepsilon_t\right).$$

By Assumption 3.A4, $|\mathcal{T}_{1T}| \to \infty$ as $T \to \infty$. As in Lemma 3.7, the pointwise law of large numbers Davidson (1994, Theorem 20.15) thus applies to each sum in the last two brackets (see Assumptions 3.B1 and 3.B2). Hence, $T^{-1}\sum_{t=1}^{T} x_t x_t^\top \mathbf{1}(z_t \in (a_1^{k^*}, b_1^{k^*})) \xrightarrow{P} \mathrm{E}(x_t x_t^\top \mathbf{1}(z_t \in (a_1^{k^*}, b_1^{k^*}))$, which is non-singular by Assumption 3.A3. Similarly, we have $T^{-1}\sum_{t=1}^{T} x_t \varepsilon_t^\top \mathbf{1}(z_t \in (a_1^{k^*}, b_1^{k^*})) \xrightarrow{P} 0$ by Assumption 3.A1 and hence $\hat{\beta}_{1,T}^{(0,k^*)} \xrightarrow{P} \beta_1^0$ as $T \to \infty$. Analogously, one can show that $\hat{\beta}_{2,T}^{(0,k^*)} \xrightarrow{P} \beta_2^0$, and consequently, it follows $\hat{\beta}_T^{(0,k^*)} \to \beta^0$ as $T \to \infty$.

Part 2. Let $\mathcal{Z}$ be the support of z_t, $\mathcal{Z}_T^j = \bigcup_{i=1}^J (s_i - \zeta_T, s_i + \zeta_T)$, and $\mathcal{Z}_T^c = \mathcal{Z} \setminus \mathcal{Z}_T^j$. The first claim is proved in two parts. First, we show that $\hat{w}_T(z, \hat{\beta}_T)$ converges to $\hat{w}_T(z, \beta^0)$ in probability uniformly on $\mathcal{Z}_T^c$. Second, we argue that $P(z_t \in \mathcal{Z}_T^j) \to 0$ as $T \to +\infty$.

By the assumption of the theorem, $\check{\beta}_T$ is consistent, and thus for any $\delta > 0$, $P(\|\check{\beta}_T - \beta^0\| < \delta) \to 1$ as $T \to +\infty$. Consequently, we have to show only that the difference $|\hat{w}_T(z, \beta) - \hat{w}_T(z, \beta^0)|$ converges to zero in probability uniformly in a neighborhood $\|\beta - \beta^0\| < \delta$ for some $\delta > 0$.

For $z \in \mathcal{Z}_T^c$ and any $\delta > 0$, we have

$$\sup_{\|\beta-\beta^0\|<\delta} \sup_{z\in\mathcal{Z}_T^c} \left|\hat{w}_T(z,\beta) - \hat{w}_T(z,\beta^0)\right| \leq \sup_{\|\beta-\beta^0\|<\delta} \sup_{z\in\mathcal{Z}_T^c} |\hat{w}_T(z,\beta) - w(z,\beta)| \tag{3.17}$$

$$+ \sup_{z\in\mathcal{Z}_T^c} \left|\hat{w}_T(z,\beta^0) - w(z,\beta^0)\right| \tag{3.18}$$

$$+ \sup_{\|\beta-\beta^0\|<\delta} \sup_{z\in\mathcal{Z}_T^c} \left|w(z,\beta^0) - w(z,\beta)\right|. \tag{3.19}$$

The three terms on the right hand side will be shown to converge to 0 in probability. Consider the first two terms (3.17) and (3.18), where the second term is a special case of the first one. By applying the generic uniform convergence theorem of Andrews (1992, Theorem 1), the uniform pointwise consistency and the stochastic equicontinuity of $\hat{w}_T(\cdot, \beta)$ (Assumptions 3.C2 and 3.C3) together with the compactness of $\mathcal{B}$ (Assumption 3.B1) imply that the first two terms (3.17) and (3.18) are asymptotically negligible in probability.

Regarding the third term (3.19), the mean value theorem can be applied here because $w(z, \beta)$ is differentiable in β for any $z \in \mathcal{Z}_T^c$ (Assumption 3.C4). Further by Assumption 3.C4, $\sup_{z\in\mathcal{Z}_T^c} \sup_{\|\beta-\beta^0\|<\delta} \|\partial w(z,\beta)/\partial\beta\|$ is bounded by some positive constant K_w uniformly in T. Hence, the mean value theorem implies

$$\sup_{||\beta-\beta^0||<\delta} \sup_{z\in\mathcal{Z}_T^c} \left|w(z,\beta^0) - w(z,\beta)\right| < K_w\delta.$$

Combining the results of all three terms and letting $\delta \to 0$ leads to

$$\sup_{z\in\mathcal{Z}_T^c} \left|\hat{w}_T(z,\check{\beta}_T) - \hat{w}_T(z,\beta^0)\right| \xrightarrow{P} 0 \tag{3.20}$$

as $T \to +\infty$, which completes the the first part of the proof.

Let us now consider the probability of $z_t \in \mathcal{Z}_T^j$. Assuming that the density of z_t is bounded by some constant K_f (Assumption 3.C5), it follows that

$$P(z \in \mathcal{Z}_T^j) \leq \sum_{j=1}^{J} P(z \in (s_j - \zeta_T, s_j + \zeta_T)) \leq 2JK_f\zeta_T,$$

and since $\zeta_T \to 0$ as $T \to \infty$, $P(z \in \mathcal{Z}_T^j) \to 0$ as $T \to +\infty$. Combined with (3.20), this implies that $|\hat{w}_T(z_t, \check{\beta}_T) - \hat{w}_T(z_t, \beta^0)| \xrightarrow{P} 0$, and because of $\hat{w}_T$ being uniformly bounded on $\mathcal{Z} \times \mathcal{B}$ (Assumption 3.C2), $\mathrm{E}[\hat{w}_T(z_t, \check{\beta}_T) - \hat{w}_T(z_t, \beta^0)]^2 \to 0$ as $T \to +\infty$.

Next, the remaining two claims about the consistency of $\hat{w}_T(z_t, \check{\beta}_T)$ follow by exactly the same arguments as above since the decomposition (3.17)–(3.19) becomes now

$$\sup_{\|\beta-\beta^0\|<\delta} \sup_{z \in \mathcal{Z}_T^c} \left|\hat{w}_T(z, \beta) - w_T(z, \beta^0)\right| \leq \sup_{\|\beta-\beta^0\|<\delta} \sup_{z \in \mathcal{Z}_T^c} |\hat{w}_T(z, \beta) - w(z, \beta)| \tag{3.21}$$

$$+ \sup_{\|\beta-\beta^0\|<\delta} \sup_{z \in \mathcal{Z}_T^c} \left|w(z, \beta^0) - w(z, \beta)\right|. \tag{3.22}$$

that is, it is equivalent to terms (3.17) and (3.19), leaving (3.18) out.

Part 3. The previous two points imply that, at least for one pair of intervals $(a_1^{k^*}, b_1^{k^*})$ and $(a_2^{k^*}, b_2^{k^*})$, the initial estimators $\hat{\beta}_T^{(0,k^*)}$ and $\hat{w}_T^{(0,k^*)}$ are consistent. The criterion $S_{(k)}^2$ for the selection of the final estimate is defined as the sum of squared residuals. Evaluated at $\hat{\beta}_T^{(0,k^*)}$ and $\hat{w}_T^{(0,k^*)}$, it equals $S_{(k^*)}^2 = T^{-1} \sum_{t=1}^{T} g(d_t; \hat{\beta}_T^{(0,k^*)}, \hat{w}_T^{(0,k^*)})$. It can be further decomposed as

$$\frac{1}{T} \sum_{t=1}^{T} \left\{ g(d_t; \hat{\beta}_T^{(0,k^*)}, \hat{w}_T^{(0,k^*)}) - \mathrm{E}g(d_t; \hat{\beta}_T^{(0,k^*)}, \hat{w}_T^{(0,k^*)}) \right\} \tag{3.23}$$

$$+ \mathrm{E}g(d_t; \hat{\beta}_T^{(0,k^*)}, \hat{w}_T^{(0,k^*)}) - \mathrm{E}g(d_t; \beta^0, w^0) \tag{3.24}$$

$$+ \mathrm{E}g(d_t; \beta^0, w^0), \tag{3.25}$$

where the first term (3.23) converges to 0 in probability by Lemma 3.7. The second term is asymptotically negligible in probability as $T \to \infty$ too because of the (uniform) consistency of $\hat{\beta}_T^{(0,k^*)}$ and $\hat{w}_T^{(0,k^*)}$: the uniform boundedness of the estimates (Assumptions

3.A2 and 3.C2) and decomposition

$$
\begin{aligned}
&\mathrm{E}g(d_t;\hat{\beta}_T^{(0,k^*)},\hat{w}_T^{(0,k^*)}) - \mathrm{E}g(d_t;\beta^0,w^0)\\
&= \mathrm{E}\Big\{y_t^2 - 2y_t x_t^\top[\hat{\beta}_{1,T}^{(0,k^*)}(1-\hat{w}_T^{(0,k^*)}(z_t)) + \hat{\beta}_{2,T}^{(0,k^*)}\hat{w}_T^{(0,k^*)}(z_t)]\\
&\quad + [\hat{\beta}_{1,T}^{(0,k^*)}(1-\hat{w}_T^{(0,k^*)}(z_t)) + \hat{\beta}_{2,T}^{(0,k^*)}\hat{w}_T^{(0,k^*)}(z_t)]^\top x_t x_t^\top\\
&\quad [\hat{\beta}_{1,T}^{(0,k^*)}(1-\hat{w}_T^{(0,k^*)}(z_t)) + \hat{\beta}_{2,T}^{(0,k^*)}\hat{w}_T^{(0,k^*)}(z_t)]\Big\}\\
&\quad - \mathrm{E}\big\{y_t^2 - 2y_t x_t^\top[\beta_1^0(1-w^0(z_t)) + \beta_2^0 w^0(z_t)]\\
&\quad - [\beta_1^0(1-w^0(z_t)) + \beta_2^0 w^0(z_t)]^\top x_t x_t^\top[\beta_1^0(1-w^0(z_t)) + \beta_2^0 w^0(z_t)]\big\}
\end{aligned}
$$

lead to (3.24) being negligible since (using one representative term of the decomposition)

$$
\begin{aligned}
&\mathrm{E}\left|2y_t x_t^\top \hat{\beta}_{2,T}^{(0,k^*)}\hat{w}_T^{(0,k^*)}(z_t) - 2y_t x_t^\top \beta_2^0 w^0(z_t)\right|\\
&= \mathrm{E}\left|2y_t x_t^\top[\hat{\beta}_{2,T}^{(0,k^*)} - \beta_2^0]\hat{w}_T^{(0,k^*)}(z_t) - 2y_t x_t^\top \beta_2^0[\hat{w}_T^{(0,k^*)}(z_t) - w^0(z_t)]\right|\\
&\le 2M\mathrm{E}|y_t x_t^\top[\hat{\beta}_{2,T}^{(0,k^*)} - \beta_2^0]| + 2\mathrm{E}|\beta_2^{0\top} x_t y_t[\hat{w}_T^{(0,k^*)}(z_t) - w^0(z_t)]|\\
&\le 2M\sqrt{\mathrm{E}\|y_t x_t\|^2 \mathrm{E}\|\hat{\beta}_{2,T}^{(0,k^*)} - \beta_2^0\|^2} + 2\|\beta_2^0\|\sqrt{\mathrm{E}\|y_t x_t\|^2 \mathrm{E}[\hat{w}_T^{(0,k^*)}(z_t) - w^0(z_t)]^2}\\
&= \mathrm{o}(1) + \mathrm{o}(1),
\end{aligned}
$$

as $T \to \infty$, where the last inequality employs the Cauchy inequality and the last equality uses the existence of the second moments (Assumption 3.B1) and the consistency (weak and in mean) established in points 1 and 2 of this theorem by the boundedness of the parameters and transition functions. Hence, criterion $S^2_{(k^*)}$ behaves as $\mathrm{E}g(d_t;\beta^0,w^0) + \mathrm{o_p}(1)$ as $T \to \infty$ and reaches asymptotically the minimum possible value by Theorem 3.2.

Among $k = 1,\ldots,\kappa$, there is at least one pair k^* of intervals leading to consistent estimates $\hat{\beta}_T^{(0,k^*)}$ and $\hat{w}_T^{(0,k^*)}$ and criterion $S^2_{(k^*)} = \mathrm{E}g(d_t;\beta^0,w^0) + \mathrm{o_p}(1)$ as $T \to \infty$. It remains to prove that any other estimator that can be selected in step 3 of the algorithm is also consistent. If k is an index in $1,\ldots,\kappa$ such that $\hat{\beta}_T^{(0,k)}$ and $\hat{w}_T^{(0,k)}$ are selected, $\hat{\beta}_T^{(0)} = \hat{\beta}_T^{(0,k)}$ and $\hat{w}_T^{(0)} = \hat{w}_T^{(0,k)}$, the corresponding sum of squared errors $S^2_{(k)} \le S^2_{(k^*)}$. For the estimator $(\hat{\beta}_T^{(0)},\hat{w}_T^{(0)})$, it holds that $\min_{k\in 1,\ldots,\kappa} S^2_{(k)} \le S^2_{(k^*)}$, and for any $\eta > 0$ and $\epsilon > 0$, there is a sufficiently large T_η such that $P(\min_{k\in 1,\ldots,\kappa} S^2_{(k)} \le S^2_{(k^*)} < \mathrm{E}g(d_t;\beta^0,w^0) + \eta/2) > 1-\epsilon$ for all $T > T_\eta$. This property however implies the weak consistency of the estimator

$(\hat{\beta}_T^{(0)}, \hat{w}_T^{(0)})$ as shown in the proof of Theorem 3.3. □

Lemma 3.8. *Under Assumptions 3.A–3.D,* $\check{Q}_T \xrightarrow{P} Q^0$ *and* $\hat{Q}_T^0 \xrightarrow{P} Q^0$ *as* $T \to +\infty$, *where* $\check{Q}_T = \frac{1}{T}\sum_{t=1}^T (\check{\omega}_t \otimes x_t)(\check{\omega}_t \otimes x_t)^\top$, $\hat{Q}_T^0 = \frac{1}{T}\sum_{t=1}^T (\omega_t^0 \otimes x_t)(\omega_t^0 \otimes x_t)^\top$, ω_t^0 *and* Q^0 *are defined in Assumption 3.D, and* $\check{\omega}_t = [1 - \check{w}_T(z_t), \check{w}_T(z_t)]^\top$ *is based on an estimator* $\check{w}_T$ *of* w^0 *such that* $\mathrm{E}[\check{w}_T(z_t) - w^0(z_t)]^2 \to 0$ *as* $T \to +\infty$.

Proof. Suppose Assumptions 3.A–3.D hold. As in Lemma 3.7, it is possible to apply the law of large numbers (Davidson, 1994, Theorem 20.15) because of the satisfied mixing conditions (Assumptions 3.B1 and Theorem 14.1 in Davidson, 1994) and existence of the sufficiently high moments (Assumptions 3.B2 along with Assumption 3.B3 and 3.C2) allow application of the law of large numbers (Davidson, 1994, Theorem 20.15), which implies that $\hat{Q}_T^0 \xrightarrow{P} Q^0$ as $T \to \infty$. Since $\check{Q}_T - Q^0 = (\check{Q}_T - \hat{Q}_T^0) + (\hat{Q}_T^0 - Q^0)$, we just have to show that $\check{Q}_T - \hat{Q}_T^0 = \mathrm{o_p}(1)$ as $T \to \infty$. For any $\epsilon > 0$, the Markov and Cauchy-Schwartz inequalities imply for the klth elements of $\check{Q}_T$ and $\hat{Q}_T^0$, $k, l = p+1, \dots, 2p$, that

$$
\begin{aligned}
&P(|\check{Q}_{T,kl} - \hat{Q}_{T,kl}^0| > \epsilon) = P\left(\left|\frac{1}{T}\sum_{t=1}^T [\check{w}_T^2(z_t) - {w^0}^2(z_t)](x_{t,k}x_{t,l})\right| > \epsilon\right) \\
&\leq \frac{1}{\epsilon}\mathrm{E}\left|\frac{1}{T}\sum_{t=1}^T [\check{w}_T^2(z_t) - {w^0}^2(z_t)](x_{t,k}x_{t,l})\right| \leq \frac{1}{\epsilon}\mathrm{E}\left|[\check{w}_T^2(z_t) - {w^0}^2(z_t)](x_{t,k}x_{t,l})\right| \\
&\leq \frac{1}{\epsilon}\sqrt{\mathrm{E}\left|\check{w}_T^2(z_t) - {w^0}^2(z_t)\right|^2 \mathrm{E}\left|x_{t,k}x_{t,l}\right|^2} \\
&= \frac{1}{\epsilon}\sqrt{\mathrm{E}\left|\check{w}_T(z_t) - w^0(z_t)\right|^2 \mathrm{E}\left|\check{w}_T(z_t) + w^0(z_t)\right|^2 \mathrm{E}\left|x_{t,k}x_{t,l}\right|^2}
\end{aligned}
$$

(the argument looks analogously for $k = 1, \dots, p$ or $l = 1, \dots, p$). While the last two expectations below the square root are uniformly bounded due to the boundedness of the transition functions w^0 and $\hat{w}_T$ (Assumptions 3.A4 and 3.C2) and the existence of the second moments of $x_t x_t^\top$ (Assumption 3.B1), the first expectation converges to zero by the assumption of the lemma: $\mathrm{E}[\check{w}_T(z_t) - w^0(z_t)]^2 \to 0$ as $T \to +\infty$. As each element of the matrix difference $\check{Q}_T - \hat{Q}_T^0$ is asymptotically negligible in probability, it follows that $\check{Q}_T - \hat{Q}_T^0 = \mathrm{o_p}(1)$ and thus $\check{Q}_T - Q_T^0 = \mathrm{o_p}(1)$ as $T \to +\infty$. □

Proof of Theorem 3.5

Let $\check{\omega}_t = [1 - \check{w}_T(z_t), \check{w}_T(z_t)]^\top$ and $\omega_t^0 = [1 - \check{w}^0(z_t), \check{w}^0(z_t)]^\top$ again. By Lemma 3.8 and Assumption 3.D, it holds that

$$\begin{aligned}
&\sqrt{T}\{\hat{\beta}(\check{w}_T) - \hat{\beta}(w^0)\} \\
&= \left(\frac{1}{T}\sum_{t=1}^{T}(\check{\omega}_t \otimes x_t)(\check{\omega}_t \otimes x_t)^\top\right)^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\check{\omega}_t \otimes x_t)y_t\right) \\
&\quad - \left(\frac{1}{T}\sum_{t=1}^{T}(\omega_t^0 \otimes x_t)(\omega_t^0 \otimes x_t)^\top\right)^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\omega_t^0 \otimes x_t)y_t\right) \\
&= \left(Q^0\right)^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\{(\check{\omega}_t - \omega_t^0) \otimes x_t\}y_t\right)(1 + \mathrm{o_p}(1)).
\end{aligned}$$

Further, let $\omega_t^w = [1 - w(z_t), w(z_t)]^\top$ for any $w \in \mathcal{W}$ and $e_k = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$ be the kth standard basis vector in $\mathbb{R}^{2p}$. Considering the class of functions $\mathcal{F}_k = \{f(x_t, y_t, z_t) = e_k^\top(\omega_t^w \otimes x_t) \cdot y_t$ for any $k = 1, \ldots, 2p$, we have verified in Lemma 3.7 after substituting for y_t from (3.3) that each class $\mathcal{F}_k$ is Donsker and satisfies thus the stochastic equicontinuity condition. Additionally, $P(\check{\omega}_t \in \mathcal{W}) \to 1$ and $\mathrm{E}(\check{\omega}_t - \omega_t^0)^2 \to 0$ as $T \to +\infty$ by the assumptions of the theorem. Hence, it holds with probability arbitrarily close to 1 that

$$\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\{(\check{\omega}_t - \omega_t^0) \otimes x_t\}y_t\right\| \leq \sup_{w \in \mathcal{W}, \mathrm{E}[w(z_t) - w^0(z_t)]^2 < \delta}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\{(\omega_t^w - \omega_t^0) \otimes x_t\}y_t\right\|,$$

where the right-hand side is negligible in probability as $T \to +\infty$ and $\delta \to 0$ due to the stochastic equicontinuity of the classes of functions $\mathcal{F}_k$ corresponding to the elements of vectors $\{(\omega_t^w - \omega_t^0) \otimes x_t\}y_t$. Consequently, $\sqrt{T}\{\hat{\beta}(\check{w}_T) - \hat{\beta}(w^0)\} = \mathrm{o_p}(1)$.

The remaining results of the theorem follow directly from Theorem 3.4: as $\hat{\beta}_T^{(0,k^*)}$ and $\hat{w}_T^{(0,k^*)} = \hat{w}_T(\cdot, \hat{\beta}_T^{(0,k^*)})$ are consistent in probability and in mean, respectively, the first

claim of this theorem implies that $\hat{\beta}_T^{(1,k^*)} = \hat{\beta}_T(\hat{w}_T^{(0,k^*)}) = \hat{\beta}_T(w^0) + \mathrm{o_p}(T^{-1/2})$. Since

$$\begin{aligned}\hat{\beta}_T(w^0) - \beta^0 &= \left[\left(\frac{1}{T}\sum_{t=1}^{T}(\omega_t^0 \otimes x_t)(\omega_t^0 \otimes x_t)^\top\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}(\omega_t^0 \otimes x_t)y_t\right) - \beta^0\right] \\ &= \left(\frac{1}{T}\sum_{t=1}^{T}(\omega_t^0 \otimes x_t)(\omega_t^0 \otimes x_t)^\top\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}(\omega_t^0 \otimes x_t)\varepsilon_t\right) \qquad (3.26)\end{aligned}$$

and $T^{-1}\sum_{t=1}^{T}(\omega_t^0 \otimes x_t)(\omega_t^0 \otimes x_t)^\top \to Q^0$ by Lemma 3.8, where Q^0 is a positive definite matrix (Assumption 3.D), the consistency of $\hat{\beta}_T^{(1,k^*)}$ follows from the law of large numbers verified for sequence $\{(\omega_t^0 \otimes x_t)\varepsilon_t\}$ in the proof of Lemma 3.7 and Assumption 3.A1 implying $E[(\omega_t^0 \otimes x_t)\varepsilon_t] = 0$. Hence, the part 2 of Theorem 3.4 can be applied for $\check{\beta}_T = \hat{\beta}_T^{(1,k^*)}$ and the proof of the part 3 of Theorem 3.4 can be repeated step by step for the iterated estimators $\hat{\beta}_T^{(1,k^*)}$ and $\hat{w}_T^{(1,k^*)} = \hat{w}_T(\cdot, \hat{\beta}_T^{(1,k^*)})$ to obtain the claims of this theorem. □

Proof of Theorem 3.6

Using $\omega_t^0 = [1 - w^0(z_t), w^0(z_t)]^\top$ again and multiplying (3.26) by $\sqrt{T}$, it holds that

$$\sqrt{T}(\hat{\beta}_T(w^0) - \beta^0) = \left(\frac{1}{T}\sum_{t=1}^{T}(\omega_t^0 \otimes x_t)(\omega_t^0 \otimes x_t)^\top\right)^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\omega_t^0 \otimes x_t)\varepsilon_t\right). \quad (3.27)$$

First, $T^{-1}\sum_{t=1}^{T}(\omega_t^0 \otimes x_t)(\omega_t^0 \otimes x_t)^\top \to Q^0$ in probability by Lemma 3.8, where Q^0 is a positive definite matrix (Assumption 3.D) and is thus non-singular. Hence, the first term in (3.27) converges to Q^{0-1} in probability.

Regarding to the second term in (3.27), we can apply the central limit theorem for mixing sequences (Davidson, 1994, Corollary 24.7) as the sequence $\{x_t, z_t, \varepsilon_t\}$ is absolutely regular of size $-(2+\xi)/\xi$ (Assumption 3.B1) with zero mean (Assumption 3.A1), $w^0 \in \mathcal{W}$ is measurable, and due to boundedness of w^0, $(\omega_t^0 \otimes x_t)\varepsilon_t$ has finite $(2+\xi)$th moments (Assumption 3.B2). Consequently,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(\omega_t^0 \otimes x_t)\varepsilon_t \xrightarrow{d} N(0, V^0)$$

in distribution, where V^0 is defined in Assumption 3.D.

Combining the two results, it follows that $\sqrt{T}(\hat{\beta}_T(w^0) - \beta^0) \xrightarrow{d} N(0, Q^{0^{-1}} V^0 Q^{0^{-1}})$ as $T \to +\infty$. Additionally, if an estimator $\check{w}_T$ satisfies $\mathrm{E}[\check{w}_T(z_t) - w^0(z_t)]^2 \to 0$, Theorem 3.5 implies $\sqrt{T}\{\hat{\beta}(\check{w}_T) - \hat{\beta}(w^0)\} = \mathrm{o_p}(1)$ as $T \to \infty$, and hence, it holds for $T \to +\infty$ that

$$\sqrt{T}(\hat{\beta}_T(\check{w}_T) - \beta^0) \xrightarrow{d} N(0, Q^{0^{-1}}\, V^0 Q^{0^{-1}}). \qquad \square$$

3.9 Appendix: Verification of Assumptions 3.C

In this section, assumptions for the nonparametric jump-preserving estimator proposed in Section 3.3.2 are introduced that are sufficient for the piecewise smooth estimates (Assumption 3.C1), their uniform consistency (Assumption 3.C2), and stochastic equicontinuity (Assumption 3.C3); only the uniform boundedness of the estimated transition function is not explicitly discussed as it can be trivially imposed during estimation. As discussed in Sections 3.2 and 3.4, we verify these properties on a compact subset D of the support $\mathcal{Z}$ of z_t since the transition function has to be estimated only outside of intervals (a_1, b_1) and (a_2, b_2), that is, only on a compact subset of $\mathbb{R}$ in practically all applications.

The sufficient assumptions introduced below characterize the properties of the marginal distribution of z_t, conditional moments, and the bandwidth and kernel function of the nonparametric estimator; these assumptions cover both the local constant and local linear estimators. The assumptions stem from Čížek and Koo (2017a), and in most cases, directly correspond to the assumptions of Hansen (2008); see Čížek and Koo (2017a) for their detailed discussion. Compared to the assumptions in the main part of the paper, this leads to a slightly stronger mixing assumption as well as the identification Assumption 3.A3 strengthened to more usual $\mathrm{E}(x_t x_t^\top | z_t) > 0$ so that the existing results of Hansen (2008) can be applied.

Assumption 3.E.

3.E1. The mixing coefficients satisfy $\beta(m) = \mathrm{O}(m^{-\gamma})$ as $m \to \infty$, where $\gamma > (1 + (1 + \xi)(2 + 1/\varsigma))/\xi$ for some $\varsigma \geq 1$.

3.E2. Random variable z_t is continuously distributed with density f_z, and on the compact set $D \subseteq \mathcal{Z}$, $\inf_{z\in D} f_z(z) > 0$. The derivative of f_z is bounded and Lipschitz continuous for $z \in D$. Additionally, the partial derivative of the joint density function f of (x_t, z_t) with respect to z_t is bounded and continuous uniformly on its support except for the points $\{s_q\}_{q=1}^J$, at which the left or right partial derivatives of f with respect to Z are bounded and left or right continuous, respectively.

3.E3. Let φ_t represent any element of matrix $x_t x_t^\top$, vector $x_t\varepsilon_t$, or variable ε_t^2, and for any t, let us assume that

$$\sup_{z\in D} \mathrm{E}(|\varphi_t|^{2+\xi}|z_t = z)f_z(z) < \infty, \ \sup_{z\in D} \mathrm{E}(\varphi_t \partial \ln f(x_t, z_t)/\partial z|z_t = z) < \infty, \text{and}$$
$$\sup_{(z',z'')\in D\times D} \mathrm{E}(|\varphi_1\varphi_t||z_1 = z', z_t = z'')f_{1t}(z', z'') < \infty,$$

where $f_{1t}(z', z'')$ denotes the joint density of (z_1, z_t).

3.E4. The variance matrix $\Omega(z) = \mathrm{E}[x_t x_t^\top|z_t = z]$ is bounded and positive definite uniformly on D except for the discontinuities $\{s_q\}_{q=1}^J$, at which variance matrices $\Omega_-(s_q) = \lim_{z\uparrow s_q} \mathrm{E}[x_t x_t^\top|z_t = z]$ and $\Omega_+(s_q) = \lim_{z\downarrow s_q} \mathrm{E}[x_t x_t^\top|z_t = z]$ are bounded and positive definite.

3.E5. The kernel $K^{(c)}(\cdot)$ is a bounded three times differentiable symmetric continuous density function and has a compact support $[-1, 1]$. It is chosen so that functions $\mathcal{K}_j^{(c)}(u) = u^j K^{(c)}(u)$ and the first three derivatives of $K^{(c)}(u)$ are Lipschitz continuous for all $j = 0, 1, 2, 3$ and the following constants are well defined and finite for $j = 0, 1, 2$ and $\iota = c, r, l$:

$$\mu_j^{(\iota)} = \int_{-1}^{1} v^j K^{(\iota)}(v)\mathrm{d}v, \qquad \nu_j^{(\iota)} = \int_{-1}^{1} v^j K^{(\iota)2}(v)\mathrm{d}v,$$
$$c_0^{(\iota)} = \frac{\mu_2^{(\iota)}}{\mu_2^{(\iota)}\mu_0^{(\iota)} - \mu_1^{(\iota)2}}, \quad \text{and} \quad c_1^{(\iota)} = \frac{-\mu_1^{(\iota)}}{\mu_2^{(\iota)}\mu_0^{(\iota)} - \mu_1^{(\iota)2}}. \tag{3.28}$$

3.E6. The bandwidths h_T and u_T satisfy $u_T \to 0$, $h_T \to 0$, and $Th_T \to \infty$ as $T \to \infty$ as well as $Th_T^5 \to \bar{c} \in [0, +\infty)$ as $T \to \infty$, where $\bar{c} \geq 0$ is some constant. Moreover,

the bandwidth h_T satisfies $\ln T/(Th_T^3) = \mathrm{o}(1)$ and $\ln T/(T^\theta h_T) = \mathrm{o}(1)$, where

$$\theta = \frac{\gamma - 2 - \frac{1}{\varsigma} - \frac{1+\gamma}{1+\xi}}{\gamma + 2 - \frac{1+\gamma}{1+\xi}}.$$

Along with Assumptions 3.A and 3.B, the above stated Assumptions 3.E cover all relevant assumptions used in Čížek and Koo (2017a) and Hansen (2008); for the simplicity of notation, the symbol $\mathcal{Z}_T^c$ refers now to the intersection of set $\mathcal{Z}_T^c$ defined in Assumption 3.C with set D in Assumption 3.E2 if $\mathcal{Z}$ itself is not compact. Now, considering the varying-coefficient model (3.10) for some $\beta \in U(\beta^0, \delta)$ and small $\delta > 0$,

$$\tilde{y}_t = \tilde{x}_t m(z_t) + \varepsilon_t = \tilde{x}_t w(z_t, \beta) + \varepsilon_t,$$

the uniform consistency (Assumption 3.C2) of the local constant estimator $\hat{m}_T(z)$ defined in (3.12) in $z_t \in \mathcal{Z}_T^c$ is verified in Theorem 5 in Čížek and Koo (2017a), see also the proof of (3.32) below, and the (asymptotic) piecewise differentiability of the estimated transition functions follows directly from Assumption 3.E5 and Theorem 4 in Čížek and Koo (2017a).

Therefore, only the stochastic equicontinuity of $\hat{m}_T(z;\beta)$ in $\beta \in U(\beta^0, \delta)$ on $\mathcal{Z}_T^c$ (Assumption 3.C3) remains to be verified, where the dependence of $\hat{m}_T(z;\beta)$ on β is made explicit and stems from $\tilde{y}_t, \tilde{x}_t$ being functions of β and the true regression function is also denoted $m(z) = w(z;\beta)$ from now on to highlight its dependence of β. Considering only sequences $\zeta_T > h_T$, the corresponding left, right, and centered estimators in (3.12) are equal for $\iota = l, r, c$ to

$$\hat{a}_T^{(\iota)}(z;\beta) = \left\{ \frac{1}{T} \sum_{t=1}^{T} \tilde{x}_t \tilde{x}_t^\top K_h^{(\iota)}(z_t - z) \right\}^{-1} \frac{1}{T} \sum_{t=1}^{T} \tilde{x}_t \tilde{y}_t K_h^{(\iota)}(z_t - z), \tag{3.29}$$

where $\tilde{y}_t = y_t - x_t^\top \beta_1$, $\tilde{x}_t = x_t^\top(\beta_2 - \beta_1)$, and $\beta = (\beta_1^\top, \beta_2^\top)^\top$, see (3.12). As we consider only $\beta \in U(\beta^0, \delta)$ and $\|\beta - \tilde{\beta}\| < \delta$, Assumption 3.A2 implies that $\delta > 0$ can be so small that $\inf_{\beta \in U(\beta^0, 2\delta)} \|\beta_1 - \beta_2\| > 0$, that is, $\beta_1 \neq \beta_2$ for any β and $\tilde{\beta}$ considered.

Now, let

$$S_n^{(\iota)}(z;\beta) = T^{-1}\sum_{t=1}^{T}(\beta_2-\beta_1)^\top x_t x_t^\top(\beta_2-\beta_1)K_h^{(\iota)}(z_t-z) \quad \text{and}$$
$$T_n^{(\iota)}(z;\beta) = T^{-1}\sum_{t=1}^{T}(\beta_2-\beta_1)^\top x_t(y_t-x_t^\top\beta_1)K_h^{(\iota)}(z_t-z).$$

Then $\hat{a}_T^{(\iota)}(z;\beta) = {S_n^{(\iota)}}^{-1}(z;\beta)T_n^{(\iota)}(z;\beta)$. By Lemma 3 in Čížek and Koo (2017a), it follows for $T\to\infty$ that

$$\begin{aligned}
&\sup_{z\in\mathcal{Z}_T^c,\beta\in U(\beta^0,2\delta)}\left\|S_n^{(\iota)}(z;\beta)-(\beta_2-\beta_1)^\top\mu_0^{(\iota)}f_z(z)\Omega(z)(\beta_2-\beta_1)\right\| \\
&= \sup_{z\in\mathcal{Z}_T^c,\beta\in U(\beta^0,2\delta)}\left\|(\beta_2-\beta_1)^\top\left\{T^{-1}\sum_{t=1}^{T}x_tx_t^\top K_h^{(\iota)}(z_t-z)-\mu_0^{(\iota)}f_z(z)\Omega(z)\right\}(\beta_2-\beta_1)\right\| \\
&\le \sup_{\beta\in U(\beta^0,2\delta)}\|\beta_2-\beta_1\|^2\sup_{z\in\mathcal{Z}_T^c}\left\|T^{-1}\sum_{t=1}^{T}x_tx_t^\top K_h^{(\iota)}(z_t-z)-\mu_0^{(\iota)}f_z(z)\Omega(z)\right\| \xrightarrow{P} 0.
\end{aligned}$$

A similar argument can be used for $T_n^{(\iota)}(z;\beta)$ after substituting $y_t - x_t^\top\beta_1 = x_t^\top(\beta_2-\beta_1)w(z_t;\beta)+\varepsilon_t$ from model (3.10). Specifically under Assumption 3.E, $w(z;\beta) = \mathrm{E}[(\beta_2-\beta_1)^\top x_t(y_t-x_t^\top\beta_1)|z_t=z]/\mathrm{E}[(\beta_2-\beta_1)^\top x_t x_t^\top(\beta_2-\beta_1)|z_t=z]$ defined in (3.5) can be rewritten in the following form: $w(z;\beta) = (\beta_2-\beta_1)^\top\{\mathrm{E}[x_ty_t|z_t=z]-\mathrm{E}[x_tx_t^\top|z_t=z]\beta_1\}/(\beta_2-\beta_1)^\top\mathrm{E}[x_tx_t^\top|z_t=z](\beta_2-\beta_1)$. For $\beta\in U(\beta^0,2\delta)$ and $z\in\mathcal{Z}_T^c$, $w(z;\beta)$ is therefore differentiable in z by Assumption 3.E2 and 3.A1 and $w(z;\beta)$ and its derivative are uniformly bounded in z and β by Assumption 3.E3 (see page 125 for more details). The mean value theorem thus implies $w(z_t;\beta) = w(z;\beta)+\mathrm{o}(1), T\to\infty$, uniformly in $z\in\mathcal{Z}_T^c$, $|z_t-z|\le h_T<\zeta_T$, and $\beta\in U(\beta^0,2\delta)$ (see Assumption 3.E5). We can therefore

write

$$
\begin{aligned}
T_n^{(\iota)}(z;\beta) &= T^{-1}\sum_{t=1}^{T}(\beta_2-\beta_1)^\top x_t(y_t - x_t^\top\beta_1)K_h^{(\iota)}(z_t-z) \\
&= T^{-1}\sum_{t=1}^{T}(\beta_2-\beta_1)^\top x_t\{w(z_t;\beta)x_t^\top(\beta_2-\beta_1)+\varepsilon_t\}K_h^{(\iota)}(z_t-z) \\
&= (\beta_2-\beta_1)^\top\cdot T^{-1}\sum_{t=1}^{T}x_tx_t^\top K_h^{(\iota)}(z_t-z)\cdot(\beta_2-\beta_1)w(z;\beta)(1+\mathrm{o}(1)) \qquad (3.30)\\
&\quad + (\beta_2-\beta_1)^\top\cdot T^{-1}\sum_{t=1}^{T}x_t\varepsilon_tK_h^{(\iota)}(z_t-z), \qquad (3.31)
\end{aligned}
$$

where we substituted for $y_t - x_t^\top\beta_1$ from (3.10). The second term (3.31) is asymptotically negligible by Čížek and Koo (2017a, Lemma 3) uniformly in $z\in\mathcal{Z}_T^c$. The same lemma also implies for the first term (3.30) (without the o(1) term) that for $T\to\infty$

$$
\begin{aligned}
&\sup_{z\in\mathcal{Z}_T^c,\beta\in U(\beta^0,2\delta)}\Big\|(\beta_2-\beta_1)^\top\cdot T^{-1}\sum_{t=1}^{T}x_tx_t^\top K_h^{(\iota)}(z_t-z)\cdot(\beta_2-\beta_1)w(z;\beta) \\
&\qquad -(\beta_2-\beta_1)^\top\mu_0^{(\iota)}f_z(z)\Omega(z)(\beta_2-\beta_1)w(z;\beta)\Big\| \\
&= \sup_{z\in\mathcal{Z}_T^c,\beta\in U(\beta^0,2\delta)}\Big\|(\beta_2-\beta_1)^\top\left\{T^{-1}\sum_{t=1}^{T}x_tx_t^\top K_h^{(\iota)}(z_t-z)-\mu_0^{(\iota)}f_z(z)\Omega(z)\right\} \\
&\qquad \cdot(\beta_2-\beta_1)w(z;\beta)\Big\| \\
&\le \sup_{z\in\mathcal{Z}_T^c,\beta\in U(\beta^0,2\delta)}\|\beta_2-\beta_1\|^2\,|w(z;\beta)|\left\|T^{-1}\sum_{t=1}^{T}x_tx_t^\top K_h^{(\iota)}(z_t-z)-\mu_0^{(\iota)}f_z(z)\Omega(z)\right\| \xrightarrow{P} 0
\end{aligned}
$$

due to the boundedness of the parameter space and $w(z;\beta)$.

Subsequently, estimator $\hat{a}_T^{(\iota)}(z;\beta) = [S_n^{(\iota)}(z;\beta)]^{-1}T_n^{(\iota)}(z;\beta)$ converges to $[\mu_0^{(\iota)}f_z(z)\Omega(z)]^{-1}$ $\mu_0^{(\iota)}f_z(z)\Omega(z)w(z;\beta) = w(z;\beta)$ uniformly in $\beta\in U(\beta^0,2\delta)$ and $z\in\mathcal{Z}_T^c$. Since the uniform limits of $S_n^{(\iota)}(z;\beta)$, $T_n^{(\iota)}(z;\beta)$, and $\hat{a}_T^{(\iota)}(z;\beta)$ are independent of $\iota = c,l,r$, it follows as in the proof of Theorem 5 in Čížek and Koo (2017a) for $T\to\infty$ that

$$
\sup_{z\in\mathcal{Z}_T^c,\beta\in U(\beta^0,2\delta)}|\hat{m}_T(z;\beta)-w(z;\beta)|\xrightarrow{P}0. \qquad (3.32)
$$

Given this uniform convergence result, the stochastic equicontinuity Assumption 3.C3

can be verified by proving the uniform continuity of $w(z;\beta) = (\beta_2 - \beta_1)^\top\{\mathrm{E}[x_t y_t|z_t = z] - \mathrm{E}[x_t x_t^\top|z_t = z]\beta_1\}/(\beta_2 - \beta_1)^\top\mathrm{E}[x_t x_t^\top|z_t = z](\beta_2 - \beta_1)$ in β since for $\beta \in U(\beta^0, \delta)$ and $\|\tilde{\beta} - \beta\| < \tilde{\delta} < \delta$

$$|\hat{m}_T(z;\beta) - \hat{m}_T(z;\tilde{\beta})| \leq |\hat{m}_T(z;\beta) - w(z;\beta)| + |\hat{m}_T(z;\tilde{\beta}) - w(z;\tilde{\beta})| + |w(z;\beta) - w(z;\tilde{\beta})|.$$

Hence, it remains us to verify for $\tilde{\delta} \to 0$ that

$$\sup_{z\in\mathcal{Z}_T^c,\beta\in U(\beta^0,\delta),\tilde{\beta}\in U(\beta^0,\tilde{\delta})} |w(z;\beta) - w(z;\tilde{\beta})| \to 0. \tag{3.33}$$

Difference $|w(z;\beta) - w(z;\tilde{\beta})|$ can be rewritten as

$$\left|\frac{(\beta_2 - \beta_1)^\top\{\mathrm{E}[x_t y_t|z_t = z] - \mathrm{E}[x_t x_t^\top|z_t = z]\beta_1\}}{(\beta_2 - \beta_1)^\top\mathrm{E}[x_t x_t^\top|z_t = z](\beta_2 - \beta_1)} - \frac{(\tilde{\beta}_2 - \tilde{\beta}_1)^\top\{\mathrm{E}[x_t y_t|z_t = z] - \mathrm{E}[x_t x_t^\top|z_t = z]\tilde{\beta}_1\}}{(\tilde{\beta}_2 - \tilde{\beta}_1)^\top\mathrm{E}[x_t x_t^\top|z_t = z](\tilde{\beta}_2 - \tilde{\beta}_1)}\right|$$

and

$$\left|\frac{(\beta_2 - \beta_1)^\top\{\mathrm{E}[x_t y_t|z_t = z] - \mathrm{E}[x_t x_t^\top|z_t = z]\beta_1\}(\tilde{\beta}_2 - \tilde{\beta}_1)^\top\mathrm{E}[x_t x_t^\top|z_t = z](\tilde{\beta}_2 - \tilde{\beta}_1)}{(\beta_2 - \beta_1)^\top\mathrm{E}[x_t x_t^\top|z_t = z](\beta_2 - \beta_1)(\tilde{\beta}_2 - \beta_1)^\top\mathrm{E}[x_t x_t^\top|z_t = z](\tilde{\beta}_2 - \tilde{\beta}_1)} - \frac{(\beta_2 - \beta_1)^\top\mathrm{E}[x_t x_t^\top|z_t = z](\beta_2 - \beta_1)(\tilde{\beta}_2 - \tilde{\beta}_1)^\top\{\mathrm{E}[x_t y_t|z_t = z] - \mathrm{E}[x_t x_t^\top|z_t = z]\tilde{\beta}_1\}}{(\beta_2 - \beta_1)^\top\mathrm{E}[x_t x_t^\top|z_t = z](\beta_2 - \beta_1)(\tilde{\beta}_2 - \beta_1)^\top\mathrm{E}[x_t x_t^\top|z_t = z](\tilde{\beta}_2 - \tilde{\beta}_1)}\right|.$$

Given that the nominator is a quadratic function of β and $\tilde{\beta}$ on a bounded set, we only have prove that (i) the conditional expectations are uniformly continuous in z and (ii) the nominator is uniformly bounded and the denominator is uniformly bounded away from zero. The latter point follows directly from the compactness of the parameter space, $\inf_{\beta\in U(\beta^0,2\delta)} \|\beta_1 - \beta_2\| > 0$ stated earlier, and Assumption 3.E4. The first point, the uniform continuity of the expectations $\mathrm{E}[x_t y_t|z_t = z]$ and $\mathrm{E}[x_t x_t^\top|z_t = z]$ in $z \in \mathcal{Z}_T^c$,

follows from Assumption 3.E2 because for example the first derivative

$$\begin{aligned}\frac{\partial \mathrm{E}[x_t x_t^\top | z_t = z]}{\partial z} &= \frac{\partial}{\partial z} \int\int x x^\top \frac{f(x,z)}{f(z)} \mathrm{d}x \\ &= \int\int x x^\top \frac{f(z)\partial f(x,z)/\partial z - f(x,z)\partial f(z)/\partial z}{f^2(z)} \mathrm{d}x \\ &= \int\int x x^\top \left[\frac{f(x,z)\partial \ln f(x,z)/\partial z}{f(z)} - \frac{f(x,z)\partial \ln f(z)/\partial z}{f(z)} \right] \mathrm{d}x \\ &= \mathrm{O}(1)\mathrm{E}[x_t x_t^\top \frac{\partial \ln f(x,z)}{\partial z} | z_t = z] + \mathrm{O}(1)\mathrm{E}[x_t x_t^\top | z_t = z]\end{aligned}$$

is uniformly bounded on $\mathcal{Z}_T^c$ due to its compactness and Assumption 3.E3. Hence, (3.33) is verified and the stochastic equicontinuity condition follows by (3.32) as mentioned above.

Chapter 4

Functional Coefficient Models with Endogenous Variables

4.1 Introduction

Instrumental variable (IV) models provide a useful framework which correctly accounts for endogeneity and identifies causal relationships among several economic variables. Parametric IV models are sometimes too restrictive due to their tight functional forms and their misspecification of can lead to inconsistent estimates and wrong inference. To relax such a restriction, nonparametric IV models were introduced by Newey and Powell (2003) and further developed by Newey et al. (1999), Hall and Horowitz (2005), and Blundell et al. (2007). Similarly to ordinary nonparametric models, nonparametric IV models suffer from the 'curse of dimensionality' problem as the number of regressors is large. To alleviate this problem, semiparametric IV models were developed (see e.g., Ai and Chen, 2003; Park, 2003; Cai et al., 2006).

Among the semiparametric IV models, we consider a functional coefficient IV model which is linear in endogenous structural regressors with their coefficients given by unknown functions of some observed transition variables. The functional coefficient IV models were first introduced by Cai et al. (2006) who assumed the transition variables to be exogenous so that the ill-posed inverse problem in a general nonparametric IV framework is avoided. They proposed a two-stage estimation procedure. First, the expectations of endogenous

regressors conditional on a set of instrumental variables are estimated nonparametrically. Second, a local linear regression of the response variable on the estimated conditional expectations of the regressors is performed. Later, Cai and Xiong (2012) and Cai et al. (2017) studied a partially varying-coefficient IV model with both constant and functional coefficients and developed estimation procedures, which achieve the $\sqrt{n}$-convergence rate of the estimates of the constant coefficients. Recently, Su and Hoshino (2015) considered a sieve-based quantile regression estimation of functional coefficient IV models and established uniform consistency and asymptotic normality of the nonparametric estimators.

The above mentioned functional coefficient IV literature relies on the assumption that the transition variables are exogenous. In this chapter, we therefore consider a model similar to Cai et al. (2006)'s model, but now with endogenous transition variables and weakly dependent observations. In such a case, the ill-posed inverse problem will still be present if using the orthogonality condition in Cai et al. (2006, equation (1)). To resolve such a problem, we employ an alternative orthogonality condition similarly to Newey et al. (1999) and Su and Ullah (2008) for nonparametric IV models (see the orthogonality condition (4.2) and its discussion in Section 4.2).

The rest of the chapter is structured as follows. The model and the identification results are presented in Section 4.2. In Section 4.3, we propose a two-stage estimation procedure, and in Section 4.4, establish asymptotic normality of the proposed estimator. Section 4.5 provides a simulation study to investigate its finite sample properties and compare them to Cai et al. (2006)'s two-stage estimation. Then, an empirical example is provided for studying the marginal return to schooling. Proofs of the main theorems and related lemmas are collected in Sections 4.7 and 4.8.

4.2 Model specification and identification

We consider a partially linear functional coefficient model of the following form:

$$\begin{cases} Y = \sum_{l=0}^{d_1} a_l(X_2)X_{1l} + \sum_{l=1}^{r_1} \gamma_l Z_{1l} + \varepsilon = X_1^\top a(X_2) + Z_1^\top \gamma + \varepsilon, \\ X = \Pi(Z) + U, \quad X = (X_{11}, \ldots, X_{1d_1}, X_2^\top)^\top, \quad Z = (Z_1^\top, Z_2^\top)^\top, \\ \mathrm{E}(U|Z) = 0, \end{cases} \tag{4.1}$$

where Y is a dependent variable, $X_{10} \equiv 1$, $X_1 = (X_{10}, X_{11}, \ldots, X_{1d_1})^\top$ is a $(d_1+1)\times 1$ vector of covariates, X_2 is a $d_2 \times 1$ vector of transition variables, $a(\cdot) = (a_0(\cdot), \ldots, a_{d_1}(\cdot))^\top$ is a $(d_1+1)\times 1$ vector of unknown coefficient functions, Z_1 is an $r_1 \times 1$ vector of exogenous random variables, $\gamma = (\gamma_1, \ldots, \gamma_{r_1})^\top$ is an $r_1 \times 1$ vector of constant coefficients, Z_2 is an $r_2 \times 1$ vector of instrumental variables, $\Pi(\cdot) = (\Pi_1(\cdot), \ldots, \Pi_d(\cdot))^\top$ is a $d \times 1$ vector of functions of an $r \times 1$ vector Z, $d = d_1 + d_2$ and $r = r_1 + r_2$, ε and U are disturbances, and $^\top$ denotes transpose of a matrix or vector. In this chapter, both variables X_1 and X_2 are allowed to be endogenous. We further impose X_2 to be a scalar variable, $d_2 = 1$, although the proposed estimation procedure in Section 4.3 can be readily extended to multiple transition variables.

Model (4.1) includes many popular IV models. When X_2 is exogenous, but not X_1, the model is reduced to (partially linear) functional coefficient IV models studied by Cai et al. (2006), Cai and Xiong (2012), and Su et al. (2013). If X_1 is further a binary endogenous variable, model (4.1) becomes the nonparametric IV model considered in Das (2005). Caner and Hansen (2004) studied a threshold IV model, where $a(\cdot)$ is a parametric threshold function of an exogenous transition variable and X_1 is endogenous. Recently, a smooth transition IV model, which assumes $a(\cdot)$ to be a logistic function, with both endogenous variables X_1 and X_2 was analyzed by Areosa et al. (2011).

Model (4.1) is different from the ordinary functional coefficient model in the sense that $\mathrm{E}(Y|X, Z_1) \neq X_1^\top a(X_2) + Z_1^\top \gamma$ when $\mathrm{E}(\varepsilon|X, Z_1) \neq 0$. Accordingly, the coefficient functions $\{a_l(\cdot)\}_{l=0}^{d_1}$ cannot be consistently estimated by projecting Y on $X_1^\top a(X_2)$ and $Z_1^\top \gamma$.

To retrieve the coefficient functions, we impose the following conditional mean independence restriction:

$$\mathrm{E}(\varepsilon|U,Z) = \mathrm{E}(\varepsilon|U), \tag{4.2}$$

which is also used in some nonparametric IV literature (e.g., Newey et al., 1999; Su and Ullah, 2008; Han, 2014). Model (4.1) under (4.2) implies that for $\lambda(U) = \mathrm{E}(\varepsilon|U)$,

$$\begin{aligned}
\mathrm{E}(Y|X,Z,U) &= X_1^\top a(X_2) + Z_1^\top \gamma + \mathrm{E}(\varepsilon|X,Z,U) \\
&= X_1^\top a(X_2) + Z_1^\top \gamma + \mathrm{E}(\varepsilon|\Pi(Z)+U,Z,U) \\
&= X_1^\top a(X_2) + Z_1^\top \gamma + \mathrm{E}(\varepsilon|U,Z) \\
&= X_1^\top a(X_2) + Z_1^\top \gamma + \lambda(U) \\
&= X_1^\top b(X_2,U) + Z_1^\top \gamma,
\end{aligned} \tag{4.3}$$

where $b(x_2,u) = [a_0(x_2)+\lambda(u), a_1(x_2), \cdots, a_{d_1}(x_2)]^\top$ is identical to the nonparametric component $a(x_2)$ from an ordinary functional coefficient model except that its first element is replaced by a sum of functions $a_0(x_2)$ and $\lambda(u)$. The above equation indicates that $a(\cdot)$ could be retrieved from a local polynomial fitting of Y on X, Z_1, and U. Since variable U is not observable, we suggest a two-stage estimation procedure. In the first stage, the predicted residual U is obtained by regressing X on Z. In the second stage, $a(x_2)$ is estimated by regressing Y on X, Z_1, and the predicted residual U locally around x_2. The estimation procedure is discussed in details in Section 4.3.

Notice that the orthogonality condition (4.2) requires that $\mathrm{E}(\varepsilon|U,Z)$ depends on U only, which is different from the orthogonality condition $\mathrm{E}(\varepsilon|Z)=0$ commonly imposed on the functional coefficient IV literature, like Cai et al. (2006), Cai and Li (2008), and Cai and Xiong (2012). It is more general than requiring ε and U are independent of Z. If U is independent of Z and $\mathrm{E}(\varepsilon)=0$, then $\mathrm{E}[\varepsilon|Z] = \mathrm{E}[\mathrm{E}[\varepsilon|Z,U]|Z] = \mathrm{E}[\mathrm{E}[\varepsilon|U]|Z] = \mathrm{E}(\varepsilon) = 0$. If no additional restrictions assumed, neither functional coefficient IV models is more general than the other.

A typical application of model (4.1) under (4.2) is a demand model. For example, Y can be the purchased quantities of apples by a household, X_1 the price of apples, X_2 the

household expenditure on fruits, Z_1 the household income, Z_2 the age of family head. The unobserved residuals ε and U include demand shocks and individual preferences. Endogeneity comes from expenditure on fruits being a variable affecting the price elasticity. The condition $E[\varepsilon|U,Z] = E[\varepsilon|U]$ means that the average unexplained variation of the expenditure on apples depends on variation of age only through the unexplained variation of the expenditure on fruits.

If the residual U can be recovered from $X - \Pi(Z)$, the identification of $a(\cdot)$ in model (4.1) with the orthogonality condition (4.2) is equivalent to the identification under equation (4.3). The following theorem provides a sufficient condition for identification of $a(\cdot)$, which is similar to the rank condition for identification in a linear structural equation with endogenous covariates.

Theorem 4.1. *Suppose that the variables X, Z, and U have compact supports and continuously differentiable distribution functions and let $\Pi_+(Z) = \mathrm{E}((Z_1^\top, X^\top)^\top|Z) = (\mathrm{E}(Z_1^\top|Z), \Pi^\top(Z))^\top$. If the functions $a(\cdot)$, $\lambda(\cdot)$, and $\Pi(\cdot)$ are differentiable for all x_2, u, and z, respectively, and with probability one, $\mathrm{rank}(\partial\Pi_+(z)/\partial z^\top)^\top = r_1 + d_1 + 1$ for all z on its support, then $\{a_l(\cdot)\}_{l=1}^{d_1}$ and γ are identified, and $a_0(\cdot)$ and $\lambda(\cdot)$ are identified up to additive constants. If we further assume $\mathrm{E}(\varepsilon) = 0$, then $a_0(\cdot)$ and $\lambda(\cdot)$ are also identified.*

Theorem 4.1 requires that the number of instrumental variables in Z_2 is at least as large as the number of the nonconstant variables in X_1 and X_2. It is similar to the identification conditions in the nonparametric simultaneous equations models (Newey et al., 1999, Theorem 2.3) and for Cai et al.'s functional coefficient IV model (Cai et al., 2006, Theorem 1).

4.3 Estimation

In this section, we propose a two-stage estimator of the coefficient function $a_l(\cdot)$, $l = 0, \ldots, d_1$, in model (4.1) under the conditional mean independence restriction (4.2) and the identification condition $\mathrm{E}(\varepsilon) = 0$ based upon local polynomial fitting and marginal integration techniques.

Let us first suppose that U is observable. Then model (4.3) becomes a special case of the semiparametric additive coefficient model in Xue and Yang (2006). It is clear that the functions $b_l(\cdot,\cdot)$, $l = 0,\ldots,d_1$, in (4.3) can be estimated consistently by a local polynomial fitting of an ordinary varying-coefficient model. Under the identification condition $\mathrm{E}(\varepsilon) = \mathrm{E}(\lambda(U)) = 0$, the coefficient function $a_l(x_2)$ is a partial mean of $b_l(x_2,U)$, $a_l(x_2) = \mathrm{E}[b_l(x_2,U)]$, for any fixed point x_2 and $l = 0,\ldots d_1$. Following Xue and Yang (2006), we employ the marginal integration method to estimate $a_l(x_2)$:

$$\tilde{a}_l(x_2) = \frac{1}{n}\sum_{i=1}^{n}\tilde{b}_l(x_2,U_i), \tag{4.4}$$

where $\tilde{b}_l(\cdot,\cdot)$ is some consistent estimator of $b_l(\cdot,\cdot)$.

Since U is unobservable, a two-stage estimation procedure is proposed. In the first step, we estimate $\Pi_m(Z_i)$ nonparametrically for $m = 1,\ldots,d$ and $i = 1,\ldots,n$. Denoting the estimates $\hat{\Pi}(Z_i) = [\hat{\Pi}_1(Z_i),\ldots,\hat{\Pi}_d(Z_i)]^\top$, we can estimate the residuals $\hat{U}_i = (\hat{U}_{1i},\ldots,\hat{U}_{di})^\top$ by $\hat{U}_{mi} = X_{mi} - \hat{\Pi}_m(Z_i)$. In the second step, we estimate the coefficient function $b_l(x_2,U_i)$ for $i = 1,\ldots,n$ using the estimated residuals $\hat{U}_i$ in place of the unobservable U_i and recover $a_l(x_2)$ using equation (4.4).

To obtain the first stage estimates $\hat{\Pi}_m(Z_i)$ for $m = 1,\ldots,d$ and $i = 1,\ldots,n$, we apply the local p_1th-order polynomial fitting and leave-one-out techniques. Assuming that $\Pi_m(\cdot)$ has Lipschitz continuous p_1 partial derivatives, $\Pi_m(\cdot)$ can be approximated locally at Z_i by a p_1th degree polynomial. For Z_k in a neighborhood of Z_i,

$$\Pi_m(Z_k) \approx \sum_{0\le|\mathbf{q}|\le p_1}\frac{1}{\mathbf{q}!}D^{\mathbf{q}}\Pi_m(Z_i)(Z_k - Z_i)^{\mathbf{q}},$$

where

$$\mathbf{q} = (q_1,\ldots,q_r)^\top,\quad \mathbf{q}! = \prod_{j=1}^{r}q_j!,\quad |\mathbf{q}| = \sum_{j=1}^{r}q_j,\quad z^{\mathbf{q}} = \prod_{j=1}^{r}z_j^{q_j},$$

$$\sum_{0\le|\mathbf{q}|\le p_1} = \sum_{j=0}^{p_1}\sum_{\substack{q_1=0\\ q_1+\cdots+q_r=j}}^{j}\cdots\sum_{q_r=0}^{j},\quad \text{and}\quad D^{\mathbf{q}}\Pi_m(z) = \frac{\partial^{|\mathbf{q}|}\Pi_m(z)}{\partial z_1^{q_1}\cdots\partial z_r^{q_r}}.$$

Following the first stage estimation procedure in Cai et al. (2006), we leave out the ith observation and use all other observations to estimate $\Pi_m(Z_i)$. The advantage of using the leave-one-out technique is eliminating the dependence between the functions $\{\hat{\Pi}_m(Z_i)\}_{m=1}^d$ and the ith observation (Y_i, X_i, Z_i), which might complicate the second-stage estimation. Given the data set $\{X_{mk}, Z_k\}_{k=1,\ldots,n,k\neq i}$, we consider the local polynomial estimator minimizing the following weighted sum of squared residuals:

$$\frac{1}{nh_1^r}\sum_{k=1,k\neq i}^{n} K_1\left(\frac{Z_k - Z_i}{h_1}\right)\left[X_{mk} - \sum_{0\leq|\mathbf{q}|\leq p_1}\theta_{\mathbf{q}}^{(m)}(Z_k - Z_i)^{\mathbf{q}}\right]^2,$$

where $K_1(\cdot)$ is a kernel function on $\mathbb{R}^r$ and $h_1 = h_1(n) > 0$ is a scalar bandwidth. The local leave-one-out estimator of $\Pi_m(Z_i)$ is defined as $\hat{\Pi}_m(Z_i) = \hat{\theta}_0^{(m)}$, where $\{\hat{\theta}_{\mathbf{q}}^{(m)}\}_{0\leq|\mathbf{q}|\leq p_1}$ minimizes the above weighted sum of squared residuals.

Following the notation of Masry (1996), let $N_j = (j+r-1)!/(j!(r-1)!)$ be the number of distinct r-tuples $\mathbf{q}$ with $|\mathbf{q}| = j$. Arrange the N_j r-tuples as a sequence in a lexicographical order (with highest priority to last position so that $(0,\ldots,0,j)$ is the first element in the sequence and $(j,0,\ldots,0)$ is the last element) and denote ϕ_j^{-1} this one-to-one map. Let $\rho_1(v) = [1, \rho_{1,1}^\top(v), \ldots, \rho_{1,p_1}^\top(v)]^\top$ in which $\rho_{1,j}(v)$ is a $N_j \times 1$ subvector with its lth element given by $[\rho_{1,j}(v)]_l = v^{\phi_j(l)}$. Then the estimator of $\Pi_m(Z_i)$ can be expressed as

$$\hat{\Pi}_m(Z_i) = e_1^\top \bar{S}_n^{-1}(Z_i)\bar{T}_n(Z_i), \tag{4.5}$$

where $e_1 = (1,0,\ldots,0)^\top$,

$$\bar{S}_n(Z_i) = \frac{1}{nh_1^r}\sum_{k=1,k\neq i}^{n} K_1\left(\frac{Z_k - Z_i}{h_1}\right)\rho_1\left(\frac{Z_k - Z_i}{h_1}\right)\rho_1^\top\left(\frac{Z_k - Z_i}{h_1}\right),$$

and

$$\bar{T}_n(Z_i) = \frac{1}{nh_1^r}\sum_{k=1,k\neq i}^{n} K_1\left(\frac{Z_k - Z_i}{h_1}\right)\rho_1\left(\frac{Z_k - Z_i}{h_1}\right)X_{mk}.$$

Given the estimate $\hat{\Pi}_m(Z_i)$ in (4.5), one can obtain the estimated residuals U_{mi} by $\hat{U}_{mi} = X_{mi} - \hat{\Pi}_m(Z_i)$ for $m = 1,\ldots,d$ and $i = 1,\ldots,n$.

Moving to the second stage, we derive first an infeasible estimator of $a_l(x_2)$, assuming $\{U_j\}_{j=1}^n$ are observed, by combining the local polynomial fitting and marginal integration (4.4). Given $\{U_j\}_{j=1}^n$ and assuming the coefficient function $a_l(\cdot)$ has Lipschitz continuous (p_2+1)th derivatives, we can approximate the coefficient function $a_l(\cdot)$ locally at a fixed point x_2. By the Taylor expansion for X_{2j} in a neighborhood of x_2, we have

$$a_l(X_{2j}) \approx \sum_{q=0}^{p_2} \frac{\partial^q a_l(x_2)}{\partial x_2^q}(X_{2j}-x_2)^q.$$

Combined with the local-constant approximation of $\lambda(\cdot)$, the regression function in (4.3) can therefore be estimated by minimizing the following locally weighted sum of squared residuals:

$$\frac{1}{ngh_2^d}\sum_{j=1}^{n}\left[Y_j-\sum_{l=0}^{d_1}\left\{\sum_{q=0}^{p_2}\theta_{lq}(X_{2j}-x_2)^q\right\}X_{1lj}-Z_{1j}^{\top}\vartheta\right]^2 \times L\left(\frac{X_{2j}-x_2}{g}\right)K_2\left(\frac{U_j-U_i}{h_2}\right), \tag{4.6}$$

where $L(\cdot)$ is a univariate kernel function, $K_2(\cdot)$ is a d-variate kernel function of order q_2, and $g=g(n)>0$ and $h_2=h_2(n)>0$ are scalar bandwidths for simplicity. Let $\{\{\tilde{\theta}_{lq}\}_{q=0}^{p_2}\}_{l=0}^{d_1}$ and $\tilde{\vartheta}$ be the minimizers of equation (4.6). The local estimator of $b_l(x_2,U_i)$ is then given by $\tilde{b}_l(x_2,U_i)=\tilde{\theta}_{l0}$. After denoting vectors $\rho_2(v)=(1,v,\ldots,v^{p_2})^{\top}$ and

$$\mathcal{X}_j=\begin{pmatrix}\rho_2\left(\frac{X_{2j}-x_2}{g}\right)\otimes X_{1j}\\ Z_{1j}\end{pmatrix}, \tag{4.7}$$

the estimator of $b_l(x_2,U_i)$ can be written as

$$\tilde{b}_l(x_2,U_i)=e_{l+1}^{\top}\tilde{S}_n^{-1}(x_2,U_i)\tilde{T}_n(x_2,U_i),$$

where e_l is a $(p_2+1)d_1\times 1$ vector with all entries 0 except the lth element being 1,

$$\tilde{S}_n(x_2,U_i)=\frac{1}{ngh_2^d}\sum_{j=1}^{n}L\left(\frac{X_{2j}-x_2}{g}\right)K_2\left(\frac{U_j-U_i}{h_2}\right)\mathcal{X}_j\mathcal{X}_j^{\top},$$

and

$$\tilde{T}_n(x_2, U_i) = \frac{1}{ngh_2^d} \sum_{j=1}^{n} L\left(\frac{X_{2j} - x_2}{g}\right) K_2\left(\frac{U_j - U_i}{h_2}\right) \mathcal{X}_j Y_j.$$

Finally, for any given fixed point x_2, we construct the marginal integration estimator of $a_l(x_2)$, $l = 1, \ldots, d_1$, according to equation (4.4):

$$\tilde{a}_l(x_2) = \frac{1}{n} \sum_{i=1}^{n} \tilde{b}_l(x_2, U_i) = \frac{1}{n} \sum_{i=1}^{n} e_{l+1}^{\top} \tilde{S}_n^{-1}(x_2, U_i) \tilde{T}_n(x_2, U_i). \tag{4.8}$$

In the above estimation procedure, we fit a local constant for $\lambda(U)$ instead of a higher order local polynomial for two reasons. First, the computation for a local constant fitting is less cumbersome when the dimension of U is large. Second, the approximation bias of the proposed estimator due to the local fitting for $\lambda(U)$ can be negligible by taking a small bandwidth h_2 and a higher order kernel $K_2(\cdot)$.

Note that the constant coefficients γ_l, $l = 1, \ldots, r_1$, can be treated as functions of x_2 and u, i.e. $\{\gamma_l(x_2, u)\}_{l=1}^{r_1}$. The minimizer $\tilde{\vartheta}$ from the objective function (4.6) is an estimator of $\gamma(x_2, U_i)$, $\tilde{\gamma}_n(x_2, U_i) = \tilde{\vartheta}$, which uses only data points in the neighborhood of x_2 and U_i and ignore the fact that the functions $\{\gamma_l(x_2, U_i)\}_{l=1}^{r_1}$ are actually constant. Again, by employing the marginal integration technique, a more efficient estimator for γ that employs all all data points can be obtained:

$$\tilde{\gamma} = \frac{1}{n} \sum_{i=1}^{n} \tilde{\gamma}_n(X_{2i}, U_i). \tag{4.9}$$

Note though that deriving its asymptotic properties is beyond the scope of this work. More details on such an estimating procedure for constant coefficients can be found in Zhang et al. (2002) and Cai and Xiong (2012).

Since the residual series $\{U_j\}_{j=1}^{n}$ is unobserved, the above estimators $\tilde{a}_l(x_2)$ and $\tilde{\gamma}$ are both infeasible. The following feasible estimators of $a_l(x_2)$ and γ follow after substituting $\{U_j\}_{j=1}^{n}$ by the estimated residual series $\{\hat{U}_j\}_{j=1}^{n}$ in the infeasible estimators (4.8) and

(4.9):

$$\hat{a}_l(x_2) = \frac{1}{n}\sum_{i=1}^{n} e_{l+1}^{\top}\hat{S}_n^{-1}(x_2, \hat{U}_i)\hat{T}_n(x_2, \hat{U}_i) \tag{4.10}$$

and

$$\hat{\gamma} = \frac{1}{n}\sum_{i=1}^{n} \hat{\gamma}_n(X_{2i}, \hat{U}_i),$$

where $\hat{S}_n(x_2, \hat{U}_i)$, $\hat{T}_n(x_2, \hat{U}_i)$, and $\hat{\gamma}_n(X_{2i}, \hat{U}_i)$ are defined analogously to $\tilde{S}_n(x_2, U_i)$, $\tilde{T}_n(x_2, U_i)$, and $\tilde{\gamma}_n(X_{2i}, U_i)$, respectively, but with the series $\{\hat{U}_j\}_{j=1}^n$ in place of $\{U_j\}_{j=1}^n$.

4.4 Distribution theory

4.4.1 Asymptotic properties and assumptions

Let $\mathcal{F}_a^b$ be the σ-algebra generated by $\{\xi_i; a \le i \le b\}$. The α-mixing coefficient of the process $\{\xi_i\}_{i=-\infty}^{\infty}$ is defined as

$$\alpha(m) = \sup\{|P(A \cap B) - P(A)P(B)| : A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_m^{\infty}\}.$$

If $\alpha(m) \to 0$ as $m \to \infty$, the process $\{\xi_i\}_{i=-\infty}^{\infty}$ is called strong mixing or α-mixing.

First, we make the following assumptions to derive the asymptotic results of the infeasible estimator $\tilde{a}_l(x_2)$ defined in (4.8).

Assumption 4.A. Let vectors $W = (X_1^{\top}, Z_1^{\top})^{\top}$ and $V = (X_2, U^{\top})^{\top}$ and random variable $\epsilon = Y - X_1^{\top}a(X_2) - Z_1^{\top}\gamma - \lambda(U)$.

4.A1. The kernels $K_2(\cdot)$ and $L(\cdot)$ are bounded symmetric functions with compact supports $[-1,1]^d$ and $[-1,1]$, respectively, where the functions $\mathcal{K}_{2,j}(\cdot) = (\cdot)^j K_2(\cdot)$ for all $j \in [0, 2p_2+1]$ and $L(\cdot)$ are Lipschitz continuous, $\int K_2(v)\mathrm{d}v = \int L(v)\mathrm{d}v = 1$, and p_2 is given in condition 4.A2. Furthermore, $K_2(\cdot)$ is a d-variate function of order q_2, that is, $\int K_2(v)v_1^{l_1}\cdots v_d^{l_d}\mathrm{d}v = 0$, for $1 \le l_1 + \cdots + l_d \le q_2 - 1$.

4.A2. The (p_2+1)th partial derivatives of $a(\cdot)$ and the q_2th partial derivatives of $\lambda(\cdot)$ are uniformly bounded and Lipschitz continuous.

4.A3. The process $\{X_{1i}, X_{2i}, Z_{1i}, U_i, \epsilon_i\}_{i=1}^n$ is strictly stationary and strong mixing with α-mixing coefficients $\alpha(m), m \in \mathbb{N}$, that satisfy $\alpha(m) \leq C_2 m^{-s_2}$, where $C_2 < \infty$, and for some $\delta_2 > 0$ and $\eta_2 > 0$,

$$s_2 > \frac{1 + (1+\delta_2)\left\{2 + d + \dfrac{d+1}{\eta_2}\right\}}{\delta_2}.$$

4.A4. The variables X_1, X_2, Z_1, and U have compact supports D_{X_1}, D_{X_2}, D_{Z_1}, and D_U, respectively. The joint probability density function $f(\cdot,\cdot,\cdot,\cdot)$ of (X_1, X_2, Z_1, U) has bounded partial derivatives with respect to U and X_2 uniformly on D_{X_1}, D_{X_2}, D_{Z_1}, and D_U. The marginal densities $f_{X_2U}(\cdot,\cdot)$ of X_2 and U, $f_U(\cdot)$ of U, and $f_{X_2}(\cdot)$ of X_2 are uniformly bounded and $\inf_{x_2 \in D_{X_2}, u \in D_U} f_{X_2U}(x_2, u) > 0$. The q_2th partial derivatives of the marginal density f_U are bounded and continuous.

4.A5. The partial derivative of the conditional variance $\mathrm{E}(\epsilon^2 | X_1 = x_1, X_2 = x_2, Z_1 = z_1, U = u) = \sigma_\epsilon^2(x_1, x_2, z_1, u)$ with respect to X_2 is bounded and continuous uniformly on D_{X_1}, D_{X_2}, D_{Z_1}, and D_U.

4.A6. Let ω represent 1 or any element of matrix $WW^\top$, vector $W\epsilon$, and variable ϵ^2. It has to satisfy the following moment conditions:

(i) $\mathrm{E}|\omega|^{2+\delta_2} < \infty$,

(ii) $\sup_{v \in D_{X_2} \times D_U} \mathrm{E}(|\omega|^{2+\delta_2} | V = v) f_{X_2U}(v) < \infty$

(iii) for all $j \in \mathbb{N}$,

$$\sup_{(v_0, v_j) \in (D_{X_2} \times D_U)^2} \mathrm{E}(|\omega_0 \omega_j| | V_0 = v_0, V_j = v_j) f_{V_0 V_j}(v_0, v_j) < \infty,$$

where $f_{V_0 V_j}(v_0, v_j)$ denote the joint density of V_0 and V_j.

4.A7. The covariance matrix $E(WW^\top | X_2 = x_2, U = u)$ is uniformly bounded. The infimum of eigenvalues of $E(WW^\top | X_2 = x_2, U = u)$ is strictly positive uniformly in $x_2 \in D_{X_2}$ and $u \in D_U$. Additionally, the moment matrix of the kernel $L(\cdot)$, $M_2 = \int \rho_2(v) \rho_2^\top(v) L(v) \mathrm{d}v$, is nonsingular.

4.A8. The bandwidths $g = g(n) \to 0$ and $h_2 = h_2(n) \to 0$ as $n \to \infty$ satisfy that

(i) $ngh_2^{2q_2} \to 0$,

(ii) $ng^{\delta_2/(2+\delta_2)} h_2^{2d(1+\delta_2)/(2+\delta_2)} \to \infty$,

(iii) $ng^{2p_2+3} \to \bar{c} \in [0, +\infty)$, and

(iv) $n^{\theta_2} g h_2^d / \ln n \to \infty$, where

$$\theta_2 = \frac{s_2 - 2 - d - \dfrac{d+1}{\eta_2} - \dfrac{1+s_2}{1+\delta_2}}{s_2 + 2 - d - \dfrac{1+s_2}{1+\delta_2}}.$$

Assumptions 4.A1–4.A8 are similar to those imposed on the additive coefficient model by Xue and Yang (2006). One noteworthy difference is that Xue and Yang (2006) assumed a β-mixing process to apply Lemma 2 of Yoshihara (1976). We relax this to an α-mixing process by applying Lemma 1.2 of Sun and Chiang (1997) and Lemma C.2 of Gao and King (2004). Another difference is that instead of assuming the mixing coefficients to have geometric decay and Cramer's moment condition as it in Xue and Yang (2006), we provide a trade-off among the mixing decay rate, moment and bandwidth conditions in Assumptions 4.A3, 4.A6, 4.A8(ii), and 4.A8(iv). Since we employ a local constant fitting for $\lambda(U)$, Assumptions 4.A2 and 4.A4 require $\lambda(\cdot)$ and the marginal density f_U to have a certain degree of smoothness. Additionally, Assumption 4.A8(i) requires the bandwidth $g = g(n)$ to be chosen in a way such that the bias from the local constant fitting for $\lambda(U)$ becomes negligible. Assumption 4.A8(iii) determines the optimal convergence rate in Theorem 4.2 which establishes the asymptotic normality result for the infeasible estimator of $a_l(\cdot)$.

Combining Assumptions 4.A3 and 4.A8(iv), we get $\theta_2 \in (0, 1]$. Assumption 4.A8(iv) is a strengthening of the condition $ngh_2^d \to \infty$. If variables X_1 and Z_1 are both bounded, we can take $\delta_2 = \infty$. Then, Assumption 4.A8(iv) simplifies to

$$s_2 > 2 + d + \frac{d+1}{\eta_2} \quad \text{and} \quad \theta_2 = \frac{s_2 - 2 - d + (d+1)/\eta_2}{s_2 + 2 - d}.$$

If $\eta_2 = \infty$, this further reduces to $s_2 > 2 + d$ and $\theta_2 = (s_2 - 2 - d)/(s_2 + 2 - d)$, where θ_2 becomes smaller as $s_2 - d$ decreases. In other words, if the decay rate of the

mixing coefficients is small or the dimension of (X_1, X_2) is high, the convergence rates for bandwidths g and h_2 are required to be small. If the mixing coefficients have geometric decay ($s_2 = \infty$), then $\theta_2 = 1$ and Assumption 4.A8(iv) is equivalent to $ngh_2^d \to \infty$ for all $\eta_2 > 0$ and d. Given this, we provide later in equation (4.14) a feasible range of the convergence rate for bandwidth h_2 such that g achieves the optimal rate $\mathrm{O}(n^{1/(2p_2+1)})$ and Assumptions 4.A8(i)–(iv) are satisfied. See Section 4.4.3 for details.

Define the covariance matrix

$$S(x_2, u) = \begin{pmatrix} S_{XX}(x_2, u) & S_{XZ}(x_2, u) \\ S_{ZX}(x_2, u) & S_{ZZ}(x_2, u) \end{pmatrix},$$

where

$$\begin{aligned} S_{XX}(x_2, u) &= \left\{ \int \rho_2(v)\rho_2^\top(v)L(v)\mathrm{d}v \right\} \otimes \mathrm{E}[X_1 X_1^\top | X_2 = x_2, U = u], \\ S_{XZ}(x_2, u) &= S_{ZX}^\top(x_2, u) = \left\{ \int \rho_2(v)L(v)\mathrm{d}v \right\} \otimes \mathrm{E}[X_1 Z_1^\top | X_2 = x_2, U = u], \\ S_{ZZ}(x_2, u) &= \mathrm{E}[Z_1 Z_1^\top | X_2 = x_2, U = u]. \end{aligned}$$

Theorem 4.2. *Suppose that (4.3) holds and* $\mathrm{E}(\lambda(U)) = \mathrm{E}(\varepsilon) = 0$. *Under Assumptions 4.A1–4.A8, it holds for* $l = 0, \ldots, d_1$ *and any fixed point* $x_2 \in D_{X_2}$ *that, as* $n \to \infty$,

$$\sqrt{ng}\{\tilde{a}_l(x_2) - a_l(x_2) - g^{p_2+1}\eta_l(x_2)\} \xrightarrow{d} N\{0, \sigma_l^2(x_2)\},$$

where

$$\sigma_l^2(x_2) = f_{X_2}(x_2)\mathrm{E}\left[\frac{f_U^2(U)}{f_{X_2U}^2(x_2, U)} \sigma_\epsilon^2(X_1, x_2, Z_1, U) \int \mathcal{L}_l^2(v, x_2, U, X_1, Z_1)\,\mathrm{d}v \right] \tag{4.11}$$

and

$$\eta_l(x_2) = \int v^{p_2+1}\mathrm{E}[\mathcal{L}_l(v, x_2, U, X_1, Z_1)X_1^\top]\mathrm{d}v \frac{a^{(p_2+1)}(x_2)}{(p_2+1)!}$$

with

$$\mathcal{L}_l(v, x_2, U, X_1, Z_1) = e_{l+1}^\top S^{-1}(x_2, U) \begin{pmatrix} \rho_2(v) \otimes X_1 \\ Z_1 \end{pmatrix} L(v).$$

The expression of the asymptotic variance in (4.11) suggests a direct estimator by estimating f_{X_2}, f_U, f_{X_2U}, σ_ϵ^2, and $\mathcal{L}_l^2$ via local polynomial smoothing. Detail discussion of the estimation for $\sigma_l^2(x_2)$ is provided in Section 4.4.2. Next, we list the assumptions required to prove that the difference between the infeasible estimator $\tilde{a}(\cdot)$ in (4.8) and the feasible two-stage estimator $\hat{a}(\cdot)$ in (4.10) is asymptotically negligible.

Assumption 4.B.

4.B1. The r-variate kernel $K_1(\cdot)$ is bounded symmetric function with $\int K_1(v)dv = 1$ and has compact support $[-1, 1]^r$. The functions $\mathcal{K}_{1,\mathbf{q}}(v) = v^{\mathbf{q}} K_1(v)$ for all $\mathbf{q}$ with $0 \leq |\mathbf{q}| \leq 2p_1 + 1$ are Lipschitz continuous, where p_1 is defined in condition 4.B2. Further, the kernel $K_2(\cdot)$ from the second stage has bounded second order derivative $K_2(\cdot)$.

4.B2. The $(p_1 + 1)$th partial derivatives of the function $\Pi_m(\cdot)$ are uniformly bounded and Lipschitz continuous for $m = 1, \ldots, d$.

4.B3. The process $\{X_i, Z_{1i}, Z_{2i}\}_{i=1}^n$ is strictly stationary and strong mixing with α-mixing coefficients $\alpha(m), m \in \mathbb{N}$, that satisfy $\alpha(m) \leq C_1 m^{-s_1}$, where $C_1 < \infty$ and s_1 satisfies for some $\delta_1 > 0$ and $\eta_1 > 0$

$$s_1 > \frac{1 + (1 + \delta_1)\left\{1 + r + \frac{r}{\eta_1}\right\}}{\delta_1}.$$

4.B4. The vector of variables $Z = (Z_1^\top, Z_2^\top)^\top$ has a compact support D_Z. The marginal density $f_Z(\cdot)$ of Z is uniformly continuous and bounded such that $\inf_{z \in D_Z} f_Z(z) > 0$.

4.B5. Let ϖ represent any element of vector U. It has to satisfy the following moment conditions:

(i) $\mathrm{E}|\varpi|^{2+\delta_1} < \infty$,

(ii) $\sup_{z\in D_Z} \mathrm{E}(|\varpi|^{2+\delta_1}|Z=z)f_Z(z) < \infty$,

(iii) for all $i \in \mathbb{N}$,

$$\sup_{(z_0,z_i)\in D_Z^2} \mathrm{E}(|\varpi_0\varpi_i||Z_0=z_0, Z_j=z_i)f_{Z_0Z_j}(z_0,z_i) < \infty,$$

where $f_{Z_0Z_i}(z_0,z_i)$ denotes the joint density of (Z_0,Z_i).

4.B6. The matrix $M_1 = \int \rho_1(v)\rho_1^\top(v)K_1(v)\mathrm{d}v$ is nonsingular.

4.B7. The bandwidths $g = g(n) \to 0$, $h_2 = h_2(n) \to 0$, and $h_1 = h_1(n) \to 0$ as $n \to \infty$ satisfy that

(i) $n^{\theta_1}h_1^r/\ln n \to \infty$, where

$$\theta_1 = \frac{s_1 - 1 - r - \dfrac{r}{\eta_1} - \dfrac{1+s_1}{1+\delta_1}}{s_1 + 3 - r - \dfrac{1+s_1}{1+\delta_1}},$$

(ii) $ngh_1^{2(p_1+1)} \to 0$, $ng^{\frac{2\delta}{2+\delta}-1}h_1^{\frac{2r\delta}{2+\delta}}h_2^{\frac{2d\delta}{2+\delta}+2} \to \infty$, $ng^{\frac{2}{2+\delta}}h_1^{\frac{2r}{2+\delta}}h_2^{\frac{2d}{2+\delta}} \to \infty$, and

(iii) $ng^{-1}h_1^{2r}h_2^4/(\ln n)^2 \to \infty$, $ngh_1^{4(p_1+1)}/h_2^4 \to 0$,

where $\delta = \max\{\delta_1,\delta_2\}$ with δ_1 and δ_2 given in Assumptions 4.B3 and 4.A3, respectively.

Assumptions 4.B1–4.B5 and 4.B7(i) are commonly imposed in literature on the local polynomial estimation to establish uniform rate of convergence; see for example Masry (1996) and Hansen (2008). Assumptions 4.B7(ii)–4.B7(iii) are similar to those imposed on the estimation of nonparametric simultaneous equations models in Su and Ullah (2008, Assumption A5). Assumption 4.B7(ii) requires that the estimation bias from the first stage nonparametric estimation should be $\mathrm{o_p}(1/\sqrt{ng})$. According to Masry (1996), $\max_{1\le k\le n}\|\hat{\Pi}(Z_k) - \Pi(Z_k)\| = \mathrm{O_p}(v_{1n} + h_1^{p_1+1})$, where $\|\cdot\|$ is the Euclidean norm and $v_{1n} = \sqrt{\ln n/(nh_1^r)}$. These approximation errors due to the use of the estimated residual series $\{\hat{U}_i\}_{i=1}^n$ are accounted for in Assumption 4.B7(iii): $(h_2^{-1}(v_{1n}+h_1^{p_1+1}))^2 = \mathrm{o}(1/\sqrt{ng})$, where the appearance of h_2^{-1} comes from the use of the Taylor expansion. Note that Assumption 4.B7(iii) also implies that $h_2^{-1}(v_{1n} + h_1^{p_1+1}) = \mathrm{o}(1)$. Section 4.4.3 provides an

example for bandwidth sequences g, h_2 and h_1 which satisfies Assumptions 4.A8 and 4.B7.

Theorem 4.3. *Suppose that model (4.1) is satisfied with the orthogonality condition (4.2) and the identification condition* $\mathrm{E}(\varepsilon) = 0$. *If Assumptions 4.A1–4.A8 and 4.B1–4.B7 hold, then for* $l = 0, \ldots, d_1$ *and any* $x_2 \in D_{X_2}$

$$\sqrt{ng}\{\hat{a}_l(x_2) - \tilde{a}_l(x_2)\} \xrightarrow{P} 0$$

as $n \to \infty$.

Following directly from Theorem 4.2 and 4.3, we finally establish the asymptotic distribution of the proposed two-stage estimator $\hat{a}(x_2)$ in the following theorem.

Theorem 4.4. *Suppose that (4.1) and (4.2) hold and* $\mathrm{E}(\varepsilon) = 0$. *Under Assumptions 4.A1–4.A8 and 4.B1–4.B7, it holds for* $l = 0, \ldots, d_1$ *and any fixed point* $x_2 \in D_{X_2}$ *that as* $n \to \infty$

$$\sqrt{ng}\{\hat{a}_l(x_2) - a_l(x_2) - g^{p_2+1}\eta_l(x_2)\} \xrightarrow{d} N\{0, \sigma_l^2(x_2)\}.$$

4.4.2 Covariance matrix estimation

Rearranging the terms in (4.11), the asymptotic variance $\sigma_l^2(x_2)$ is the $(l+1)$-th diagonal element of the following covariance matrix

$$\mathrm{E}\left(\frac{f_{X_2}(x_2)f_U^2(U)}{f_{X_2U}(x_2,U)} \cdot \frac{S^{-1}(x_2,U)}{f_{X_2U}(x_2,U)} f_{X_2U}(x_2,U)\Omega(x_2,U)\frac{S^{-1}(x_2,U)}{f_{X_2U}(x_2,U)}\right), \tag{4.12}$$

where

$$\Omega(x_2,u) = \begin{pmatrix} \Omega_{XX}(x_2,u) & \Omega_{XZ}(x_2,u) \\ \Omega_{ZX}(x_2,u) & S_{ZZ}(x_2,u) \end{pmatrix},$$

$$\Omega_{XX}(x_2,u) = \left\{\int \rho_2(v)\rho_2^\top(v)L^2(v)\mathrm{d}v\right\} \otimes \mathrm{E}[X_1X_1^\top\epsilon^2|X_2 = x_2, U = u],$$

$$\Omega_{XZ}(x_2,u) = \Omega_{ZX}^\top(x_2,u) = \left\{\int \rho_2(v)L^2(v)\mathrm{d}v\right\} \otimes \mathrm{E}[X_1Z_1^\top\epsilon^2|X_2 = x_2, U = u].$$

Using the above expression, we construct a estimator of $\sigma_l^2(x_2)$:

$$\hat{\sigma}_l^2(x_2) = \left[\frac{1}{n}\sum_{i=1}^{n} \hat{\omega}_n(x_2, \hat{U}_i)\hat{S}_n^{-1}(x_2, \hat{U}_i)\hat{\Omega}_n(x_2, \hat{U}_i)\hat{S}_n^{-1}(x_2, \hat{U}_i)\right]_{(l+1,l+1)}, \tag{4.13}$$

where

$$\hat{\omega}_n(x_2, \hat{U}_i) = \frac{\left[n^{-1}g^{-1}\sum_{j=1}^{n} K\left(\frac{X_{2j}-x_2}{g}\right)\right]\left[n^{-1}h_2^{-d}\sum_{j=1}^{n} L\left(\frac{\hat{U}_j-\hat{U}_i}{h_2}\right)\right]^2}{n^{-1}g^{-1}h_2^{-d}\sum_{j=1}^{n} K\left(\frac{X_{2j}-x_2}{g}\right) L\left(\frac{\hat{U}_j-\hat{U}_i}{h_2}\right)},$$

$$\hat{\Omega}_n(x_2, \hat{U}_i) = \frac{1}{ngh_2^d}\sum_{j=1}^{n} L\left(\frac{X_{2j}-x_2}{g}\right) K_2\left(\frac{\hat{U}_j-\hat{U}_i}{h_2}\right) \mathcal{X}_j\mathcal{X}_j^\top\hat{\epsilon}_j^2,$$

$$\hat{\epsilon}_j = Y_j - \mathcal{X}_j^\top\hat{S}_n^{-1}(X_{2j}, \hat{U}_j)\hat{T}_n(X_{2j}, \hat{U}_j),$$

and $\hat{S}_n^{-1}(x_2, \hat{U}_i)$ is defined in (4.10), which is a consistent estimator of $f_{X_2U}(x_2, \hat{U}_i)S(x_2, \hat{U}_i)$ by Lemma 4.15. Similar to the proof of consistency for $\hat{S}_n(x_2, \hat{U}_i)$, $\hat{\Omega}_n(x_2, u)$ can be shown to be a consistent estimator of $f_{X_2U}(x_2, u)\Omega(x_2, u)$. Since every term in equation (4.13) are consistent estimators of its corresponding term in (4.12), consistency of the variance estimator $\hat{\sigma}_l^2(x_2)$ can be proven easily.

4.4.3 Discussion

The proposed estimator $\hat{a}_l(\cdot)$ is consistent with a convergence rate depending only on the sample size n and the bandwidth $g = g(n)$ for well-chosen first- and second-stage bandwidth sequences $h_1 = h_1(n)$ and $h_2 = h_2(n)$. Like some other kernel-based multi-stage nonparametric procedures (e.g. Xiao et al., 2003; Su and Ullah, 2008), our first stage nonparametric estimation does not have any impact on the asymptotic variance of our final stage estimators. However, such an observation does not hold in general. The asymptotic variance of the two-stage estimator in Cai et al. (2006) does depend on the variation of the estimated reduced form in the first stage and the covariation between the first and second stage.

According to Theorem 4.4, the asymptotically optimal bandwidth of g, denoted by g_{opt}, minimizes the total asymptotic mean integrated squared error (AMISE) of $\{\hat{a}_l(x_2)\}_{l=0}^{d_1}$:

$$\sum_{l=0}^{d_1} \text{AMISE}(\hat{a}_l) = g^{2(p_2+1)} \sum_{l=0}^{d_1} \int \eta_l^2(x_2)\mathrm{d}x_2 + \frac{1}{ng} \sum_{l=0}^{d_1} \int \sigma_l^2(x_2)\mathrm{d}x_2.$$

The optimal bandwidth g_{opt} is found to be

$$g_{\text{opt}} = \left\{ \frac{\sum_{l=0}^{d_1} \int \sigma_l^2(x_2)\mathrm{d}x_2}{2n(p_2+1)\sum_{l=0}^{d_1} \int \eta_l^2(x_2)\mathrm{d}x_2} \right\}^{1/(2p_2+3)}.$$

If we select g that achieves its optimal rate of convergence, Assumption 4.A8 implies that we should choose the other second stage bandwidth h_2 such that $h_2 \propto n^{-\alpha^*}$, where

$$\underline{\alpha}^* \equiv \frac{p_2+1}{q_2} < \alpha^*\gamma^* < \frac{2p_2+3+(p_2+1)\delta}{d(1+\delta)} \equiv \bar{\alpha}^* \tag{4.14}$$

and $\gamma^* = 2p_2 + 3$. Further, by Assumption 4.B7, the first stage bandwidth should be chosen such that $h_1 \propto n^{-\beta^*}$ with

$$\underline{\beta}^* \equiv \max\left\{ \frac{p_2+1}{p_1+1}, \frac{p_2+1+2\alpha^*\gamma^*}{2(p_1+1)} \right\} < \beta^*\gamma^* < \frac{p_2+2-2\alpha^*\gamma^*}{r} \equiv \bar{\beta}^*. \tag{4.15}$$

For example, when $p_2 = 1$, equations (4.14) and (4.15) become

$$\frac{2}{5q_2} < \alpha^* < \frac{5+2\delta}{5d(1+\delta)} \quad \text{and} \quad \max\left\{ \frac{2}{5(p_1+1)}, \frac{1+5\alpha^*}{5(p_1+1)} \right\} < \beta^* < \frac{3-10\alpha^*}{5r}.$$

From the above inequality, the convergence rates for bandwidths h_1 and h_2 are required to be small when d and r are large. The existences of α^* and β^* are ensured if

$$\frac{2(1+\delta)}{5+2\delta} < \frac{q_2}{d} \quad \text{and} \quad \frac{\max\{2, 1+5\alpha^*\}}{3-10\alpha^*} < \frac{p_1+1}{r}.$$

The orders of kernel K_2 and first-stage local fitting should be sufficiently large for high dimensional (X_1, X_2) and (Z_1, Z_2), respectively. If X_1 and Z_1 are both bounded, take $\delta = \infty$, α^* exists if the order of kernel K_2 is larger than the number of non-constant variables in X_1 and X_2, i.e. $d < q_2$.

4.4.4 Bandwidth selection

The nature of the multi-stage estimation complicates bandwidth selection in a semiparametric model. We propose an *ad hoc* bandwidth selection method for i.i.d observations, similar to that discussed in Cai et al. (2006). First, we apply the leave-one-out least squares cross-validation on the first stage fitting to select the bandwidth $\hat{h}_1^{\mathrm{CV}}$. Since $\hat{h}_1^{\mathrm{CV}}$ might have a different rate of convergence from what we desire, we set the data-driven choice of h_1 equal to $\hat{h}_1 = \hat{h}_1^{\mathrm{CV}} n^{1/(2p_1+2+r)-(\underline{\alpha}^*+\bar{\alpha}^*)/(2\gamma^*)}$, where $\underline{\alpha}^*, \bar{\alpha}^*$ and γ^* are defined in (4.14). Additionally, we take the bandwidth for the nonparametric estimated residuals to be $\hat{h}_2 = \hat{h}_1^{\mathrm{CV}} n^{1/(2p_1+2+r)-(\underline{\beta}^*+\bar{\beta}^*)/(2\gamma^*)}$ with $\underline{\beta}^*$ and $\bar{\beta}^*$ given in (4.15). Once we obtain $\hat{h}_1$ and $\hat{h}_2$, we can keep their values fixed and select the remaining bandwidth g via the least squares cross-validation in the second stage:

$$\tilde{g} = \arg\min_{g} \sum_{i=1}^{n} \left[Y_i - \tilde{b}_{0,-i}(X_{2i}, U_i) - \sum_{l=1}^{d_1} \tilde{a}_{l,-i}(X_{2i}) X_{1li} - \sum_{l=1}^{r_1} \tilde{\gamma}_{l,-i} Z_{1li}^{\top} \right]^2,$$

where $\tilde{a}_{l,-i}(\cdot)$, $\tilde{\gamma}_{l,-i}$, and $\tilde{b}_{0,-i}(\cdot,\cdot)$ are the estimates $\tilde{a}_l(\cdot)$, $\tilde{\gamma}_l$, and $\tilde{b}_0(\cdot,\cdot)$ based on all data except of the ith observation. After obtaining $\tilde{g}$, we can adjust the data-driven choice of h_2 to be $\tilde{h}_2 = \tilde{g} n^{1/\gamma^* - (\underline{\beta}^*+\bar{\beta}^*)/(2\gamma^*)}$.

The above mentioned bandwidth selection method requires the use of leave-one-out cross-validation based on i.i.d observations. For weak dependent observations, where regressors do not contain lag dependent variables, leave-one-out cross-validation can still be used. When regressors contain lagged dependent variables, we suggest the modified multi-fold cross-validation method discussed in Cai et al. (2000) be used.

4.5 Simulation and empirical studies

In this section, some simulated examples are used to investigate the finite sample properties of the proposed two-stage estimator. It is compared with the ordinary local linear estimator that ignores the endogeneity issue and the two-stage estimator from Cai et al. (2006), which allows for endogenous regressors in X_1, but not in the transition variable X_2. A real data example is provided in Section 4.5.3.

We evaluate the performance of all these estimators in terms of the mean absolute deviation error (MADE):

$$\mathrm{MADE}_l = \frac{1}{n_{\mathrm{grid}}} \sum_{k=1}^{n_{\mathrm{grid}}} |\tilde{a}_l(x_{2k}) - a_l(x_{2k})|,$$

where x_{2k} $(k = 1, \ldots, n_{\mathrm{grid}})$ are the grid points. For each example, evaluation is based on 500 samples of sizes $n = 200$, 400, and 800.

In all examples, we employ local fifth degree polynomial fittings ($p_1 = 5$) in the first stage and a local linear fitting ($p_2 = 1$) in the second stage for the proposed estimator. We use the second-order Epanechnikov kernel $K(v) = 0.75(1-v^2)I(|v| \leq 1)$ and its product form as a multivariate kernel with bandwidth sequences: $h_1(n) = 2n^{-1/14}$, $h_2(n) = 2n^{-5/24}$, and $g(n) = 2n^{-1/5}\sigma_{X_2}$, where σ_{X_2} is the standard derivation of variable X_2 (the variables Z and $\hat{U}$ are standardized to have the unit variance). These bandwidths are chosen such that inequalities (4.14) and (4.15) are satisfied. The same $g(n)$ is used as the bandwidth for the ordinary local linear estimator and the estimator of Cai et al. (2006) in the second stage. Since Cai et al. (2006) employed local linear fittings in the first stage and required the first stage bandwidth to be $h_1(n) = \mathrm{o}(g(n))$, we take $h_1(n) = 2n^{-2/9}$ in that case instead.

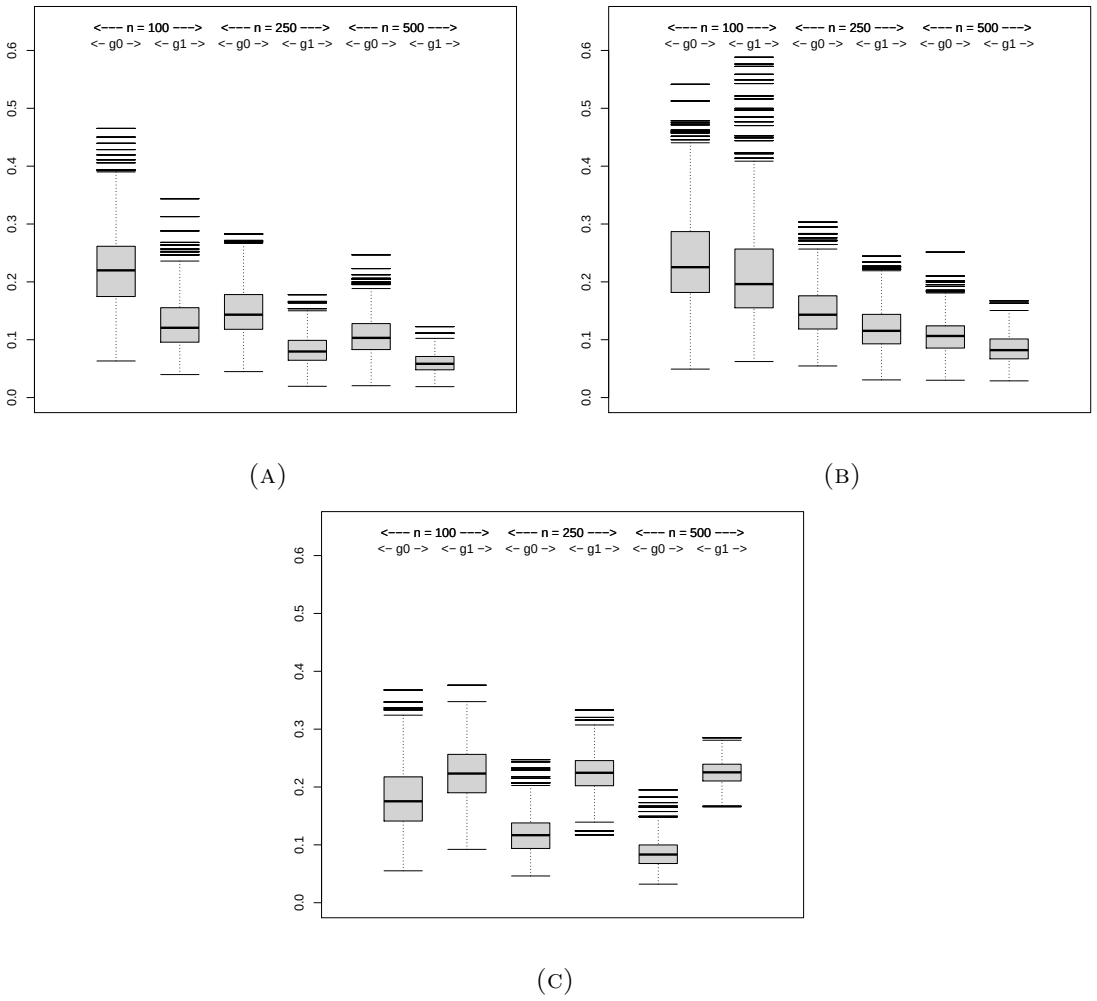
4.5.1 Example 1: iid observations

First, we consider a functional coefficient IV model (Cai et al., 2006, Example 2):

$$\begin{cases} Y = g_0(Z_1) + g_1(Z_1)X_1 + \varepsilon, \\ X_1 = 2\sin(Z_1 + Z_2) + U, \end{cases}$$

where the coefficient functions are defined as

$$g_0(z_1) = \cos(z_1) \quad \text{and} \quad g_1(z_1) = (1 + 0.1z_1)\exp\{-(0.5z_1 - 1.5)^2\},$$



(A) (B)

(C)

FIGURE 4.1: Simulation results for Example 1. Displayed in (a), (b) and (c) are boxplots of the 500 MADE values of the proposed estimator, Cai et al. (2006)'s estimator, and the ordinary local linear estimator, respectively.

Z_1 and Z_2 are independently generated from a uniform distribution $U[2, 8]$, and the errors ε and U are jointly generated from a bivariate normal distribution:

$$\begin{pmatrix} \varepsilon \\ U \end{pmatrix} \sim N\left(0, \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}\right).$$

It is clear that ε and U are independent of Z_1 and Z_2. Consequently, $\mathrm{E}(\varepsilon|Z_1, Z_2) = 0$, $\mathrm{E}(\varepsilon|U, Z_1, Z_2) = \mathrm{E}(\varepsilon|U) = 0.7U$, and $\mathrm{E}(\varepsilon) = 0$. This implies that the identification conditions of the proposed estimator and of Cai et al. (2006) are both satisfied.

For each sample size, Figure 4.1 depicts boxplots of the 500 MADE values of the proposed two-stage estimators in Figure 4.1(a), the two-stage estimator of Cai et al. (2006) in Figure

4.1(b), and the ordinary local linear estimator in Figure 4.1(c), respectively. Although it seems that the ordinary local linear estimator for $g_0(\cdot)$ is not biased in Figure 4.1(c), the MADE values of the ordinary local linear estimator of $g_1(\cdot)$ converge to a positive constant as the sample size increases. This confirms that ignoring endogeneity leads to an inconsistent result. On the contrary, we observe that the MADE values of both two-stage estimators for $g_0(\cdot)$ and $g_1(\cdot)$ do converge to zero in Figure 4.1(a) and 4.1(b). Furthermore, the MADE values of our estimator are more or less the same as values of Cai et al.'s estimator, which suggests that our estimator performs almost equally well as Cai et al.'s estimator. For comparison, Figure 4.2 displays the plots of three estimates for $g_0(\cdot)$ and $g_1(\cdot)$ from a typical sample for each sample size. The typical sample is selected such that its total MADE value equals to the median in the 500 replications.

4.5.2 Example 2: weakly dependent observations

In the second example, we generate weakly dependent data according to

$$\begin{cases} Y_t = \cos(X_{2t})X_{1t} + \varepsilon_t, \\ X_{1t} = \cos(0.5Z_{1t} + 0.6Z_{2t}) + U_{1t}, \\ X_{2t} = Z_{2t} + \sin(0.2Z_{2t}) + U_{2t}, \end{cases}$$

where the errors ε_t, U_{1t}, and U_{2t}, and the instruments Z_{1t} and Z_{2t} are generated as

$$\varepsilon_t = 0.5\omega_t + 0.3\nu_{1t}, \quad U_{1t} = 1.8\omega_t + 0.6\nu_{2t}, \quad U_{2t} = 0.5\omega_t + 0.2\nu_{3t},$$
$$Z_{1t} = 0.7Z_{1t-1} + \nu_{4t}, \quad \text{and} \quad Z_{2t} = 1 + 0.5Z_{2t-1} + \nu_{5t}.$$

Here, ω_t, ν_{1t}, ν_{2t}, ν_{3t}, ν_{4t}, and ν_{5t} are independent normal random variables with zero mean and unit variance. It is easy to see that the above design satisfies the identification conditions of Theorem 4.1: $\mathrm{E}(\varepsilon_t|U_{1t}, U_{2t}, Z_{1t}, Z_{2t}) = \mathrm{E}(\varepsilon|U_{1t}, U_{2t})$ and $\mathrm{E}(\varepsilon_t) = 0$. Since the errors U_{2t} and ε_t are correlated, the transition variable X_{2t} is endogenous, which violates the conditions in Cai et al. (2006). A study of weak instruments is provided in Section 4.9, where we investigate how the performance of our proposed estimator changes under different correlations between X_2 and Z_2.

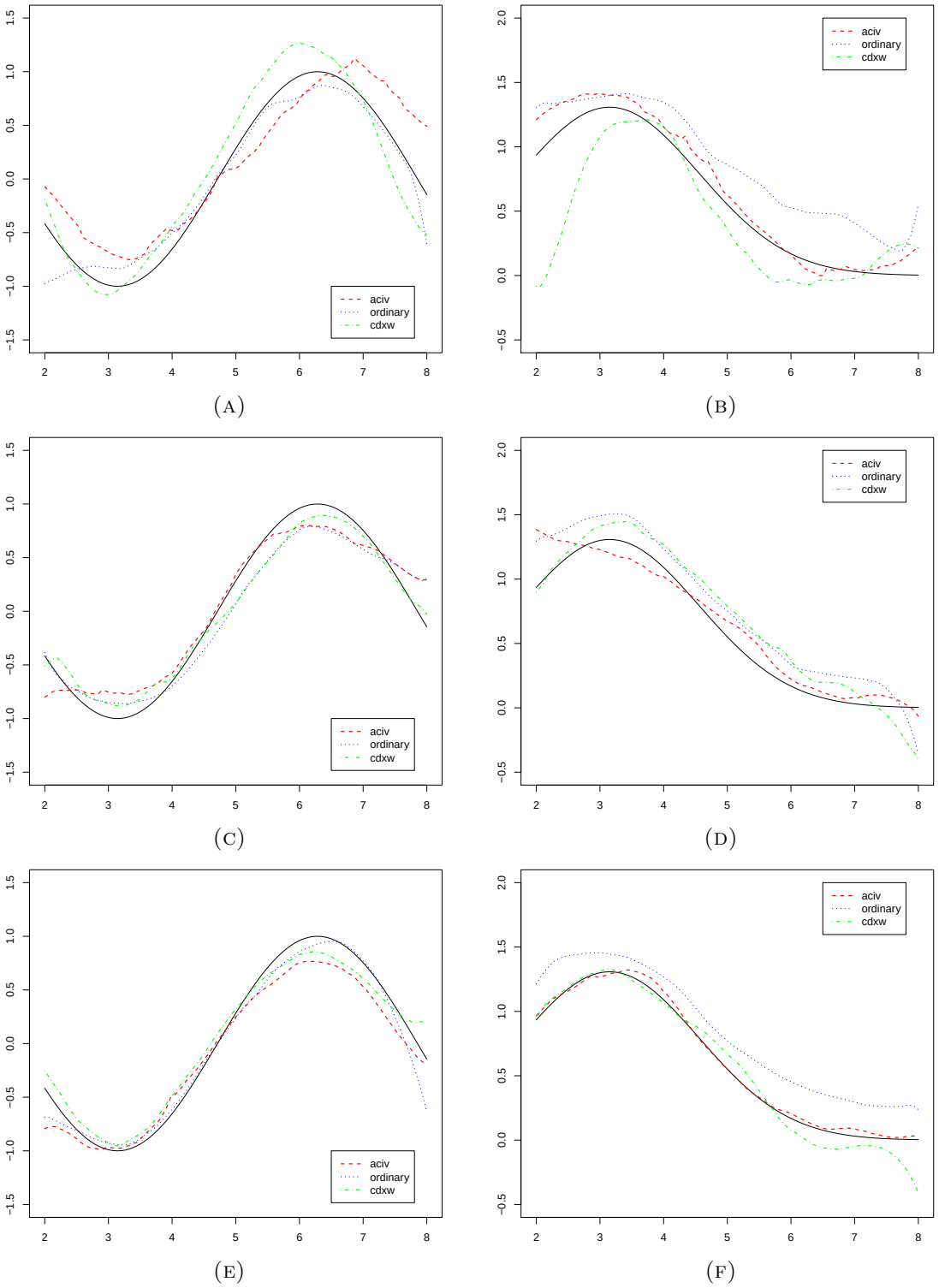


FIGURE 4.2: Simulation results for Example 1. Figures (a), (c) and (e) give the plots of the true coefficient functions $g_0(\cdot)$ (in solid line), the proposed estimator (dashed line), the estimator in Cai et al. (2006) (dashed-dotted line), and the ordinary local linear estimator (dotted line) for $n = 100$, 250 and 500, respectively. Figures (b), (d) and (f) give the plots the three estimators of the coefficient function $g_1(\cdot)$ for $n = 100$, 250 and 500, respectively.

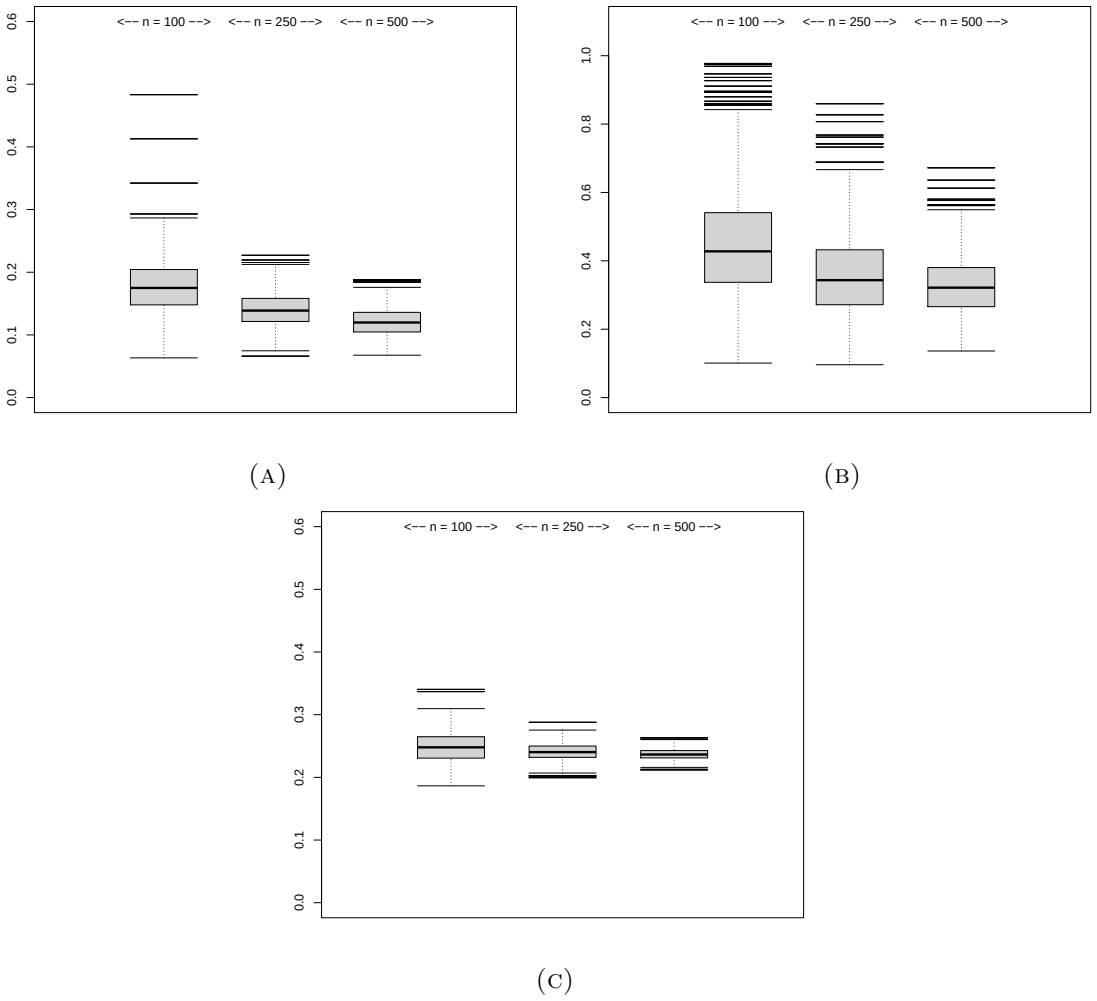


FIGURE 4.3: Simulation results for Example 2. Displayed in (a), (b) and (c) are boxplots of the 500 MADE values of the proposed estimator, Cai et al. (2006)'s estimator, and the ordinary local linear estimator, respectively.

Figure 4.3 presents boxplots of the 500 MADE values for each sample size. In Figure 4.3(a), the MADE values of the proposed two-stage estimator shrink toward zero as sample size increases. This phenomenon does not hold for the ordinary local linear estimation and two-stage estimation from Cai et al. (2006). In Figure 4.3(c), the MADE values of the ordinary local linear estimator are almost constant, and in Figure 4.3(b), the MADE values of Cai et al.'s estimator are at least twice as large as those of the proposed estimator and converge to a positive constant. The same conclusions can be drawn by using the plots of three estimates from a typical sample for each sample size in Figure 4.4. The proposed estimates closely track the true coefficient function and the biases become smaller as the sample size expands. On the other hand, the ordinary local linear estimates always overestimate the true coefficient function, whereas Cai et al.'s estimates underestimate

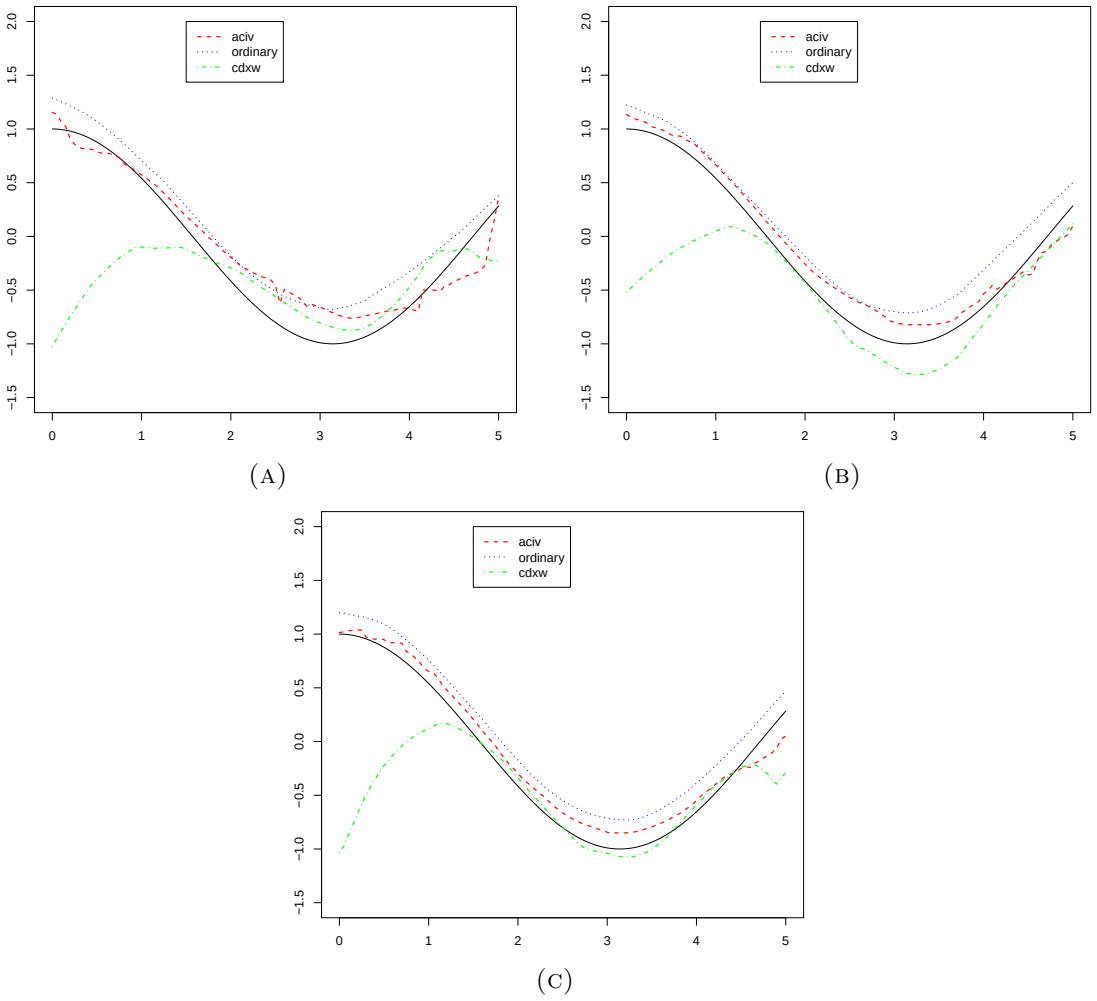
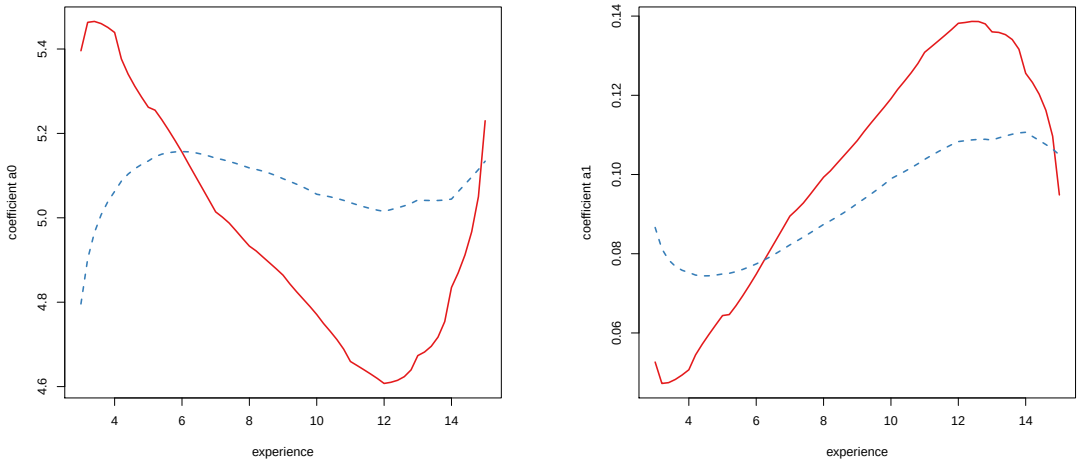


FIGURE 4.4: Simulation results for Example 2. Figures (a), (b) and (c) give the plots of the true coefficient functions (in solid line), the proposed estimator (dashed line), the estimator in Cai et al. (2006) (dashed-dotted line), and the ordinary local linear estimator (dotted line) for $n = 100$, 250 and 500, respectively.

the first one-third of the domain. These biases do not vanish as the sample size increases.

4.5.3 Example 3: real data example

In this example, we analyze the return to schooling using the wage and education data for a sample of men in 1976 from the same data set used by Card (1993). The level of education in a wage equation is endogenous if a good proxy for the individual ability is not available. To deal with the endogeneity problem of education, we considered the following IV model, which allows the marginal returns of schooling to vary with different



(A) (B)

FIGURE 4.5: Estimation results for Example 3. Figures (a) and (b) give the plots of the proposed estimator (in solid dashed line) and the ordinary local linear estimator (dashed line) for $a_0(\cdot)$ and $a_1(\cdot)$, respectively.

levels of working experience:

$$Y = a_0(X_2) + a_1(X_2)X_1 + \varepsilon,$$
$$(X_1, X_2) = \Pi(Z_1, Z_2) + U,$$

where Y is the logarithm of hourly wage, X_1 is years of schooling, and X_2 is years of potential work experience, instrumental variables Z_1 and Z_2 are measures of age and father's educational attainment, respectively. The potential work experience defined as $X_2 = Z_1 - X_1 - 6$, which is also endogenous. The objects of interests are the marginal return to work experience and schooling, $a_0(\cdot)$ and $a_1(\cdot)$.

Figures 4.5(a) and (b) show the plots of our proposed IV estimator (the solid line) and the ordinary local linear estimator without correcting endogeneity (the dashed line) of the coefficient functions $a_0(\cdot)$ and $a_1(\cdot)$. The bandwidths are set to $g = 5$ and $h_2 = \hat{\sigma}_{\hat{U}}$ for the final stage regression, and $h_1 = (\hat{\sigma}_{Z_1}, \hat{\sigma}_{Z_2})$ for the fist stage regression, where $\hat{\sigma}_X$ denotes the sample standard deviation of variable X. The plots suggest that the ordinary local linear estimators of $a_0(\cdot)$ and $a_1(\cdot)$ are positive and almost constant when the working experience is in the range of 4 to 15 years. In contrast, our proposed IV estimators suggest nonlinearities in $a_0(\cdot)$ and $a_1(\cdot)$. Figure 4.5(a) indicates that the marginal return

to working experience is positive and diminishing at the beginning and then expanding after 12 years of working experience is reached. On the other hand, Figure 4.5(b) suggests that while the marginal returns to education are positive, these returns are first increasing in experience and then decreasing after a threshold of 12 years of working experience.

4.6 Conclusion

This chapter studies a nonparametric simultaneous equations model under a functional coefficient representation for the regression function. Using the conditional mean independence restriction (4.2), the ill-posed inverse problem of nonparametric structural models is resolved, which allows the coefficients to be unknown functions of endogenous transitions variables. We propose a two-stage estimation procedure based on the local polynomial fitting and marginal integration techniques and establish its asymptotic properties. Simulation evidence suggests our estimator perform equally well as Cai et al. (2006)'s two-stage estimator in the case of an exogenous transition variable, while it exhibits reasonably good performance when the transition variable is endogenous. Future research should include optimal selection of bandwidths and allow the regressors to be a mixture of continuous and discrete variables.

4.7 Appendix: Technical lemmas

Lemma 4.5. (Sun and Chiang, 1997, Lemma 1.2) *Let* $\{\xi_i\}$ *be a* d*-dimensional strong mixing process with the mixing coefficient* $\alpha(i)$*. For any integer* $p > 1$ *and integers* $(i_1, \ldots, i_p)$ *such that* $1 \leq i_1 < i_2 < \ldots < i_p$*, let* φ *be a Borel function defined on* $\mathbb{R}^{pd}$ *such that*

$$\int |\varphi(v_1, \ldots, v_p)|^{1+\theta} \mathrm{d}F^{(1)}(v_1, \ldots, v_j) \mathrm{d}F^{(2)}(v_{j+1}, \ldots, v_p) \leq M_1$$

for some $\theta > 0$ *and* $M_1 > 0$*, where* $F^{(1)} = F_{i_1,\ldots,i_j}$ *and* $F^{(2)} = F_{i_{j+1},\ldots,i_p}$ *are the distribution functions of* $(\xi_{i_1}, \ldots, \xi_{i_j})$ *and* $(\xi_{i_{j+1}}, \ldots, \xi_{i_p})$*, respectively. Let* F *denote the distribution*

function of $(\xi_{i_1}, \ldots, \xi_{i_p})$. *Then*

$$\left| \int \varphi(v_1, \ldots, v_p) \mathrm{d}F(v_1, \ldots, v_p) - \int \varphi(v_1, \ldots, v_p) \mathrm{d}F^{(1)}(v_1, \ldots, v_j) \mathrm{d}F^{(2)}(v_{j+1}, \ldots, v_p) \right| \leq 4M_1^{1/(1+\theta)} \alpha(i_{j+1} - i_j)^{\theta/(1+\theta)}.$$

Lemma 4.6. (Gao and King, 2004, Lemma C.2(ii)) *Let* $\phi(\cdot, \cdot)$ *be a symmetric Borel function defined on* $\mathbb{R}^d \times \mathbb{R}^d$. *Let the strictly stationary process* $\{\xi_i\}$ *be defined as in Lemma 4.5. Assume that for any fixed* $v \in \mathbb{R}^d$, $\mathrm{E}[\phi(\xi_1, v)] = 0$. *Then*

$$\mathrm{E}\left\{ \sum_{1 \leq i < j \leq n} \phi(\xi_i, \xi_j) \right\}^2 \leq Cn^2 M_2^{1/(1+\theta)},$$

where $\theta > 0$ *is a fixed constant,* $C > 0$ *is a constant independent of* n *and the function* ϕ, $F(\xi_i)$ *denote the distribution function of* ξ_i, *and*

$$M_2 = \max_{1 \leq i < j \leq n} \max \left\{ \mathrm{E}|\phi(\xi_i, \xi_j)|^{2(1+\theta)}, \int \phi(\xi_i, \xi_j)^{2(1+\theta)} \mathrm{d}F(\xi_i) \mathrm{d}F(\xi_j) \right\}.$$

Lemma 4.7. (Gao and King, 2004, Lemma C.2(i)) *Let* $\psi(\cdot, \cdot, \cdot)$ *be a symmetric Borel function defined on* $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$. *Let the strictly stationary process* $\{\xi_i\}_{i=1}^n$ *be defined as in Lemma 4.5. Assume that for any fixed* $v, v^* \in \mathbb{R}^d$, $\mathrm{E}[\psi(\xi_1, v, v^*)] = 0$. *Then*

$$\mathrm{E}\left\{ \sum_{1 \leq i < j < k \leq n} \psi(\xi_i, \xi_j, \xi_k) \right\}^2 \leq Cn^3 M_3^{1/(1+\theta)},$$

where $0 < \theta < 1$ *is a small constant,* $C > 0$ *is a constant independent of* n *and the function* ψ, $M_3 = \max\{M_{31}, M_{32}, M_{33}\}$, *and*

$$M_{31} = \max_{1 \leq i < j \leq n} \max \left\{ \mathrm{E}|\psi(\xi_1, \xi_i, \xi_j)|^{2(1+\theta)}, \int \psi(\xi_1, \xi_i, \xi_j)^{2(1+\theta)} \mathrm{d}F(\xi_1) \mathrm{d}F(\xi_i) \mathrm{d}F(\xi_j) \right\},$$
$$M_{32} = \max_{1 \leq i < j \leq n} \int \psi(\xi_1, \xi_i, \xi_j)^{2(1+\theta)} \mathrm{d}F(\xi_j) \mathrm{d}F(\xi_1, \xi_i),$$
$$M_{33} = \max_{1 \leq i < j \leq n} \int \psi(\xi_1, \xi_i, \xi_j)^{2(1+\theta)} \mathrm{d}F(\xi_1) \mathrm{d}F(\xi_i, \xi_j).$$

4.8 Appendix: Proof of the theorems

Proof of Theorem 4.1

Let $\underline{x}_1 = (x_{11}, \ldots, x_{1d_1})^\top$ and $\underline{a}(x_2) = [a_1(x_2), \ldots, a_{d_1}(x_2)]^\top$. Suppose that there exists another set of parameters $\dot{a}(x_2) = [\dot{a}_0(x_2), \dot{\underline{a}}^\top(x_2)]^\top$, $\dot{\gamma}$ and $\dot{\lambda}(u)$ satisfying equation (4.3) for every x_2 and u on their compact supports. Defining $\delta_1(x_2) = a_0(x_2) - \dot{a}_0(x_2)$, $\delta_2(x_2) = \underline{a}(x_2) - \dot{\underline{a}}(x_2)$, $\delta_3 = \gamma - \dot{\gamma}$, and $\delta_4(u) = \lambda(u) - \dot{\lambda}(u)$, we have on the whole support

$$\begin{aligned} 0 &= \delta_1(x_2) + \underline{x}_1^\top \delta_2(x_2) + z_1^\top \delta_3 + \delta_4(u) \\ &= \delta_1(\Pi_2(z) + u_2) + \{\Pi_1(z) + u_1\}^\top \delta_2(\Pi_2(z) + u_2) + z_1^\top \delta_3 + \delta_4(u), \end{aligned} \tag{4.16}$$

where $[\Pi_1^\top(z), \Pi_2^\top(z)]^\top = \Pi(z)$ denotes the parts of $\Pi(z)$ corresponding to X_1 and X_2, respectively, and $(u_1^\top, u_2)^\top = u$, by using the fact that $(\underline{x}_1^\top, x_2)^\top = \Pi(z) + u$.

Let D, D_1, and D_2 denote the partial derivatives with respect to z, z_1, and z_2, respectively; for example, $D\Pi(z)$ is the Jacobian matrix of $\Pi(z)$. By the continuous differentiability assumptions for $\Pi(\cdot)$, $a(\cdot)$, and $\lambda(\cdot)$, we differentiate the identity (4.16) with respect to z_1, z_2, and u, respectively:

$$0 = D_1\Pi^\top(z) \begin{pmatrix} \delta_2(x_2) \\ D\delta_1(x_2) + D\delta_2^\top(x_2)\underline{x}_1 \end{pmatrix} + \delta_3, \tag{4.17}$$

$$0 = D_2\Pi^\top(z) \begin{pmatrix} \delta_2(x_2) \\ D\delta_1(x_2) + D\delta_2^\top(x_2)\underline{x}_1 \end{pmatrix}, \tag{4.18}$$

$$0 = \begin{pmatrix} \delta_2(x_2) \\ D\delta_1(x_2) + D\delta_2^\top(x_2)\underline{x}_1 \end{pmatrix} + D\delta_4(u). \tag{4.19}$$

By the full rank condition for $D_2\Pi^\top(z_2)$ given in the statement of the theorem, it follows from equation (4.18) that $\delta_2(x_2) = 0$, and thereby $D\delta_2(x_2) = 0$ and $D\delta_1(x_2) = 0$. By equations (4.17) and (4.19) and full rank $D\Pi_+^\top$, we have $\delta_3 = 0$ and $D\delta_4(u) = 0$. As $\delta_2(x_2) = 0$, $\delta_3 = 0$, and the partial derivatives of $\delta_1(x_2)$ and $\delta_4(u)$ are zero, it follows from (4.16) that $\delta_1(x_2) = c$ and $\delta_4(u) = -c$ for some constant c. If we further assume

$\mathrm{E}(\varepsilon) = \mathrm{E}\lambda(U) = 0$, we have $\mathrm{E}\delta_4(u) = 0$ and hence $c = 0$. This completes the proof of Theorem 4.1. □

Proof of Theorem 4.2

By the continuous differentiability condition for the coefficient function $a_l(\cdot)$ (Assumption 4.A2) and the (p_2+1)th order Taylor expansion for X_{2j} in a neighborhood of x_2 such that $|X_{2j} - x_2| \le g$, we have

$$\begin{aligned} Y_j =& \mathcal{X}_j^\top \beta(x_2) + \frac{g^{p_2+1}}{(p_2+1)!}\left(\frac{X_{2j}-x_2}{g}\right)^{p_2+1} X_{1j}^\top a^{(p_2+1)}(x_2) \\ &+ \lambda(U_j) + \epsilon_j + \mathrm{o}(g^{p_2+1}), \end{aligned} \tag{4.20}$$

in which $\mathcal{X}_j$ is defined in (4.7), $\beta(x_2) = [a^\top(x_2), ga'^\top(x_2), \ldots, (p_2!)^{-1}g^{p_2}a^{(p_2)\top}(x_2), \gamma^\top]^\top$, $a^{(q)}(x_2)$ is a vector consisting of the qth-order derivatives of $a_l(x_2)$, and the residual $\epsilon_j = \varepsilon_j - \lambda(U_j)$. Substituting equation (4.20) into the definition of $\tilde{a}_l(x_2)$ in (4.8) gives the following decomposition:

$$\begin{aligned} \tilde{a}_l(x_2) =& \frac{1}{n}\sum_{i=1}^n e_{l+1}^\top \tilde{S}_n^{-1}(x_2,U_i)\tilde{T}_n(x_2,U_i) \\ =& \frac{1}{n}\sum_{i=1}^n e_{l+1}^\top \tilde{S}_n^{-1}(x_2,U_i)\frac{1}{ngh_2^d}\sum_{j=1}^n L\left(\frac{X_{2j}-x_2}{g}\right)K_2\left(\frac{U_j-U_i}{h_2}\right)\mathcal{X}_j Y_j \\ =& \frac{1}{n}\sum_{i=1}^n e_{l+1}^\top \tilde{S}_n^{-1}(x_2,U_i)\frac{1}{ngh_2^d}\sum_{j=1}^n L\left(\frac{X_{2j}-x_2}{g}\right)K_2\left(\frac{U_j-U_i}{h_2}\right)\mathcal{X}_j\{\epsilon_j \\ &+ \frac{g^{p_2+1}}{(p_2+1)!}\left(\frac{X_{2j}-x_2}{g}\right)^{p_2+1} X_{1j}^\top a^{(p_2+1)}(x_2) + [\lambda(U_j)-\lambda(U_i)] \\ &+ \underbrace{\mathcal{X}_j^\top e_1}_{=1}\lambda(U_i) + \mathcal{X}_j^\top\beta(x_2) + \underbrace{\mathcal{X}_j^\top e_1}_{=1}\mathrm{o}(g^{p_2+1})\} \\ =& \frac{1}{n}\sum_{i=1}^n e_{l+1}^\top \tilde{S}_n^{-1}(x_2,U_i)[\tilde{T}_{n,1}(x_2,U_i) + \tilde{T}_{n,2}(x_2,U_i) + \tilde{T}_{n,3}(x_2,U_i)] \\ &+ e_{l+1}^\top e_1 R_n + e_{l+1}^\top\beta(x_2) + e_{l+1}^\top e_1 \mathrm{o_p}(g^{p_2+1}) \\ =& \frac{1}{n}\sum_{i=1}^n e_{l+1}^\top \tilde{S}_n^{-1}(x_2,U_i)\bar{\tilde{T}}_n(x_2,U_i) + e_{l+1}^\top e_1 R_n + a_l(x_2) + \mathrm{o_p}(g^{p_2+1}), \end{aligned} \tag{4.21}$$

where $R_n = \frac{1}{n}\sum_{i=1}^n \lambda(U_i)$, $\bar{\tilde{T}}_n(x_2, U_i) = \tilde{T}_{n,1}(x_2, U_i) + \tilde{T}_{n,2}(x_2, U_i) + \tilde{T}_{n,3}(x_2, U_i)$,

$$\tilde{T}_{n,1}(x_2, U_i) = \frac{1}{ngh_2^d}\sum_{j=1}^n L\left(\frac{X_{2j}-x_2}{g}\right) K_2\left(\frac{U_j - U_i}{h_2}\right)\mathcal{X}_j\epsilon_j, \tag{4.22}$$

$$\begin{aligned}\tilde{T}_{n,2}(x_2, U_i) = &\frac{g^{p_2+1}}{(p_2+1)!ngh_2^d}\sum_{j=1}^n L\left(\frac{X_{2j}-x_2}{g}\right) K_2\left(\frac{U_j - U_i}{h_2}\right)\mathcal{X}_j \\ &\times \left(\frac{X_{2j}-x_2}{g}\right)^{p_2+1} X_{1j}^\top a^{(p_2+1)}(x_2)\end{aligned} \tag{4.23}$$

and

$$\tilde{T}_{n,3}(x_2, U_i) = \frac{1}{ngh_2^d}\sum_{j=1}^n L\left(\frac{X_{2j}-x_2}{g}\right) K_2\left(\frac{U_j - U_i}{h_2}\right)\mathcal{X}_j[\lambda(U_j) - \lambda(U_i)].$$

To facilitate the notation, we define an operator

$$\Phi(A, B) = \frac{1}{n}\sum_{i=1}^n e_{l+1}^\top AB$$

and a covariance matrix – the population counterpart of $\tilde{S}_n(x_2, u)$ –

$$S(x_2, u) = \begin{pmatrix} S_{XX}(x_2, u) & S_{XZ}(x_2, u) \\ S_{ZX}(x_2, u) & S_{ZZ}(x_2, u)\end{pmatrix},$$

where $\rho_2(v) = (1, v, \ldots, v^{p_2})^\top$,

$$\begin{aligned}S_{XX}(x_2, u) &= \left\{\int \rho_2(v)\rho_2^\top(v)L(v)\mathrm{d}v\right\} \otimes \mathrm{E}[X_1X_1^\top | X_2 = x_2, U = u], \\ S_{XZ}(x_2, u) = S_{ZX}^\top(x_2, u) &= \left\{\int \rho_2(v)L(v)\mathrm{d}v\right\} \otimes \mathrm{E}[X_1Z_1^\top | X_2 = x_2, U = u],\end{aligned}$$

and

$$S_{ZZ}(x_2, u) = \mathrm{E}[Z_1Z_1^\top | X_2 = x_2, U = u].$$

Since $A_1^{-1} - A_2^{-1} = A_1^{-1}(A_2 - A_1)A_2^{-1}$ for nonsingular matrices A_1 and A_2, we have

$$\tilde{S}_n^{-1}(x_2, u) = \frac{S^{-1}(x_2, u)}{f_{X_2U}(x_2, u)} - \tilde{Q}_n(x_2, u), \tag{4.24}$$

where

$$\tilde{Q}_n(x_2, u) = \tilde{S}_n^{-1}(x_2, u)\left[\tilde{S}_n(x_2, u) - f_{X_2U}(x_2, u)S(x_2, u)\right]\frac{S^{-1}(x_2, u)}{f_{X_2U}(x_2, u)}.$$

Here we require $\tilde{S}_n^{-1}(x_2, u)$ to be invertible, which is shown in Lemma 4.12 for a sufficiently large n under the non-zero marginal density $f_{X_2U}(x_2, u)$ condition in Assumption 4.A4 and the rank condition for $S(x_2, u)$ in Assumption 4.A7. It follows from equations (4.21) and (4.24) that

$$\begin{aligned}\tilde{a}_l(x_2) - a_l(x_2) =& \frac{1}{n}\sum_{i=1}^{n} e_{l+1}^{\top}\left\{\frac{S^{-1}(x_2, U_i)}{f_{X_2U}(x_2, U_i)} - \tilde{Q}_n(x_2, U_i)\right\}\tilde{\bar{T}}_n(x_2, U_i) + e_{l+1}^{\top}e_1 R_n \\ &+ \mathrm{o}(g^{p_2+1}) \\ =& \sum_{c=1}^{3}\{P_c(x_2) - R_c(x_2)\} + e_{l+1}^{\top}e_1 R_n + \mathrm{o}(g^{p_2+1}),\end{aligned}$$

where for $c = 1, 2, 3$,

$$P_c(x_2) = \Phi\left(\frac{S^{-1}(x_2, U_i)}{f_{X_2U}(x_2, U_i)}, \tilde{T}_{n,c}(x_2, U_i)\right) \quad \text{and} \quad R_c(x_2) = \Phi(\tilde{Q}_n(x_2, U_i), \tilde{T}_{n,c}(x_2, U_i)).$$

We complete the proof of Theorem 4.2 by investigating the asymptotic properties of the terms R_n, $P_c(x_2)$ and $R_c(x_2)$ for $c = 1, 2, 3$, in Lemmas 4.8–4.12. □

Lemma 4.8. *Suppose Assumptions 4.A4, 4.A7, and 4.A8 hold. We have for $n \to \infty$ and $v \in [-1, 1]^d$ that*

$$S^{-1}(x_2, u + h_2 v) = \frac{f_{X_2U}(x_2, u + h_2 v)}{f_{X_2U}(x_2, u)} S^{-1}(x_2, u)(1 + \mathrm{o}(1)).$$

Proof. By the Taylor expansion of the joint density f, one has

$$\begin{aligned}
S_{XZ}(x_2, u + h_2 v) =& \int \rho_2(\dot{v}) L(\dot{v}) \mathrm{d}\dot{v} \otimes \mathrm{E}[X_1 Z_1^\top | X_2 = x_2, U = u + h_2 v] \\
=& \int \rho_2(\dot{v}) L(\dot{v}) \mathrm{d}\dot{v} \otimes \int X_1 Z_1^\top \frac{f(X_1, x_2, Z_1, u + h_2 v)}{f_{X_2U}(x_2, u + h_2 v)} \mathrm{d}X_1 \mathrm{d}Z_1 \\
=& \frac{f_{X_2U}(x_2, u)}{f_{X_2U}(x_2, u + h_2 v)} \int \rho_2(\dot{v}) L(\dot{v}) \mathrm{d}\dot{v} \\
& \otimes \int X_1 Z_1^\top \frac{f(X_1, x_2, Z_1, u)}{f_{X_2U}(x_2, u)} \mathrm{d}X_1 \mathrm{d}Z_1 + \mathrm{O}(h_2 v) \\
=& \frac{f_{X_2U}(x_2, u)}{f_{X_2U}(x_2, u + h_2 v)} S_{XZ}(x_2, u) + \mathrm{O}(h_2 v),
\end{aligned}$$

where the third equality follows from the (lower) boundedness conditions for $f_{X_2U}(x_2, u)$, the partial derivatives of the joint density f with respect to U, and $\mathrm{E}[X_1 Z_1^\top | X_2 = x_2, U = u]$ in Assumptions 4.A4 and 4.A7. Similarly, one can also show that

$$S_{XX}(x_2, u + h_2 v) = \frac{f_{X_2U}(x_2, u)}{f_{X_2U}(x_2, u + h_2 v)} S_{XX}(x_2, u) + \mathrm{O}(h_2 v)$$

and

$$S_{ZZ}(x_2, u + h_2 v) = \frac{f_{X_2U}(x_2, U_j)}{f_{X_2U}(x_2, u + h_2 v)} S_{ZZ}(x_2, u) + \mathrm{O}(h_2 v).$$

Consequently, by the invertiblity of $f_{X_2U}(x_2, u)$ and $S(x_2, u)$ in Assumptions 4.A4 and 4.A7, and the fact that $h_2 = h_2(n) \to 0$ as $n \to \infty$ in 4.A8, we have for $v \in [-1, 1]^d$,

$$S^{-1}(x_2, U_j + h_2 v) = \frac{f_{X_2U}(x_2, U_j + h_2 v)}{f_{X_2U}(x_2, U_j)} S^{-1}(x_2, U_j)(1 + \mathrm{o}(1)).$$

□

Lemma 4.9. *Under Assumptions 4.A2, 4.A3, and 4.A8, we have as $n \to \infty$,*

$$\sqrt{ng} R_n = \mathrm{O_p}(g^{1/2}) = \mathrm{o_p}(1).$$

Proof. Applying the central limit theorem for strongly mixing process (Fan and Yao, 2003, Theorem 2.21) under the identification condition $\mathrm{E}\lambda(U) = 0$, the mixing condition for U

in Assumption 4.A3, and the boundedness assumption on $\lambda(\cdot)$ in 4.A2, we have

$$\sqrt{ng}\frac{1}{n}\sum_{i=1}^{n}\lambda(U_i) = \sqrt{ng}\mathrm{O_p}(n^{-1/2}) = \mathrm{O_p}(g^{1/2}) = \mathrm{o_p}(1),$$

where the last equality follows from the fact that $g = g(n) \to 0$ as $n \to +\infty$ (Assumption 4.A8). □

Lemma 4.10. *Under Assumptions 4.A1–4.A8, for $n \to \infty$, $l = 0, \ldots, d_1$, and any fixed point $x_2 \in D_{X_2}$,*

(i) $\sqrt{ng}P_1(x_2) \xrightarrow{d} N\{0, \sigma_l^2(x_2)\}$,

(ii) $P_2(x_2) = g^{p_2+1}\eta_l(x_2) + \mathrm{o_p}(g^{p_2+1})$, *and*

(iii) $P_3(x_2) = \mathrm{O_p}(h_2^{q_2}) = \mathrm{o_p}(n^{-1/2}g^{-1/2})$,

where

$$\sigma_l^2(x_2) = f_{X_2}(x_2)\mathrm{E}\left[\frac{f_U^2(U)}{f_{X_2U}^2(x_2,U)}\sigma_\epsilon^2(X_1,x_2,Z_1,U)\int \mathcal{L}_l^2\left(v,x_2,U,X_1,Z_1\right)\mathrm{d}v\right]$$

and

$$\eta_l(x_2) = \int v^{p_2+1}\mathrm{E}[\mathcal{L}_l(v,x_2,U,X_1,Z_1)X_1^\top]\mathrm{d}v\frac{a^{(p_2+1)}(x_2)}{(p_2+1)!}$$

with

$$\sigma_\epsilon^2(x_1,x_2,z_1,u) = \mathrm{E}[\epsilon^2|X_1 = x_1, X_2 = x_2, Z_1 = z_1, U = u]$$

and

$$\mathcal{L}_l(v,x_2,U,X_1,Z_1) = e_{l+1}^\top S^{-1}(x_2,U)\begin{pmatrix}\rho_2(v)\otimes X_1\\ Z_1\end{pmatrix}L(v).$$

Proof. To show part (i), let $\xi_i = (X_{1i}^\top, X_{2i}, Z_{1i}^\top, U_i^\top, \epsilon_j)^\top$ and

$$\phi_0(\xi_i, \xi_j) = \frac{1}{f_{X_2U}(x_2, U_i)} \mathcal{L}_l \left(\frac{X_{2j} - x_2}{g}, x_2, U_i, X_{1j}, Z_{1j} \right) K_2 \left(\frac{U_j - U_i}{h_2} \right) \epsilon_j.$$

Define $\bar{\phi}_0(\xi_i, \xi_j) = \phi_0(\xi_i, \xi_j) - \mathrm{E}_i[\phi_0(\xi_i, \xi_j)]$ and $\bar{\phi}(\xi_i, \xi_j) = \{\bar{\phi}_0(\xi_i, \xi_j) + \bar{\phi}_0(\xi_j, \xi_i)\}/2$, where E_i denotes the expectation with respect to ξ_i. By construction, $\bar{\phi}(\xi_i, \xi_j)$ is symmetric and $\mathrm{E}_i[\bar{\phi}_0(\xi_i, \xi_j)] = \mathrm{E}_i[\bar{\phi}(\xi_i, \xi_j)] = 0$.

Using the definitions of $\tilde{T}_{n,1}(x_2, U_i)$ and $\mathcal{X}_j$ in (4.22) and (4.7), respectively, we have

$$\begin{aligned}
P_1(x_2) =& \Phi \left(\frac{S^{-1}(x_2, U_i)}{f_{X_2U}(x_2, U_i)}, \tilde{T}_{n,1}(x_2, U_i) \right) \\
=& \frac{1}{n^2 g h_2^d} \sum_{i=1}^n \sum_{j=1}^n \frac{e_{l+1}^\top S^{-1}(x_2, U_i)}{f_{X_2U}(x_2, U_i)} L \left(\frac{X_{2j} - x_2}{g} \right) \begin{pmatrix} \rho_2 \left(\frac{X_{2j} - x_2}{g} \right) \otimes X_{1j} \\ Z_{1j} \end{pmatrix} \\
& \times K_2 \left(\frac{U_j - U_i}{h_2} \right) \epsilon_j \\
=& \frac{1}{n^2 g h_2^d} \sum_{i=1}^n \sum_{j=1}^n \frac{1}{f_{X_2U}(x_2, U_i)} \mathcal{L}_l \left(\frac{X_{2j} - x_2}{g}, x_2, U_i, X_{1j}, Z_{1j} \right) K_2 \left(\frac{U_j - U_i}{h_2} \right) \epsilon_j
\end{aligned}$$

With the help of the notations ϕ_0, $\bar{\phi}_0$, and $\bar{\phi}$,

$$\begin{aligned}
P_1(x_2) =& \frac{1}{n^2 g h_2^d} \sum_{i=1}^n \sum_{j=1}^n \phi_0(\xi_i, \xi_j) \\
=& \frac{1}{n^2 g h_2^d} \sum_{i=1}^n \sum_{j=1}^n \{\mathrm{E}_i[\phi_0(\xi_i, \xi_j)] + \bar{\phi}_0(\xi_i, \xi_j)\} \\
=& \frac{1}{n^2 g h_2^d} \sum_{i=1}^n \sum_{j=1}^n \mathrm{E}_i[\phi_0(\xi_i, \xi_j)] + \frac{1}{n^2 g h_2^d} \sum_{i=1}^n \bar{\phi}_0(\xi_i, \xi_i) + \frac{2}{n^2 g h_2^d} \sum_{1 \le i < j \le n} \bar{\phi}(\xi_i, \xi_j) \\
=& P_{1a} + P_{1b} + P_{1c}, \qquad (4.25)
\end{aligned}$$

where P_{1a}, P_{1b}, and P_{1c} denote the first, second, and last terms in the second last equality, respectively.

For the term P_{1a}, by a change of variables ($\dot{u} = U_j + h_2 v$), we have

$$\begin{aligned}
P_{1a} =& \frac{1}{n^2 g h_2^d} \sum_{i=1}^{n} \sum_{j=1}^{n} \int \frac{f_U(\dot{u})}{f_{X_2U}(x_2, \dot{u})} \mathcal{L}_l \left(\frac{X_{2j} - x_2}{g}, x_2, \dot{u}, X_{1j}, Z_{1j} \right) K_2 \left(\frac{U_j - \dot{u}}{h_2} \right) \epsilon_j \mathrm{d}\dot{u} \\
=& \frac{1}{ngh_2^d} \sum_{j=1}^{n} \int \frac{f_U(\dot{u})}{f_{X_2U}(x_2, \dot{u})} \mathcal{L}_l \left(\frac{X_{2j} - x_2}{g}, x_2, \dot{u}, X_{1j}, Z_{1j} \right) K_2 \left(\frac{U_j - \dot{u}}{h_2} \right) \epsilon_j \mathrm{d}\dot{u} \\
=& \frac{1}{ng} \sum_{j=1}^{n} \int \frac{f_U(U_j + h_2 v)}{f_{X_2U}(x_2, U_j + h_2 v)} \mathcal{L}_l \left(\frac{X_{2j} - x_2}{g}, x_2, U_j + h_2 v, X_{1j}, Z_{1j} \right) K_2(v) \epsilon_j \mathrm{d}v.
\end{aligned}$$

According to Lemma 4.8 and its proof, the Taylor expansion of $S^{-1}(x_2, U_j + h_2)$, the q_2th order Taylor expansion of the marginal density f_U, and the bounded continuous differentiability condition for f_U (Assumption 4.A4),

$$\begin{aligned}
P_{1a} =& \frac{1}{ng} \sum_{j=1}^{n} \int \frac{f_U(U_j + v h_2)}{f_{X_2U}(x_2, U_j)} \mathcal{L}_l \left(\frac{X_{2j} - x_2}{g}, x_2, U_j, X_{1j}, Z_{1j} \right) (1 + \mathrm{o}(1)) \epsilon_j K_2(v) \mathrm{d}v \\
=& \frac{1}{ng} \sum_{j=1}^{n} \frac{f_U(U_j)}{f_{X_2U}(x_2, U_j)} \mathcal{L}_l \left(\frac{X_{2j} - x_2}{g}, x_2, U_j, X_{1j}, Z_{1j} \right) \epsilon_j + \mathrm{O_p}(h_2^{q_2}) \\
=& \frac{1}{ng} \sum_{j=1}^{n} \frac{f_U(U_j)}{f_{X_2U}(x_2, U_j)} \mathcal{L}_l \left(\frac{X_{2j} - x_2}{g}, x_2, U_j, X_{1j}, Z_{1j} \right) \epsilon_j + \mathrm{o_p}(n^{-1/2} g^{-1/2}).
\end{aligned}$$

The $\mathrm{O_p}$-term in the second equality follows from that the convergence result for the term $\frac{1}{ng} \sum_j \mathcal{X}_j \epsilon_j L \left(\frac{X_{2j} - x_2}{g} \right) = \mathrm{O_p}(1)$ by Theorem 1 in Hansen (2008) (under Assumptions 4.A1, 4.A3, 4.A4, and 4.A6), the property of the q_2th order kernel K_2 in 4.A1, the boundedness conditions for f_U in 4.A4, and the existence of the inverse of f_{X_2U} and $S(x_2, u)$ in 4.A4 and 4.A7. The last equality is due to Assumption 4.A8(i) on convergence rate of h_2.

By applying the central limit theorem for strong mixing process (Fan and Yao, 2003, Theorem 2.21) under Assumptions 4.A3–4.A7 and the first order Taylor expansion of density f and conditional variance σ_ϵ^2 with their differentiability conditions in 4.A4–4.A5, $\sqrt{ng} P_{1a}$ is asymptotically normal with mean 0 (due to the law of iterated expectation)

and variance

$$\begin{aligned}
&\frac{1}{g}\int \frac{f_U^2(\dot{u})}{f_{X_2U}^2(x_2,\dot{u})}\mathcal{L}_l^2\left(\frac{\dot{x}_2-x_2}{g},x_2,\dot{u},\dot{x}_1,\dot{z}_1\right)\sigma_\epsilon^2(\dot{x},\dot{z}_1,\dot{u})f(\dot{x},\dot{z}_1,\dot{u})\mathrm{d}\dot{x}\mathrm{d}\dot{z}_1\mathrm{d}\dot{u}\\
&\quad= f_{X_2}(x_2)\int \frac{f_U^2(\dot{u})}{f_{X_2U}^2(x_2,\dot{u})}\mathcal{L}_l^2\left(v,x_2,\dot{u},\dot{x}_1,\dot{z}_1\right)\sigma_\epsilon^2(\dot{x}_1,x_2+gv,\dot{z}_1,\dot{u})\\
&\qquad\times\frac{f(\dot{x}_1,x_2+gv,\dot{z}_1,\dot{u})}{f_{X_2}(x_2)}\mathrm{d}\dot{x}_1\mathrm{d}v\mathrm{d}\dot{z}_1\mathrm{d}\dot{u}\\
&\quad= f_{X_2}(x_2)\int \frac{f_U^2(\dot{u})}{f_{X_2U}^2(x_2,\dot{u})}\sigma_\epsilon^2(\dot{x}_1,x_2,\dot{z}_1,\dot{u})\int\mathcal{L}_l^2\left(v,x_2,\dot{u},\dot{x}_1,\dot{z}_1\right)\mathrm{d}v\\
&\qquad\times f(\dot{x}_1,\dot{z}_1,\dot{u}|x_2)\mathrm{d}\dot{x}_1\mathrm{d}\dot{z}_1\mathrm{d}\dot{u}+\mathrm{O}(g)\\
&\quad= f_{X_2}(x_2)\mathrm{E}\left[\frac{f_U^2(U)}{f_{X_2U}^2(x_2,U)}\sigma_\epsilon^2(X_1,x_2,Z_1,U)\int\mathcal{L}_l^2\left(v,x_2,U,X_1,Z_1\right)\mathrm{d}v\right]+\mathrm{O}(g)\\
&\quad=\sigma_l^2(x_2)+\mathrm{o}(1).
\end{aligned}$$

The corresponding covariances across observations are not present since their sum can be shown to be negligible in probability similar to Lemma 2(ii) in Čížek and Koo (2017a) under Assumptions 4.A3–4.A7 (see also Cai et al., 2000, Lemma A.1(b)).

For the term P_{1b}, according to the law of iterated expectation, we obtain $\mathrm{E}P_{1b}=0$. By the mixing condition in Assumption 4.A3 and Lemma 4.5 (with $\theta=\delta_2/2$),

$$\begin{aligned}
\mathrm{var}(P_{1b}) &= \frac{1}{n^4g^2h_2^{2d}}\sum_{i=1}^n\sum_{j=1}^n\mathrm{E}[\bar{\phi}_0(\xi_i,\xi_i)\bar{\phi}_0(\xi_j,\xi_j)]\\
&= \frac{1}{n^3g^2h_2^{2d}}\mathrm{E}[\bar{\phi}_0(\xi_i,\xi_i)^2]+\frac{1}{n^4g^2h_2^{2d}}\sum_{\tau=1}^{n-1}\sum_{i=1}^n\mathrm{E}[\bar{\phi}_0(\xi_i,\xi_i)\bar{\phi}_0(\xi_{i-\tau},\xi_{i-\tau})]\\
&\le Cn^{-3}g^{-2}h_2^{-2d}g+Cn^{-3}g^{-2}h_2^{-2d}g^{4/(2+\delta_2)}\sum_{\tau=1}^{n-1}(1-\tau/n)\alpha^{\delta_2/(2+\delta_2)}(\tau)\\
&= \mathrm{O}(n^{-3}g^{-1}h_2^{-2d}+n^{-3}g^{-2\delta_2/(2+\delta_2)}h_2^{-2d})\\
&= \mathrm{o}(n^{-1}g^{-1}). \qquad \text{(by Assumption 4.A8(ii))}
\end{aligned}$$

As a result, $P_{1b}=\mathrm{o_p}(n^{-1/2}g^{-1/2})$ by Chebyshev's inequality. In order to use Lemma 4.5 above, we require $\mathrm{E}|\bar{\phi}_0(\xi_i,\xi_i)|^{2+\delta_2}<\infty$, which is ensured by the compact supports of the kernels L and K_2 in Assumption 4.A1, which implies $\left\|\rho_2\left\{(X_{2j}-x_2)/g\right\}\right\|_\infty\le 1$, the existence of $S^{-1}(x_2,u)$ due to Assumption 4.A7, the boundedness of the variables X_1 and

Z_1 and the densities f_{X_2U} in Assumption 4.A4, and the moment conditions for $X_1\epsilon$ and $Z_1\epsilon$ in Assumption 4.A6.

For the term P_{1c}, by Lemma 4.5 (with $\theta = \delta_2/2$) under Assumptions 4.A3–4.A7 (such that the conditions in Lemma 4.5 are satisfied),

$$\begin{aligned}
|\mathrm{E}P_{1c}| &= \left| \frac{2}{n^2 g h_2^d} \sum_{\tau=1}^{n-1} \sum_{i=1}^{n} \mathrm{E}\bar{\phi}(\xi_i, \xi_{i-\tau}) \right| \\
&\leq Cn^{-1}g^{-1}h_2^{-d}(gh_2^d)^{2/(2+\delta_2)} \sum_{\tau=1}^{n-1}(1-\tau/n)\alpha^{\delta_2/(2+\delta_2)}(\tau) \\
&= \mathrm{O}(n^{-1}g^{-\delta_2/(2+\delta_2)}h_2^{-d\delta_2/(2+\delta_2)}) \\
&= \mathrm{o}(n^{-1/2}g^{-1/2}). \qquad \text{(by Assumption 4.A8(ii))}
\end{aligned}$$

It follows from Lemma 4.6 (with $\theta = \delta_2/2$) under Assumptions 4.A3–4.A7 that

$$\begin{aligned}
\mathrm{E}(P_{1c}^2) &\leq Cn^{-2}g^{-2}h_2^{-2d}(gh_2^d)^{2/(2+\delta_2)} \\
&= \mathrm{O}(n^{-2}g^{-2(1+\delta_2)/(2+\delta_2)}h_2^{-2d(1+\delta_2)/(2+\delta_2)}) \\
&= \mathrm{o}(n^{-1}g^{-1}). \qquad \text{(by Assumption 4.A8(ii))}
\end{aligned}$$

Hence, $P_{1c} = \mathrm{o_p}(n^{-1/2}g^{-1/2})$ by Chebyshev's inequality. This concludes part (i).

To prove part (ii), similar to the decomposition of $P_1(x_2)$ in (4.25), we split $P_2(x_2)$ into three terms – P_{2a}, P_{2b}, and P_{2c}, where the leading term P_{2a} is

$$\begin{aligned}
P_{2a} =& \frac{g^{p_2+1}}{(p_2+1)!ngh_2^d} \sum_{j=1}^{n} \int \frac{f_U(\dot{u})}{f_{X_2U}(x_2,\dot{u})} \mathcal{L}_l\left(\frac{X_{2j}-x_2}{g}, x_2, \dot{u}, X_{1j}, Z_{1j}\right) K_2\left(\frac{U_j-\dot{u}}{h_2}\right) \\
&\times \left(\frac{X_{2j}-x_2}{g}\right)^{p_2+1} X_{1j}^{\top} \mathrm{d}\dot{u} a^{(p_2+1)}(x_2) \\
=& \frac{g^{p_2+1}}{(p_2+1)!gh_2^d} \int\int \frac{f(\mathring{x},\mathring{z}_1,\mathring{u})f_U(\dot{u})}{f_{X_2U}(x_2,\dot{u})} \mathcal{L}_l\left(\frac{\mathring{x}_2-x_2}{g}, x_2, \dot{u}, \mathring{x}_1, \mathring{z}_1\right) K_2\left(\frac{\mathring{u}-\dot{u}}{h_2}\right) \\
&\times \left(\frac{\mathring{x}_2-x_2}{g}\right)^{p_2+1} \mathring{x}_1^{\top} \mathrm{d}\mathring{x}\mathrm{d}\mathring{z}_1\mathrm{d}\mathring{u}\mathrm{d}\dot{u} a^{(p_2+1)}(x_2)\{1+\mathrm{o_p}(1)\}
\end{aligned}$$

by applying the law of large number for strong mixing process in Theorem 2.20 of Fan and Yao (2003) under Assumptions 4.A1–4.A4 and 4.A6–4.A7.

By a change of variables $\mathring{x}_2 = x_2 + g\mathring{v}$ and $\mathring{u} = \dot{u} + h_2\dot{v}$ and the first-order Taylor expansion of the joint density f, we have

$$\begin{aligned}
P_{2a} =& \frac{g^{p_2+1}}{(p_2+1)!} \int\int \frac{f(\mathring{x}_1, x_2 + g\mathring{v}, \mathring{z}_1, \dot{u} + h_2\dot{v}) f_U(\dot{u})}{f_{X_2U}(x_2, \dot{u})} \mathcal{L}_l(\mathring{v}, x_2, \dot{u}, \mathring{x}_1, \mathring{z}_1) K_2(\dot{v}) \\
& \times \mathring{v}^{p_2+1} \mathring{x}_1^\top \mathrm{d}\mathring{x}_1 \mathrm{d}\mathring{v} \mathrm{d}\mathring{z}_1 \mathrm{d}\dot{v} \mathrm{d}\dot{u} a^{(p_2+1)}(x_2) \{1 + \mathrm{o_p}(1)\} \\
=& \frac{g^{p_2+1}}{(p_2+1)!} \int\int \frac{f(\mathring{x}_1, x_2, \mathring{z}_1, \dot{u}) f_U(\dot{u})}{f_{X_2U}(x_2, \dot{u})} \mathcal{L}_l(\mathring{v}, x_2, \dot{u}, \mathring{x}_1, \mathring{z}_1) K_2(\dot{v}) \\
& \times \mathring{v}^{p_2+1} \mathring{x}_1^\top \mathrm{d}\mathring{x}_1 \mathrm{d}\mathring{v} \mathrm{d}\mathring{z}_1 \mathrm{d}\dot{v} \mathrm{d}\dot{u} a^{(p_2+1)}(x_2) \{1 + \mathrm{o_p}(1) + \mathrm{O}(g + h_2)\} \\
=& \frac{g^{p_2+1}}{(p_2+1)!} \int \left\{ \int f(\mathring{x}_1, \mathring{z}_1 | x_2, \dot{u}) \mathcal{L}_l(\mathring{v}, x_2, \dot{u}, \mathring{x}_1, \mathring{z}_1) \mathring{v}^{p_2+1} \mathring{x}_1^\top \mathrm{d}\mathring{v} \mathrm{d}\mathring{x}_1 \mathrm{d}\mathring{z}_1 \right\} \\
& \times f_U(\dot{u}) \mathrm{d}\dot{u} a^{(p_2+1)}(x_2) \{1 + \mathrm{o_p}(1) + \mathrm{O}(g + h_2)\} \\
=& \frac{g^{p_2+1}}{(p_2+1)!} \int \mathring{v}^{p_2+1} E[\mathcal{L}_l(\mathring{v}, x_2, U, X_1, Z_1) X_1^\top] \mathrm{d}\mathring{v} a^{(p_2+1)}(x_2) \{1 + \mathrm{o_p}(1) + \mathrm{O}(g + h)\} \\
=& g^{p_2+1} \eta_l(x_2) + \mathrm{o_p}(g^{p_2+1}) + g^{p_2+1} \mathrm{O}(g + h_2) \\
=& g^{p_2+1} \eta_l(x_2) + \mathrm{o_p}(g^{p_2+1}).
\end{aligned}$$

The O-term in the second equality is due to the uniformly boundedness condition on the partial derivatives of the joint density f with respect to X_2 and U in Assumption 4.A4. The $\mathrm{o_p}$- and O-terms in the second last equality follow from the existence of $\eta_l(x_2)$, which is implied by the compact support, $\mathring{v} \in [-1, 1]$, of the kernel L in Assumption 4.A1, the existence of $S^{-1}(x_2, u)$ due to Assumption 4.A7, the boundedness of the variables X_1 and Z_1 and the densities f_U, f_{X_2U}, and f in Assumption 4.A4, and the existence of $a^{(p_2+1)}(\cdot)$ by Assumption 4.A2.

Analogously to the proof of $P_{1b} = \mathrm{o_p}(n^{-1/2} g^{-1/2})$ and $P_{1c} = \mathrm{o_p}(n^{-1/2} g^{-1/2})$, one can show that P_{2b} and P_{2c} are $\mathrm{o_p}(n^{-1/2} g^{-1/2})$ for $n \to \infty$ as well.

For part (iii), by the law of large number in Fan and Yao (2003, Theorem 2.20) under Assumptions 4.A1–4.A4 and 4.A6–4.A7, the leading term of $P_3(x_2)$ is given by

$$\begin{aligned}P_{3a} =& \frac{1}{ngh_2^d}\sum_{j=1}^{n}\int \frac{f_U(\dot{u})}{f_{X_2U}(x_2,\dot{u})}\mathcal{L}_l\left(\frac{X_{2j}-x_2}{g}, x_2, \dot{u}, X_{1j}, Z_{1j}\right) K_2\left(\frac{U_j-\dot{u}}{h_2}\right)\\ &\times [\lambda(U_j)-\lambda(\dot{u})]\mathrm{d}\dot{u}\\ =& \frac{1}{gh_2^d}\int\int \frac{f(\mathring{x}, \mathring{z}_1, \mathring{u})f_U(\dot{u})}{f_{X_2U}(x_2,\dot{u})}\mathcal{L}_l\left(\frac{\mathring{x}_2-x_2}{g}, x_2, \dot{u}, \mathring{x}_1, \mathring{z}_1\right) K_2\left(\frac{\mathring{u}-\dot{u}}{h_2}\right)\\ &\times [\lambda(\mathring{u})-\lambda(\dot{u})]\mathrm{d}\mathring{x}\mathrm{d}\mathring{z}_1\mathrm{d}\mathring{u}\mathrm{d}\dot{u}\{1+\mathrm{o_p}(1)\},\end{aligned}$$

which, after a change of variables $\mathring{x}_2 = x_2 + g\mathring{v}$ and $\mathring{u} = \dot{u} + h_2\dot{v}$, becomes

$$\begin{aligned}P_{3a} =& \int\int \frac{f(\mathring{x}_1, x_2+g\mathring{v}, \mathring{z}_1, \dot{u}+h_2\dot{v})f_U(\dot{u})}{f_{X_2U}(x_2,\dot{u})}\mathcal{L}_l(\mathring{v}, x_2, \dot{u}, \mathring{x}_1, \mathring{z}_1)K_2(\dot{v})\\ &\times [\lambda(\dot{u}+h_2\dot{v})-\lambda(\dot{u})]\mathrm{d}\mathring{x}_1\mathrm{d}\mathring{v}\mathrm{d}\mathring{z}_1\mathrm{d}\dot{v}\mathrm{d}\dot{u}\{1+\mathrm{o_p}(1)\}\\ =& \mathrm{O_p}(h_2^{q_2}) = \mathrm{o_p}(n^{-1/2}g^{-1/2}) \qquad \text{(by Assumption 4.A8(i))}\end{aligned}$$

by the Taylor expansion to q_2th degree of the coefficient function λ, which has bounded continuous q_2th partial derivatives, and to the first degree of the joint density f, where its partial derivatives with respect to X_2 and U are uniformly bounded. □

Lemma 4.11. *Under Assumptions 4.A1, 4.A3–4.A7, and 4.A8(iv), we have as* $n \to +\infty$,

$$\sup_{x_2\in D_{X_2}, u\in D_U}\left\|\tilde{S}_n(x_2,u) - f_{X_2U}(x_2,u)S(x_2,u)\right\| = \mathrm{O_p}(v_{2n}),$$

where $v_{2n} = \sqrt{\ln n/(ngh_2^d)} + g + h_2$.

Proof. Using the definition of $\mathcal{X}_j$ in equation (4.7), we write $\tilde{S}_n(x_2,u)$ into several partitioned block matrices:

$$\tilde{S}_n(x_2,u) = \begin{pmatrix}\tilde{S}_{XX,n}(x_2,u) & \tilde{S}_{XZ,n}(x_2,u)\\ \tilde{S}_{ZX,n}(x_2,u) & \tilde{S}_{ZZ,n}(x_2,u)\end{pmatrix},$$

where $\rho_2(v) = (1, v, \ldots, v^{p_2})^\top$,

$$\begin{aligned}\tilde{S}_{XX,n}(x_2,u) =& \frac{1}{ngh_2^d}\sum_{j=1}^{n}\rho_2\left(\frac{X_{2j}-x_2}{g}\right)\rho_2^\top\left(\frac{X_{2j}-x_2}{g}\right)\otimes X_{1j}X_{1j}^\top \\ &\times L\left(\frac{X_{2j}-x_2}{g}\right)K_2\left(\frac{U_j-u}{h_2}\right), \\ \tilde{S}_{XZ,n}(x_2,u) =& \tilde{S}_{ZX,n}^\top(x_2,u) \\ =& \frac{1}{ngh_2^d}\sum_{j=1}^{n}\rho_2\left(\frac{X_{2j}-x_2}{g}\right)\otimes X_{1j}Z_{1j}^\top L\left(\frac{X_{2j}-x_2}{g}\right)K_2\left(\frac{U_j-u}{h_2}\right),\end{aligned}$$

and

$$\tilde{S}_{ZZ,n}(x_2,u) = \frac{1}{ngh_2^d}\sum_{j=1}^{n}Z_{1j}Z_{1j}^\top L\left(\frac{X_{2j}-x_2}{g}\right)K_2\left(\frac{U_j-u}{h_2}\right).$$

By Assumptions 4.A1, 4.A3–4.A7, and 4.A8(iv), the conditions of Theorem 2 in Hansen (2008) are satisfied. Applying the uniform consistency result for a general kernel average in Hansen (2008, Theorem 2) yields

$$\sup_{x_2\in D_{X_2},u\in D_U}\left\|\tilde{S}_{XZ,n}(x_2,u)-\mathrm{E}\tilde{S}_{XZ,n}(x_2,u)\right\| = \mathrm{O_p}\left(\sqrt{\frac{\ln n}{ngh_2^d}}\right). \tag{4.26}$$

The expectation of $\tilde{S}_{XZ,n}(x_2,u)$ is

$$\begin{aligned}&\mathrm{E}\tilde{S}_{XZ,n}(x_2,u) \\ &= \frac{1}{gh_2^d}\mathrm{E}\left[\rho_2\left(\frac{X_{2j}-x_2}{g}\right)\otimes\left(X_{1j}Z_{1j}^\top\right)L\left(\frac{X_{2j}-x_2}{g}\right)K_2\left(\frac{U_j-u}{h_2}\right)\right] \\ &= \frac{1}{gh_2^d}\int\rho_2\left(\frac{X_{2j}-x_2}{g}\right)\otimes\int X_{1j}Z_{1j}^\top f(X_{1j},X_{2j},Z_{1j},U_j)\mathrm{d}X_{1j}\mathrm{d}Z_{1j} \\ &\quad\times L\left(\frac{X_{2j}-x_2}{g}\right)K_2\left(\frac{U_j-u}{h_2}\right)\mathrm{d}U_j\mathrm{d}X_{2j}.\end{aligned}$$

By a change of variables and the first-order Taylor expansion of f, we have

$$\begin{aligned}
&\mathrm{E}\tilde{S}_{XZ,n}(x_2,u)\\
&\quad=\int\rho_2(\dot{v})\otimes\int X_{1j}Z_{1j}^{\top}f(X_{1j},x_2+g\dot{v},Z_{1j},u+h_2\dot{u})\mathrm{d}X_{1j}\mathrm{d}Z_{1j}\\
&\qquad\times L(\dot{v})K_2(\dot{u})\mathrm{d}\dot{v}\mathrm{d}\dot{u}\\
&\quad=f_{X_2U}(x_2,u)\int K_2(\dot{u})\mathrm{d}\dot{u}\int\rho_2(\dot{v})L(\dot{v})\mathrm{d}\dot{v}\\
&\qquad\otimes\int X_{1j}Z_{1j}^{\top}\frac{f(X_{1j},x_2,Z_{1j},u)}{f_{X_2U}(x_2,u)}\mathrm{d}X_{1j}\mathrm{d}Z_{1j}+\mathrm{O}(g+h_2)\\
&\quad=f_{X_2U}(x_2,u)\int\rho_2(\dot{v})\ L(\dot{v})\mathrm{d}\dot{v}\otimes\mathrm{E}\left[X_{1j}Z_{1j}^{\top}|X_2=x_2,U=u\right]\\
&\qquad+\mathrm{O}(g+h_2).
\end{aligned}\tag{4.27}$$

The O-term in the second equality follows from the boundedness of the variables X_1 and Z_1, the marginal density f_{X_2U}, and the partial derivatives of f with respect to X_2 and U in Assumption 4.A4. Combining (4.26) and (4.27) yields the consistency result for $\tilde{S}_{XZ,n}(x_2,u)$:

$$\begin{aligned}
&\sup_{x_2\in D_{X_2},u\in D_U}\left\|\tilde{S}_{XZ,n}(x_2,u)-f_{X_2U}(x_2,u)\int\rho_2(\dot{v})(\dot{v})\mathrm{d}\dot{v}\otimes S_{XZ}(x_2,u)\right\|\\
&\qquad=\mathrm{O_p}(v_{2n}).
\end{aligned}$$

The consistency results for $\tilde{S}_{XX,n}(x_2,u)$ and $\tilde{S}_{ZZ,n}(x_2,u)$ can be proved in a similar way. □

Lemma 4.12. *Under Assumptions 4.A1–4.A8, as* $n\to\infty$,

(i) $\tilde{S}_n^{-1}(x_2,u)=f_{X_2U}^{-1}(x_2,u)S^{-1}(x_2,u)(1+\mathrm{o_p}(1))$ *uniformly in* $x_2\in D_{X_2}$ *and* $u\in D_U$,

(ii) $R_1(x_2)+R_2(x_2)+R_3(x_2)=\mathrm{o_p}(n^{-1/2}g^{-1/2})$.

Proof. By Lemma 4.11 and Assumption 4.A8(iv),

$$\sup_{x_2\in D_{X_2},u\in D_U}\left\|\tilde{S}_n(x_2,u)-f_{X_2U}(x_2,u)S(x_2,u)\right\|=\mathrm{O_p}(v_{2n})=\mathrm{o_p}(1),$$

and by the invertiblity of f_{X_2U} and $S(x_2,u)$ in Assumptions 4.A4 and 4.A7, respectively, $\tilde{S}_n^{-1}(x_2,u) = f_{X_2U}^{-1}(x_2,u)S^{-1}(x_2,u)(1+\mathrm{o_p}(1))$ as $n\to\infty$. Together with the asymptotic results for $P_c(x_2), c=1,2,3$, in Lemma 4.10, we have

$$
\begin{aligned}
&|R_1(x_2)+R_2(x_2)+R_3(x_2)| \\
&\quad\le \sum_{c=1}^{3}\left|\Phi\left(\tilde{S}_n^{-1}(x_2,u)\left[\tilde{S}_n(x_2,u)-f_{X_2U}(x_2,u)S(x_2,u)\right]\frac{S^{-1}(x_2,u)}{f_{X_2U}(x_2,u)},\tilde{T}_{n,c}(x_2,U_i)\right)\right| \\
&\quad\le \sum_{c=1}^{3}\left|\Phi\left(\frac{S^{-1}(x_2,u)}{f_{X_2U}(x_2,u)},\tilde{T}_{n,c}(x_2,U_i)\right)\right|\cdot \sup_{x_2\in D_{X_2},u\in D_U}\left\|\tilde{S}_n(x_2,u)-f_{X_2U}(x_2,u)S(x_2,u)\right\| \\
&\qquad\times \sup_{x_2\in D_{X_2},u\in D_U}\left\|\tilde{S}_n^{-1}(x_2,u)\right\| \\
&\quad\le \sum_{c=1}^{3}|P_c(x_2)|\cdot \sup_{x_2\in D_{X_2},u\in D_U}\left\|\tilde{S}_n(x_2,u)-f_{X_2U}(x_2,u)S(x_2,u)\right\| \\
&\qquad\times \sup_{x_2\in D_{X_2},u\in D_U}\left\|\tilde{S}_n^{-1}(x_2,u)\right\| \\
&\quad= \mathrm{O_p}(n^{-1/2}g^{-1/2})\cdot\mathrm{o_p}(1)\cdot\mathrm{O_p}(1)=\mathrm{o_p}(n^{-1/2}g^{-1/2}).
\end{aligned}
$$

□

Proof of Theorem 4.3

Before proving Theorem 4.3, we first derive the uniform convergence rate for the first stage estimator $\hat{\Pi}_m(\cdot)$ for $m=1,\ldots,d$, which will be used in several places later. From the definition of $\hat{\Pi}_m(Z_j)$ in (4.5),

$$
\begin{aligned}
\hat{\Pi}_m(Z_j) &= e_1^\top \bar{S}_n^{-1}(Z_j)\bar{T}_n(Z_j) \\
&= e_1^\top \bar{S}_n^{-1}(Z_j)\frac{1}{nh_1^r}\sum_{k\neq j}^{n}K_1\left(\frac{Z_k-Z_j}{h_1}\right)\rho_1\left(\frac{Z_k-Z_j}{h_1}\right)X_{mk},
\end{aligned}
$$

where $e_1 = (1, 0, \ldots, 0)^\top$. Since

$$\begin{aligned}
&\frac{e_1^\top \bar{S}_n^{-1}(Z_j)}{nh_1^r} \sum_{k\neq j}^{n} K_1\left(\frac{Z_k - Z_j}{h_1}\right) \rho_1\left(\frac{Z_k - Z_j}{h_1}\right) \rho_1^\top\left(\frac{Z_k - Z_j}{h_1}\right) e_l \\
&\quad = e_1^\top \bar{S}_n^{-1}(Z_j)\bar{S}_n(Z_j)e_l = e_1^\top e_l = \begin{cases} 1, & \text{if } l{=}1, \\ 0, & \text{otherwise,} \end{cases}
\end{aligned}$$

we have the following discrete moment conditions:

$$\begin{aligned}
&\frac{e_1^\top \bar{S}_n^{-1}(Z_j)}{nh_1^r} \sum_{k\neq j}^{n} K_1\left(\frac{Z_k - Z_j}{h_1}\right) \rho_1\left(\frac{Z_k - Z_j}{h_1}\right) \left(\frac{Z_k - Z_j}{h_1}\right)^{\mathbf{q}} \\
&\quad = \begin{cases} 1, & \text{if } |\mathbf{q}| = 0, \\ 0, & \text{if } 1 \le |\mathbf{q}| \le p_1. \end{cases}
\end{aligned} \tag{4.28}$$

From the above discrete moment conditions with $|\mathbf{q}| = 0$,

$$\begin{aligned}
&\hat{\Pi}_m(Z_j) - \Pi_m(Z_j) \\
&\quad = \frac{e_1^\top \bar{S}_n^{-1}(Z_j)}{nh_1^r} \sum_{k\neq j}^{n} K_1\left(\frac{Z_k - Z_j}{h_1}\right) \rho_1\left(\frac{Z_k - Z_j}{h_1}\right) \{X_{mk} - \Pi_m(Z_j)\} \\
&\quad = \frac{e_1^\top \bar{S}_n^{-1}(Z_j)}{nh_1^r} \sum_{k\neq j}^{n} K_1\left(\frac{Z_k - Z_j}{h_1}\right) \rho_1\left(\frac{Z_k - Z_j}{h_1}\right) \{\Pi_m(Z_k) - \Pi_m(Z_j) + U_{mk}\}.
\end{aligned} \tag{4.29}$$

By the p_1th order Taylor expansion of Π_m for Z_k in a h_1-neighborhood of Z_j and the Lipschitz continuity of the p_1th partial derivatives of Π_m (Assumption 4.B2),

$$\Pi_m(Z_k) - \Pi_m(Z_j) = \sum_{1\le|\mathbf{q}|\le p_1} \frac{h_1^{\mathbf{q}}}{\mathbf{q}!}(D^{\mathbf{q}}\Pi_m)(Z_j)\left(\frac{Z_k - Z_j}{h_1}\right)^{\mathbf{q}} + \mathrm{O}(h_1^{p_1+1}). \tag{4.30}$$

Using equations (4.28) with $|\mathbf{q}| = 0$ and with $|\mathbf{q}| \in [1, p_1]$, (4.29), and (4.30), we have

$$\begin{aligned}
&\hat{\Pi}_m(Z_j) - \Pi_m(Z_j) \\
&\quad = \frac{e_1^\top \bar{S}_n^{-1}(Z_j)}{nh_1^r} \sum_{k\neq j}^{n} \rho_1\left(\frac{Z_k - Z_j}{h_1}\right) K_1\left(\frac{Z_k - Z_j}{h_1}\right) \left[U_{mk} + \mathrm{O}(h_1^{p_1+1})\right] \\
&\quad = e_1^\top \bar{S}_n^{-1}(Z_j)\bar{T}_{1,n}(Z_j) + \mathrm{O_p}(h_1^{p_1+1}),
\end{aligned} \tag{4.31}$$

where

$$\bar{T}_{1,n}(Z_j) = \frac{1}{nh_1^r}\sum_{k\neq j}^{n}\rho_1\left(\frac{Z_k - Z_j}{h_1}\right)K_1\left(\frac{Z_k - Z_j}{h_1}\right)U_{mk}.$$

□

Lemma 4.13. *Under Assumptions 4.B1, 4.B3–4.B6 and 4.B7(i), as* $n \to \infty$,

(i) $\sup_{z\in D_Z}\left\|\bar{S}_n(z) - f_Z(z)M_1\right\| = \mathrm{O_p}(\sqrt{\ln n/(nh_1^r)} + h^2)$,

(ii) $\bar{S}_n^{-1}(z) = f_Z^{-1}(z)M_1^{-1}(1+\mathrm{o_p}(1))$ *uniformly in* $z \in D_Z$, *and*

(iii) $\sup_{z\in D_Z}\left\|\bar{T}_{1,n}(z)\right\| = \mathrm{O_p}(\sqrt{\ln n/(nh_1^r)})$,

where $M_1 = \int \rho_1(v)\rho_1^\top(v)K_1(v)\mathrm{d}v$.

Proof. This lemma is analogous to Lemmas 4.11 and 4.12(i), and the results follow by application of Hansen (2008) Theorem 2 under Assumptions 4.B1, 4.B3–4.B6, and 4.B7(i). □

By the invertiblity of f_Z and M_1 given in Assumptions 4.B4 and 4.B6, respectively, equation (4.31), and Lemma 4.13, one obtains the uniform convergence rate for the local polynomial estimator of $\Pi_m(\cdot)$:

$$\begin{aligned}
&\left|\hat{\Pi}_m(Z_j) - \Pi_m(Z_j)\right| \\
&\quad\le e_1^\top \sup_{z\in D_Z}\left\|\bar{S}_n^{-1}(z)\right\| \sup_{z\in D_Z}\left\|\bar{T}_{n,1}(z)\right\| + \mathrm{O_p}(h_1^{p_1+1}) \\
&\quad= \frac{e_1^\top M_1^{-1}(1+\mathrm{o_p}(1))}{\inf_{z\in D_Z} f_Z(z)}\mathrm{O_p}(v_{1n}) + \mathrm{O_p}(h_1^{p_1+1}) \\
&\quad= \mathrm{O_p}(v_{1n} + h_1^{p_1+1})
\end{aligned} \tag{4.32}$$

uniformly in $Z_j \in D_Z$, where $v_{1n} = \sqrt{\ln n/(nh_1^r)}$.

Next, similar to the proof of Theorem 4.2, we are going to split $\hat{a}_l(x_2) - \tilde{a}(x_2)$ into several terms and then investigate their asymptotic behaviors separately. To this end, using the

identity that $A_1^{-1} - A_2^{-1} = A_1^{-1}(A_2 - A_1)A_2^{-1}$ yields

$$\tilde{S}_n^{-1}(x_2, U_i) = \hat{S}_n^{-1}(x_2, \hat{U}_i) - \hat{Q}_n(x_2, \hat{U}_i), \tag{4.33}$$

where

$$\hat{S}_n(x_2, \hat{U}_i) = \frac{1}{ngh_2^d} \sum_{j=1}^{n} L\left(\frac{X_{2j} - x_2}{g}\right) K_2\left(\frac{\hat{U}_j - \hat{U}_i}{h_2}\right) \mathcal{X}_j \mathcal{X}_j^\top,$$

which is shown to be invertible in Lemma 4.16 for a sufficiently large n, and

$$\hat{Q}_n(x_2, \hat{U}_i) = \hat{S}_n^{-1}(x_2, \hat{U}_i)[\tilde{S}_n(x_2, U_i) - \hat{S}_n(x_2, \hat{U}_i)]\tilde{S}_n^{-1}(x_2, U_i).$$

By equations (4.21) and (4.33),

$$\begin{aligned}
\tilde{a}_l(x_2) =& \frac{1}{n} \sum_{i=1}^{n} e_{l+1}^\top \left[\hat{S}_n^{-1}(x_2, \hat{U}_i) - \hat{Q}_n(x_2, \hat{U}_i)\right] \bar{\bar{T}}_n(x_2, U_i) + e_{l+1}^\top e_1 R_n \\
&+ a_l(x_2) + \mathrm{o_p}(g^{p_2+1}) \\
=& \frac{1}{n} \sum_{i=1}^{n} e_{l+1}^\top \hat{S}_n^{-1}(x_2, \hat{U}_i) \bar{\bar{T}}_n(x_2, U_i) - \mathcal{R}_0(x_2) + e_{l+1}^\top e_1 R_n + a_l(x_2) \\
&+ \mathrm{o_p}(g^{p_2+1})
\end{aligned} \tag{4.34}$$

in which $\mathcal{R}_0(x_2) = \Phi(\hat{Q}_n(x_2, \hat{U}_i), \bar{\bar{T}}_n(x_2, U_i))$. Similarly to equation (4.21), one can show that

$$\begin{aligned}
\hat{a}_l(x_2) =& \frac{1}{n} \sum_{i=1}^{n} e_{l+1}^\top \hat{S}_n^{-1}(x_2, \hat{U}_i) \frac{1}{ngh_2^d} \sum_{j=1}^{n} L\left(\frac{X_{2j} - x_2}{g}\right) K_2\left(\frac{\hat{U}_j - \hat{U}_i}{h_2}\right) \mathcal{X}_j \{\epsilon_j \\
&+ \frac{g^{p_2+1}}{(p_2+1)!} \left(\frac{X_{2j} - x_2}{g}\right)^{p_2+1} X_{1j}^\top a^{(p_2+1)}(x_2) + [\lambda(U_j) - \lambda(U_i)] \\
&+ \underbrace{\mathcal{X}_j^\top e_1}_{=1} \lambda(U_i) + \mathcal{X}_j^\top \beta(x_2) + \underbrace{\mathcal{X}_j^\top e_1}_{=1} \mathrm{o}(g^{p_2+1})\} \\
=& \frac{1}{n} \sum_{i=1}^{n} e_{l+1}^\top \hat{S}_n^{-1}(x_2, \hat{U}_i) \bar{\hat{T}}_n(x_2, \hat{U}_i) + e_{l+1}^\top e_1 R_n + a_l(x_2) + \mathrm{o_p}(g^{p_2+1}),
\end{aligned} \tag{4.35}$$

where

$$\hat{T}_{n,1}(x_2,\hat{U}_i)=\frac{1}{ngh_2^d}\sum_{j=1}^{n}L\left(\frac{X_{2j}-x_2}{g}\right)K_2\left(\frac{\hat{U}_j-\hat{U}_i}{h_2}\right)\mathcal{X}_j\epsilon_j,$$

$$\hat{T}_{n,2}(x_2,\hat{U}_i)=\frac{g^{p_2+1}}{(p_2+1)!ngh_2^d}\sum_{j=1}^{n}L\left(\frac{X_{2j}-x_2}{g}\right)K_2\left(\frac{\hat{U}_j-\hat{U}_i}{h_2}\right)\mathcal{X}_j \times\left(\frac{X_{2j}-x_2}{g}\right)^{p_2+1}X_{1j}^{\top}a^{(p_2+1)}(x_2),$$

and

$$\hat{T}_{n,3}(x_2,\hat{U}_i)=\frac{1}{ngh_2^d}\sum_{j=1}^{n}L\left(\frac{X_{2j}-x_2}{g}\right)K_2\left(\frac{\hat{U}_j-\hat{U}_i}{h_2}\right)\mathcal{X}_j[\lambda(U_j)-\lambda(U_i)].$$

Note that $\lambda(\cdot)$ within $\hat{T}_{n,3}(x_2,\hat{U}_i)$ is a function of the latent U_j and U_j instead of the estimated $\hat{U}_j$ and $\hat{U}_j$. Besides, R_n in equation (4.35) is the same R_n as in equation (4.34).

Again by the fact that $A_1^{-1}-A_2^{-1}=A_1^{-1}(A_2-A_1)A_2^{-1}$, one obtains

$$\hat{S}_n^{-1}(x_2,\hat{U}_i)=\frac{S^{-1}(x_2,U_i)}{f_{X_2U}(x_2,U_i)}-\mathcal{Q}_n(x_2,\hat{U}_i), \tag{4.36}$$

where

$$\mathcal{Q}_n(x_2,\hat{U}_i)=\hat{S}_n^{-1}(x_2,\hat{U}_i)\left[\hat{S}_n(x_2,\hat{U}_i)-f_{X_2U}(x_2,U_i)S(x_2,U_i)\right]\frac{S^{-1}(x_2,U_i)}{f_{X_2U}(x_2,U_i)}.$$

It follows from (4.34), (4.35), and (4.36) that

$$\begin{aligned}\hat{a}_l(x_2)-\tilde{a}_l(x_2)=&\frac{1}{n}\sum_{i=1}^{n}e_l^{\top}\hat{S}_n^{-1}(x_2,\hat{U}_i)[\bar{\hat{T}}_n(x_2,\hat{U}_i)-\bar{\tilde{T}}_n(x_2,U_i)]+\mathcal{R}_0(x_2)+\mathrm{o}(g^{p_2+1})\\=&\sum_{c=1}^{3}\{\mathcal{P}_c(x_2)-\mathcal{R}_c(x_2)\}+\mathcal{R}_0(x_2)+\mathrm{o}(g^{p_2+1}),\end{aligned}$$

where for $c=1,2,3$,

$$\mathcal{P}_c(x_2)=\Phi\left\{\frac{S^{-1}(x_2,U_i)}{f_{X_2U}(x_2,U_i)},\hat{T}_{n,c}(x_2,\hat{U}_i)-\tilde{T}_{n,c}(x_2,U_i)\right\}$$

and

$$\mathcal{R}_c(x_2) = \Phi\{\mathcal{Q}_n(x_2, U_i), \hat{T}_{n,c}(x_2, \hat{U}_i) - \tilde{T}_{n,c}(x_2, U_i)\}.$$

We complete the proof of Theorem 4.3 by showing the terms $\mathcal{R}_0(x_2)$, $\mathcal{P}_c(x_2)$, and $\mathcal{R}_c(x_2)$, for $c = 1, 2, 3$, are all $\mathrm{o_p}(n^{-1/2}g^{-1/2})$ in Lemmas 4.14–4.18. □

First, we define variables

$$K_n^*(x_2, U_i) = \frac{1}{ngh_2^d} \sum_{j=1}^{n} L\left(\frac{X_{2j} - x_2}{g}\right) K^*\left(\frac{U_j - U_i}{h_2}\right),$$
$$T_{n,1}^*(x_2, U_i) = \frac{1}{ngh_2^d} \sum_{j=1}^{n} L\left(\frac{X_{2j} - x_2}{g}\right) K_2'\left(\frac{U_j - U_i}{h_2}\right) \mathcal{X}_j \epsilon_j,$$

and the population counterpart for $T_{n,1}^*(x_2, U_i)$

$$T^*(x_2, U_i) = \begin{pmatrix} \int \rho_2(u) K_2'(u) \mathrm{d}u \otimes \mathrm{E}[X_1 \epsilon | X_2 = x_2, U = U_i] \\ \mathrm{E}[Z_1 \epsilon | X_2 = x_2, U = U_i] \end{pmatrix},$$

where K_2' is the derivative of the kernel K_2, and K^* is a d-variate product kernel such that

$$K^*(u) = \frac{1}{4.1}\left(\mathbf{1}\{\|u\|_\infty \le 2\} + \mathbf{1}\{\|u\|_\infty \in (2, 2.1]\} \prod_{m=1}^{d} \{1 - 10(|u_m| - 2)\}\right).$$

Lemma 4.14. *Under Assumptions 4.A1, 4.A3–4.A4, 4.A6, 4.A8(iv), and 4.B1, we have as* $n \to \infty$,

(i) $\sup_{x_2 \in D_{X_2}, u \in D_U} |K_n^*(x_2, u) - f_{X_2U}(x_2, u)| = \mathrm{O_p}(v_{2n})$,

(ii) $\sup_{x_2 \in D_{X_2}, u \in D_U} \left\|T_{n,1}^*(x_2, u) - T^*(x_2, u) f_{X_2U}(x_2, u)\right\| = \mathrm{O_p}\left(\sqrt{\ln n/(ngh_2^d)}\right)$,

where $v_{2n} = \sqrt{\ln n/(ngh_2^d)} + g + h_2$.

Proof. Clearly, the kernel K^* is bounded and Lipschitz continuous. Together with Assumptions 4.A1, 4.A3–4.A4, 4.A6(iii) (with $\omega = 1$), and 4.A8(iv), the conditions in Hansen (2008) Theorem 6 are satisfied. By the uniform convergence result for kernel

density estimator (Hansen, 2008, Theorem 6), we completes the proof of (i). Part (ii) is similar to Lemma 4.11, and the convergence result follows by application of Hansen (2008) Theorem 2 under Assumptions 4.A1, 4.A3–4.A4, 4.A6, and 4.A8(iv), the continuously differentiability condition on K' (Assumption 4.B1). □

Lemma 4.15. *If Assumptions 4.A1–4.A8 and 4.B1–4.B7 are satisfied, then as* $n \to \infty$,

$$\text{(i)} \quad \sup_{x_2 \in D_{X_2}, U_i \in D_U} \left\| \hat{S}_n(x_2, \hat{U}_i) - \tilde{S}_n(x_2, U_i) \right\| = \mathrm{o_p}(1) \text{ and}$$

$$\text{(ii)} \quad \sup_{x_2 \in D_{X_2}, U_i \in D_U} \left\| \hat{S}_n(x_2, \hat{U}_i) - f_{X_2 U}(x_2, U_i) S(x_2, U_i) \right\| = \mathrm{o_p}(1).$$

Proof. Applying the mean value theorem to the kernel K_2, in which its partial derivatives are bounded under Assumption 4.B1, we have

$$\begin{aligned}
&\left\| \hat{S}_n(x_2, \hat{U}_i) - \tilde{S}_n(x_2, U_i) \right\| \\
&\quad \le \frac{1}{ngh_2^d} \sum_{j=1}^{n} \left\| \mathcal{X}_j \mathcal{X}_j^\top \right\| L\left(\frac{X_{2j} - x_2}{g}\right) \left| K_2\left(\frac{\hat{U}_j - \hat{U}_i}{h_2}\right) - K_2\left(\frac{U_j - U_i}{h_2}\right) \right| \\
&\quad \le \frac{1}{ngh_2^d} \sum_{j=1}^{n} \left\| \mathcal{X}_j \mathcal{X}_j^\top \right\| L\left(\frac{X_{2j} - x_2}{g}\right) K^*\left(\frac{U_j - U_i}{h_2}\right) \\
&\qquad \times \max_v |K_2'(v)| \cdot h_2^{-1} \cdot \max_{i,j} \|(\hat{U}_j - U_j) - (\hat{U}_i - U_i)\| \\
&\quad \le \max_j \left\| \mathcal{X}_j \mathcal{X}_j^\top \right\| K_n^*(x_2, U_i) \max_v |K_2'(v)| \cdot h_2^{-1} \max_{i,j} \|(\hat{U}_j - U_j) - (\hat{U}_i - U_i)\|
\end{aligned}$$

uniformly in $x_2 \in D_{X_2}$ and $U_i \in D_U$. Since the variables X_1, X_2, and Z_1 are bounded under Assumption 4.A4, $\max_i \left\| \mathcal{X}_j \mathcal{X}_j^\top \right\| < \infty$. By the uniform convergence result for $\hat{\Pi}_m(z)$ in (4.32),

$$\max_{1 \le j \le n} \|\hat{U}_j - U_j\| \le d \max_{1 \le m \le d} \max_{1 \le j \le n} |\hat{\Pi}_m(Z_j) - \Pi_m(Z_j)| = \mathrm{O_p}(v_{1n} + h_1^{p_1+1}),$$

where $v_{1n} = \sqrt{\ln n/(nh_1^r)}$. Together with the convergence result for $K_n^*(x_2, u)$ by Lemma 4.14(i) and the bounded density f_{X_2U} (Assumption 4.A4), we complete the proof of (i):

$$\begin{aligned}
&\sup_{x_2\in D_{X_2}, U_i\in D_U} \left\|\hat{S}_n(x_2,\hat{U}_i) - \tilde{S}_n(x_2,U_i)\right\| \\
&\quad \le C \sup_{x_2\in D_{X_2}, u\in D_U} |f_{X_2U}(x_2,u)|\,(1+\mathrm{o_p}(1))\cdot h_2^{-1}(v_{1n}+h_1^{p_1+1}) \\
&\quad = \mathrm{O_p}(h_2^{-1}(v_{1n}+h_1^{p_1+1})) = \mathrm{o_p}(1) \qquad \text{(by Assumption 4.B7(iii))} \qquad (4.37)
\end{aligned}$$

for some $C > 0$. Combining (4.37) with Lemma 4.11 gives the result in part (ii). □

Lemma 4.16. *Under Assumptions 4.A1–4.A8 and 4.B1–4.B7, as $n \to \infty$,*

(i) $\hat{S}_n^{-1}(x_2, U_i) = f_{X_2U}^{-1}(x_2, U_i)S^{-1}(x_2, U_i)(1 + \mathrm{o_p}(1))$ *uniformly in* $x_2 \in D_{X_2}$, $U_i \in D_U$,

(ii) $\mathcal{R}_0(x_2) = \mathrm{o_p}(n^{-1/2}g^{-1/2})$.

Proof. By Lemma 4.15 and Assumption 4.A8(iv),

$$\sup_{x_2\in D_{X_2}, u\in D_U} \left\|\hat{S}_n(x_2,u) - f_{X_2U}(x_2,u)S(x_2,u)\right\| = \mathrm{O_p}(\sqrt{\ln n/(ngh_2^d)} + g + h_2) = \mathrm{o_p}(1).$$

Then, part (i) follows from the invertiblity of f_{X_2U} and $S(x_2, u)$ in Assumptions 4.A4 and 4.A7, respectively. To show (ii), we write

$$\begin{aligned}
|\mathcal{R}_0(x_2)| &= \left|\Phi(\hat{Q}_n(x_2,\hat{U}_i), \bar{\tilde{T}}_n(x_2,U_i))\right| \\
&= \left|\frac{e_{l+1}^\top}{n}\sum_{i=1}^n \hat{S}_n^{-1}(x_2,\hat{U}_i)[\tilde{S}_n(x_2,U_i) - \hat{S}_n(x_2,\hat{U}_i)]\tilde{S}_n^{-1}(x_2,U_i)\bar{\tilde{T}}_n(x_2,U_i)\right| \\
&\le e_{l+1}^\top \sup_{x_2\in D_{X_2}, U_i\in D_U}\left\|\hat{S}_n^{-1}(x_2,\hat{U}_i)\right\| \sup_{x_2\in D_{X_2}, U_i\in D_U}\left\|\tilde{S}_n(x_2,U_i) - \hat{S}_n(x_2,\hat{U}_i)\right\| \\
&\quad \times \left|\frac{1}{n}\sum_{i=1}^n \tilde{S}_n^{-1}(x_2,U_i)\bar{\tilde{T}}_n(x_2,U_i)\right|.
\end{aligned}$$

By part(i), Lemma 4.15, and Assumptions 4.A4 and 4.A7,

$$\begin{aligned}|\mathcal{R}_0(x_2)| \leq & e_{l+1}^\top \frac{\sup_{x_2\in D_{X_2}, U_i\in D_U} \left\| S_n^{-1}(x_2,\hat{U}_i)\right\| (1+\mathrm{o_p}(1))}{\inf_{x_2\in D_{X_2}, U_i\in D_U} f_{X_2U}(x_2,U_i)} \mathrm{o_p}(1) \\ & \times \left| \frac{1}{n}\sum_{i=1}^n \tilde{S}_n^{-1}(x_2,U_i)\bar{\bar{T}}_n(x_2,U_i)\right| \\ = & e_{l+1}^\top \mathrm{O_p}(1)\mathrm{o_p}(1) \left| \frac{1}{n}\sum_{i=1}^n \tilde{S}_n^{-1}(x_2,U_i)\bar{\bar{T}}_n(x_2,U_i)\right|.\end{aligned}$$

The claim (ii) now follows from the asymptotic normality result in Theorem 4.2:

$$|\mathcal{R}_0(x_2)| \leq e_{l+1}^\top \mathrm{O_p}(1)\mathrm{o_p}(1)\mathrm{O_p}(n^{-1/2}g^{-1/2}) = \mathrm{o_p}(n^{-1/2}g^{-1/2}).$$

□

Lemma 4.17. *Under Assumptions 4.A1–4.A8 and 4.B1–4.B7, for $n\to\infty$ and $c=1,2,3$,*

$$\mathcal{P}_c(x_2) = \mathrm{o_p}(n^{-1/2}g^{-1/2}).$$

Proof. Here we only consider the case of the term $\mathcal{P}_1(x_2)$ as $\mathcal{P}_2(x_2)$ and $\mathcal{P}_3(x_2)$ can be proven in a similar manner. By the first-order Taylor expansion of the kernel K_2 under Assumption 4.B1, the conditions for the convergence rates of the bandwidths in 4.B7(iii), and (4.32)

$$\begin{aligned}& K_2\left(\frac{\hat{U}_j-\hat{U}_i}{h_2}\right) - K_2\left(\frac{U_j-U_i}{h_2}\right) \\ & = K_2'^\top\left(\frac{U_j-U_i}{h_2}\right)\left[\frac{\hat{U}_j-U_j}{h_2} - \frac{\hat{U}_i-U_i}{h_2}\right] + \mathrm{O_p}(h_2^{-1}v_{1n} + h_2^{-1}h_1^{p_1+1})^2 \\ & = \frac{1}{h_2}K_2'^\top\left(\frac{U_j-U_i}{h_2}\right)[\hat{\Pi}(Z_i) - \Pi(Z_i) - \hat{\Pi}(Z_j) + \Pi(Z_j)] + \mathrm{o_p}(n^{-1/2}g^{-1/2}). \quad (4.38)\end{aligned}$$

By equation (4.31), the convergence results for $\bar{S}_n^{-1}(z)$ and $\bar{T}_{n,1}(z)$ in Lemma 4.13, and the invertiblity of f_Z and M_1 in Assumptions 4.B3 and 4.B6,

$$\begin{aligned}
&\hat{\Pi}_m(Z_j) - \Pi_m(Z_j) \\
&\quad = e_1^\top \bar{S}_n^{-1}(Z_j)\bar{T}_{n,1}(Z_j) + \mathrm{O_p}(h_1^{p_1+1}) \\
&\quad = \frac{e_1^\top M_1^{-1}}{f_Z(Z_j)} \frac{1}{nh_1^r} \sum_{k\neq j} \rho_1\left(\frac{Z_k - Z_j}{h_1}\right) K_1\left(\frac{Z_k - Z_j}{h_1}\right) U_{mk} \\
&\qquad + \frac{e_1^\top M_1^{-1}\mathrm{o_p}(1)}{f_Z(Z_j)} \mathrm{O_p}(v_{1n}) + \mathrm{O_p}(h_1^{p_1+1}) \\
&\quad = \frac{e_1^\top M_1^{-1}}{f_Z(Z_j)} \frac{1}{nh_1^r} \sum_{k\neq j} \rho_1\left(\frac{Z_k - Z_j}{h_1}\right) K_1\left(\frac{Z_k - Z_j}{h_1}\right) U_{mk} + \mathrm{O_p}(v_{1n} + h_1^{p_1+1}),
\end{aligned}$$

which implies that

$$\begin{aligned}
&\hat{\Pi}(Z_i) - \Pi(Z_i) - \hat{\Pi}(Z_j) + \Pi(Z_j) \\
&\quad = \frac{e_1^\top M_1^{-1}}{nh_1^r} \sum_{k\neq j\neq i} \left[\frac{\rho_1\left(\frac{Z_k - Z_i}{h_1}\right)}{f_Z(Z_i)} K_1\left(\frac{Z_k - Z_i}{h_1}\right) - \frac{\rho_1\left(\frac{Z_k - Z_j}{h_1}\right)}{f_Z(Z_j)} K_1\left(\frac{Z_k - Z_j}{h_1}\right) \right] U_k \\
&\qquad + \mathrm{o_p}(h_2 \cdot n^{-1/2} g^{-1/2})
\end{aligned} \tag{4.39}$$

by Assumption 4.B7(iii). Combining equations (4.38) and (4.39) yields

$$\begin{aligned}
\mathcal{P}_1(x_2) =& \frac{e_{l+1}^\top}{n} \sum_{i=1}^n \frac{S^{-1}(x_2, U_i)}{f_{X_2U}(x_2, U_i)} \left[\hat{T}_{n,1}(x_2, \hat{U}_i) - \tilde{T}_{n,1}(x_2, U_i)\right] \\
=& \frac{e_{l+1}^\top}{n^2 g h_2^d} \sum_{i=1}^n \frac{S^{-1}(x_2, U_i)}{f_{X_2U}(x_2, U_i)} \sum_{j=1}^n L\left(\frac{X_{2j} - x_2}{g}\right) \left[K_2\left(\frac{\hat{U}_j - \hat{U}_i}{h_2}\right) \right. \\
& \left. - K_2\left(\frac{U_j - U_i}{h_2}\right) \right] \mathcal{X}_j \epsilon_j \\
=& \mathcal{P}_1^{(1)} - \mathcal{P}_1^{(2)} + \mathrm{o_p}(n^{-1/2} g^{-1/2})
\end{aligned} \tag{4.40}$$

where $\mathcal{L}_l$ is defined in Lemma 4.10,

$$\begin{aligned}\mathcal{P}_{n,1}^{(1)} =& \frac{1}{n^3 g h_2^{d+1} h_1^r} \sum_{k\neq i\neq j} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{f_{X_2U}(x_2, U_i)} \mathcal{L}_l \left(\frac{X_{2j} - x_2}{g}, x_2, U_i, X_{1j}, Z_{1j} \right) \epsilon_j \\ & \times K_2' \left(\frac{U_j - U_i}{h_2} \right)^\top U_k \frac{e_1^\top M_1^{-1} \rho_1 \left(\frac{Z_k - Z_i}{h_1} \right)}{f_Z(Z_i)} K_1 \left(\frac{Z_k - Z_i}{h_1} \right),\end{aligned}$$

and

$$\begin{aligned}\mathcal{P}_{n,1}^{(2)} =& \frac{1}{n^3 g h_2^{d+1} h_1^r} \sum_{k\neq i\neq j} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{f_{X_2U}(x_2, U_i)} \mathcal{L}_l \left(\frac{X_{2j} - x_2}{g}, x_2, U_i, X_{1j}, Z_{1j} \right) \epsilon_j \\ & \times K_2' \left(\frac{U_j - U_i}{h_2} \right)^\top U_k \frac{e_1^\top M_1^{-1} \rho_1 \left(\frac{Z_k - Z_j}{h_1} \right)}{f_Z(Z_j)} K_1 \left(\frac{Z_k - Z_j}{h_1} \right).\end{aligned}$$

The $\mathrm{o_p}$-term in equation (4.40) follows from equations the convergence result for $T_{n,1}^*(x_2, u)$ in Lemma 4.14. We also write the sum over k in the terms $\mathcal{P}_{n,1}^{(1)}$ and $\mathcal{P}_{n,1}^{(2)}$ without $i = j$, since the corresponding summands in are canceled out for $i = j$. We complete the proof for $\mathcal{P}_1(x_2)$ by showing $\mathcal{P}_{n,1}^{(1)} = \mathrm{o_p}(n^{-1/2} g^{-1/2})$ and $\mathcal{P}_{n,1}^{(2)} = \mathrm{o_p}(n^{-1/2} g^{-1/2})$.

To show $\mathcal{P}_{n,1}^{(1)} = \mathrm{o_p}(n^{-1/2} g^{-1/2})$, we first let $\zeta_j = (Y_i, X_i^\top, Z_i^\top, U_i^\top)^\top$ and

$$\begin{aligned}\psi_0(\zeta_k, \zeta_i, \zeta_j) =& \frac{1}{f_{X_2U}(x_2, U_i)} \mathcal{L}_l \left(\frac{X_{2j} - x_2}{g}, x_2, U_i, X_{1j}, Z_{1j} \right) \epsilon_j K_2'^\top \left(\frac{U_j - U_i}{h_2} \right) U_k \\ & \times \frac{e_1^\top M_1^{-1} \rho_1 \left(\frac{Z_k - Z_i}{h_1} \right)}{f_Z(Z_i)} K_1 \left(\frac{Z_k - Z_i}{h_1} \right).\end{aligned}$$

Define $\bar{\psi}_0(\zeta_k, \zeta_i, \zeta_j) = \psi_0(\zeta_k, \zeta_i, \zeta_j) - \mathrm{E}_k[\psi_0(\zeta_k, \zeta_i, \zeta_j)]$ and

$$\begin{aligned}\bar{\psi}(\zeta_k, \zeta_i, \zeta_j) =& \frac{1}{6} \{ \bar{\psi}_0(\zeta_i, \zeta_j, \zeta_k) + \bar{\psi}_0(\zeta_i, \zeta_k, \zeta_j) + \bar{\psi}_0(\zeta_j, \zeta_i, \zeta_k) \\ & + \bar{\psi}_0(\zeta_j, \zeta_k, \zeta_i) + \bar{\psi}_0(\zeta_k, \zeta_i, \zeta_j) + \bar{\psi}_0(\zeta_k, \zeta_j, \zeta_i) \},\end{aligned}$$

where E_k denotes expectation with respect to ζ_k only. By construction, $\bar{\psi}(\cdot,\cdot,\cdot)$ is symmetric and $\mathrm{E}_k\bar{\psi}(\zeta_k,\zeta_i,\zeta_j)=0$. Using the above notation, we have

$$\begin{aligned}\mathcal{P}_{n,1}^{(1)} &= \frac{1}{n^3 g h_2^{d+1} h_1^r}\sum_{k\neq i\neq j}\sum_{i=1}^{n}\sum_{j=1}^{n}\psi_0(\zeta_k,\zeta_i,\zeta_j)\\ &= \frac{1}{n^3 g h_2^{d+1} h_1^r}\sum_{k\neq i\neq j}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}_k[\psi_0(\zeta_k,\zeta_i,\zeta_j)] + \frac{6}{n^3 g h_2^{d+1} h_1^r}\sum_{1\le k<i<j\le n}\bar{\psi}(\zeta_k,\zeta_i,\zeta_j)\\ &= \mathcal{P}_{n,1a}^{(1)} + \mathcal{P}_{n,1b}^{(1)},\end{aligned}$$

where

$$\mathcal{P}_{n,1a}^{(1)} = \frac{1}{n^2 g h_2^{d+1} h_1^r}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}_k[\psi_0(\zeta_k,\zeta_i,\zeta_j)]$$

and

$$\mathcal{P}_{n,1b}^{(1)} = \frac{6}{n^3 g h_2^{d+1} h_1^r}\sum_{1\le k<i<j\le n}\bar{\psi}(\zeta_k,\zeta_i,\zeta_j).$$

For the term $\mathcal{P}_{n,1a}^{(1)}$,

$$\begin{aligned}\mathcal{P}_{n,1a}^{(1)} =& \frac{1}{n^2 g h_2^{d+1} h_1^r}\sum_{j=1}^{n}\sum_{i=1}^{n}\frac{1}{f_{X_2U}(x_2,U_i)}\mathcal{L}_l\left(\frac{X_{2j}-x_2}{g},x_2,U_i,X_{1j},Z_{1j}\right)\epsilon_j\\ &\times K_2'^{\top}\left(\frac{U_j-U_i}{h_2}\right)\frac{e_1^{\top}M_1^{-1}}{f_Z(Z_i)}\mathrm{E}_k\left[\rho_1\left(\frac{Z_k-Z_i}{h_1}\right)\mathrm{E}_k\left(U_k|Z_k\right)K_1\left(\frac{Z_k-Z_i}{h_1}\right)\right]\\ =&0\end{aligned}$$

by the law of iterated expectation. By applying Lemma 4.5 (with $\theta=\delta/2$) under the mixing condition for ζ_i in Assumptions 4.A3 and 4.B3, it holds for $k<i<j$ and $\delta=\max\{\delta_1,\delta_2\}$ that

$$\begin{aligned}|\mathrm{E}\bar{\psi}_0(\zeta_i,\zeta_j,\zeta_k)| =&|\mathrm{E}\{\psi_0(\zeta_i,\zeta_j,\zeta_k)-\mathrm{E}_k\psi_0(\zeta_i,\zeta_j,\zeta_k)\}|\\ =&|\mathrm{E}\psi_0(\zeta_i,\zeta_j,\zeta_k)-\mathrm{E}_{ij}\mathrm{E}_k\psi_0(\zeta_i,\zeta_j,\zeta_k)|\\ \le& C(gh_2^d h_1^r)^{2/(2+\delta)}\alpha^{\delta/(2+\delta)}(i-k),\end{aligned}$$

where E_{ij} refers to expectation with respect to both ζ_i and ζ_j. To apply Lemma 4.5

above, we require $\mathrm{E}_{ij}\mathrm{E}_k|\psi_0(\zeta_i,\zeta_j,\zeta_k)|^{2+\delta}$ to be bounded, which is ensured by the compact supports of the bounded kernels K_1, L, and the derivative of K_2 in Assumptions 4.A1 and 4.B1, the existence of $S^{-1}(x_2,u)$ and M^{-1} due to Assumptions 4.A7 and 4.B6, the boundedness of ζ_i and density f_{X_2U} by Assumptions 4.A4 and 4.B4, and the moment conditions for $X_1\epsilon$ and $Z_1\epsilon$ in Assumption 4.A6. Similarly, we can show that $|\mathrm{E}\bar{\psi}_0(\cdot,\cdot,\cdot)| \leq C(gh_2^dh_1^r)^{2/(2+\delta)}\alpha^{\delta/(2+\delta)}(i-k)$ holds for different order of inputs ζ_i, ζ_j, and ζ_k and $\delta = \max\{\delta_1,\delta_2\}$. Consequently, we obtain $|\mathrm{E}\bar{\psi}(\zeta_k,\zeta_i,\zeta_j)| \leq C(gh_2^dh_1^r)^{2/(2+\delta)}\alpha^{\delta/(2+\delta)}(i-k)$ and for $\delta = \max\{\delta_1,\delta_2\}$,

$$
\begin{aligned}
|\mathrm{E}\mathcal{P}_{n,1b}^{(1)}| =& \frac{6}{n^3gh_2^{d+1}h_1^r}\sum_{1\leq k<i<j\leq n}|\mathrm{E}\bar{\psi}(\zeta_k,\zeta_i,\zeta_j)| \\
\leq& Cn^{-3}g^{-1}h_2^{-d-1}h_1^{-r}(gh_2^dh_1^r)^{2/(2+\delta)}\sum_{1\leq k<i<j\leq n}\alpha^{\delta/(2+\delta)}(i-k) \\
=& Cn^{-3}g^{-1}h_2^{-d-1}h_1^{-r}(gh_2^dh_1^r)^{2/(2+\delta)}\sum_{\tau=1}^{n-2}\sum_{l=\tau+2}^{n}(n+1-l)\cdot\alpha^{\delta/(2+\delta)}(\tau) \\
\leq& Cn^{-1}g^{-\delta/(2+\delta)}h_2^{-d\delta/(2+\delta)-1}h_1^{-r\delta/(2+\delta)} \\
&\times\sum_{\tau=1}^{n-2}(1-\tau/n)(1-\tau/n-1/n)\alpha^{\delta/(2+\delta)}(\tau) \\
=& \mathrm{O}(n^{-1}g^{-\delta/(2+\delta)}h_2^{-d\delta/(2+\delta)-1}h_1^{-r\delta/(2+\delta)}) \\
=& \mathrm{o}(n^{-1/2}g^{-1/2}). \qquad \text{(by Assumption 4.B7(ii))}
\end{aligned}
$$

According to Lemma 4.7 (with $\theta = \delta/2$) under Assumptions 4.A1–4.A7 and 4.B1–4.B6,

$$
\begin{aligned}
\mathrm{E}(\mathcal{P}_{n,1b}^{(1)})^2 &\leq Cn^{-3}g^{-2}h_1^{-2r}h_2^{-2d-2}(gh_1^rh_2^d)^{2/(2+\delta)} \\
&= \mathrm{O}(n^{-3}g^{-2(1+\delta)/(2+\delta)}h_1^{-2r(1+\delta)/(2+\delta)}h_2^{-2d(1+\delta)/(2+\delta)-2}) \\
&= \mathrm{o}(n^{-1}g^{-1}). \qquad \text{(by Assumption 4.B7(ii))}
\end{aligned}
$$

Therefore, $\mathcal{P}_{n,1}^{(1)} = \mathrm{o}_\mathrm{p}(n^{-1/2}g^{-1/2})$ by Chebyshev's inequality. One can also show that $\mathcal{P}_{n,1}^{(2)} = \mathrm{o}_\mathrm{p}(n^{-1/2}g^{-1/2})$ in a similar way. This completes the proof for the result $\mathcal{P}_1(x_2) = \mathrm{o}_\mathrm{p}(n^{-1/2}g^{-1/2})$. □

Lemma 4.18. *Under Assumptions 4.A1–4.A8 and 4.B1–4.B7, for $n \to \infty$ and $x_2 \in D_{X_2}$,*

$$\mathcal{R}_1(x_2) + \mathcal{R}_2(x_2) + \mathcal{R}_3(x_2) = \mathrm{o_p}(n^{-1/2}g^{-1/2}).$$

Proof. Similarly to the proof of Lemma 4.12, by the invertibility of f_{X_2U} and $S(x_2, u)$ in Assumptions 4.A4 and 4.A7 and the convergence results for $\hat{S}_n(x_2, u)$, $\hat{S}_n^{-1}(x_2, u)$, and $\mathcal{P}_c(x_2)$ in Lemmas 4.15, 4.16, and 4.17, respectively, for $c = 1, 2, 3$, we can write

$$\begin{aligned}
\mathcal{R}_c(x_2) =& \Phi\left\{\hat{S}_n^{-1}(x_2, \hat{U}_i)\left[\hat{S}_n(x_2, \hat{U}_i) - f_{X_2U}(x_2, U_i)S(x_2, U_i)\right]\frac{S^{-1}(x_2, U_i)}{f_{X_2U}(x_2, U_i)},\right. \\
& \left.\hat{T}_{n,c}(x_2, \hat{U}_i) - \tilde{T}_{n,c}(x_2, U_i)\right\} \\
=& \mathrm{O_p}(1)\mathrm{o_p}(1)\underbrace{\Phi\left\{\frac{S^{-1}(x_2, U_i)}{f_{X_2U}(x_2, U_i)}, \hat{T}_{n,c}(x_2, \hat{U}_i) - \tilde{T}_{n,c}(x_2, U_i)\right\}}_{=\mathcal{P}_c(x_2)} \\
=& \mathrm{O_p}(1)\mathrm{o_p}(1)\mathrm{o_p}(n^{-1/2}g^{-1/2}) = \mathrm{o_p}(n^{-1/2}g^{-1/2}).
\end{aligned}$$

□

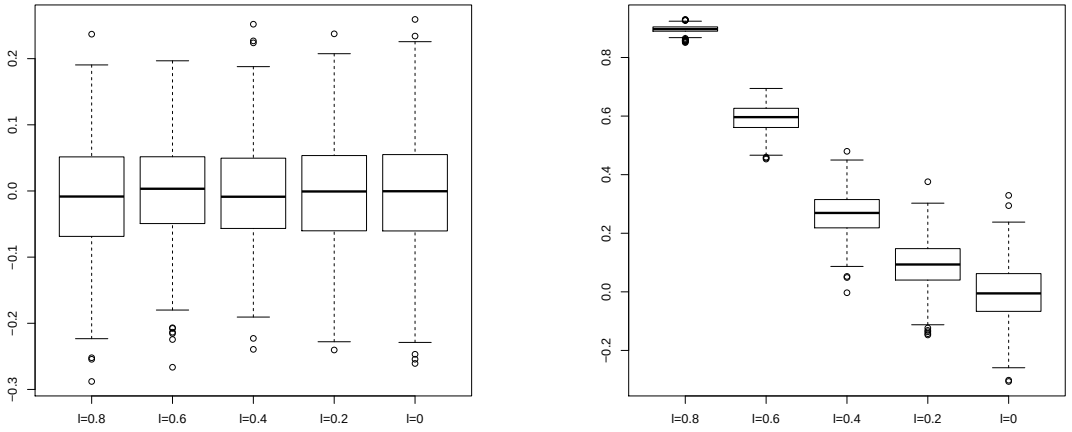
4.9 Appendix: Example 2: weak instruments

In this example, we investigate how the performance of our proposed estimator changes for weak instruments. We consider the same weakly dependent process as in Example 2 in Section 4.5.2, except the generating process for X_{2t} is replaced by

$$X_{2t} = 1.25l \cdot \{Z_{2t} + \sin(0.2 \cdot Z_{2t})\} + 5(1 - l)U_{2t},$$

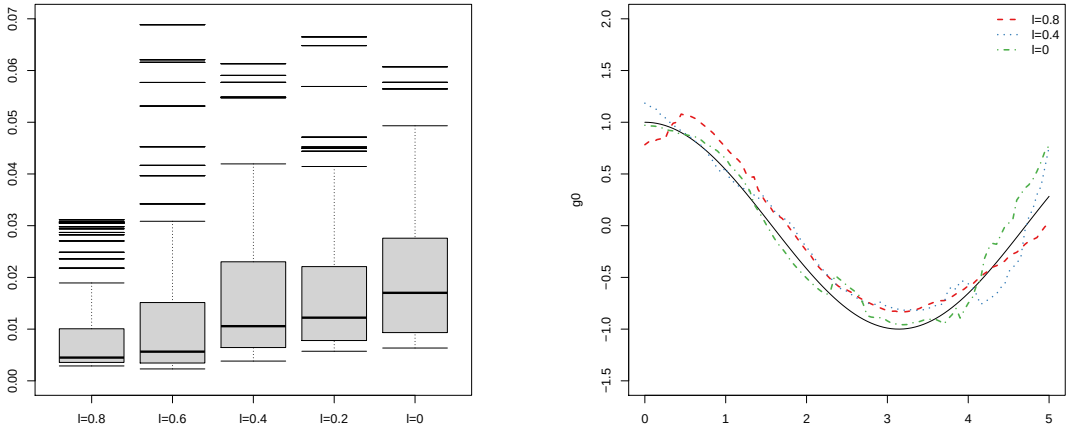
for $l = \{0, 0.2, 0.4, 0.6, 0.8\}$. Figure 4.6 demonstrates how the correlations between endogenous variable and exogenous variables change for different values of l. The correlations between X_2 and Z_1 across 500 replications more or less remain in the range of $(-0.2, 0.2)$ for different l's, but the median correlation between X_2 and Z_2 decreases from 0.9 to 0 as l becomes smaller. In other words, Z_2 is a weaker instrument for smaller l.

Simulation results are summarized in Figures 4.7(a) and (b). Figure 4.7(b) displays the plots of our proposed estimates for the coefficient function g_0 from a sample such that



(A) (B)

FIGURE 4.6: Correlations for $l = \{0, 0.2, 0.4, 0.6, 0.8\}$. Figures (a) and (b) give the boxplots of the correlations between endogenous variable X_2 and exogenous variables Z_1 and Z_2, respectively.



(A) (B)

FIGURE 4.7: Simulation results for Example 4. Figure (a) shows the mean squared errors for different values of l. Figure (b) provides the plots of the estimates for $l = 0, 0.4, 0.8$.

its correlation between X_2 and Z_2 equals to the median in the 500 replications. In figure 4.7(b), the estimates for different values of l are closed. However, the mean squared errors for smaller l is larger in Figure 4.7(a), which indicates our proposed IV estimate perform worse for weaker instruments.

Bibliography

Ahmad, I., S. Leelahanon, and Q. Li (2005). Efficient estimation of a semiparametric partially linear varying coefficient model. *The Annals of Statistics 33*, 258–283.

Ai, C. and X. Chen (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica 71*(6), 1795–1843.

Andrews, D. W. (1992). Generic uniform convergence. *Econometric Theory 8*(2), 241–257.

Arcones, M. A. and B. Yu (1994). Central limit theorems for empirical and u-processes of stationary mixing sequences. *Journal of Theoretical Probability 7*, 47–72.

Areosa, W. D., M. McAleer, and M. C. Medeiros (2011). Moment-based estimation of smooth transition regression models with endogenous variables. *Journal of Econometrics 165*, 100–111.

Bai, J. and P. Perron (1998). Estimating and testing linear models with multiple structural changes. *Econometrica 66*(1), 47–78.

Bai, J. and P. Perron (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics 18*(1), 1–22.

Bierens, H. J. (1982). Consistent model specification tests. *Journal of Econometrics 20*(1), 105 – 134.

Blundell, R., X. Chen, and D. Kristensen (2007). Semi-nonparametric IV estimation of shape-invariant engel curves. *Econometrica 75*(6), 1613–1669.

Boldea, O. and A. Hall (2013). Estimation and inference in unstable nonlinear least squares models. *Journal of Econometrics 172*, 158–167.

Cai, Z., M. Das, H. Xiong, and X. Wu (2006). Functional coefficient instrumental variables models. *Journal of Econometrics 133*(1), 207–241.

Cai, Z., J. Fan, and R. Li (2000). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association 95*(451), 888–902.

Cai, Z., J. Fan, and Q. Yao (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association 95*(451), 941–956.

Cai, Z., Y. Fang, M. Lin, and J. Su (2017). Inferences for a partially varying coefficient model with endogenous regressors. *Journal of Business & Economic Statistics 0*(0), 1–13.

Cai, Z. and Q. Li (2008). Nonparametric estimation of varying coefficient dynamic panel data models. *Econometric Theory 24*(5), 1321–1342.

Cai, Z. and H. Xiong (2012). Partially varying coefficient instrumental variables models. *Statistica Neerlandica 66*(2), 85–110.

Caner, M. and B. E. Hansen (2004). Instrumental variable estimation of a threshold model. *Econometric Theory 20*(5), 813–843.

Card, D. (1993). Using Geographic Variation in College Proximity to Estimate the Return to Schooling. Working Papers 696, Princeton University, Department of Economics, Industrial Relations Section.

Casas, I. and I. Gijbels (2012). Unstable volatility: the break-preserving local linear estimator. *Journal of Nonparametric Statistics 24*(4), 883–904.

Chan, K. S. and H. Tong (1986). On estimating thresholds in autoregressive models. *Journal of Time Series Analysis 7*, 179–190.

Chen, B. and Y. Hong (2012). Testing for smooth structural changes in time series models via nonparametric regression. *Econometrica 80*(3), 1157–1183.

Chen, R. and R. S. Tsay (1993). Functional-coefficient autoregressive models. *Journal of the American Statistical Association 88*(421), 298–308.

Čížek, P. and C. Koo (2017a). Jump-preserving functional-coefficient models for nonlinear time series. CentER Discussion Paper 2017-017, Tilburg University.

Čížek, P. and C. Koo (2017b). Semiparametric transition models. Unpublished manuscript, Tilburg University.

Clements, M. P. and H.-M. Krolzig (1998). A comparison of the forecast performance of markov-switching and threshold autoregressive models of us gnp. *Econometrics Journal 1*, C47–C75.

Das, M. (2005). Instrumental variables estimators of nonparametric models with discrete endogenous regressors. *Journal of Econometrics 124*(2), 335–361.

Davidson, J. (1994). *Stochastic Limit Theory*. Oxford University Press.

Dedecker, J. and S. Louhichi (2002). Maximal inequalities and empirical central limit theorems. In T. M. H. Dehling and M. Sorensen (Eds.), *Empirical Process Techniques for Dependent Data*, pp. 137–160. Basel: Birhäuser.

Fan, J. and T. Huang (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli 11*(6), 1031–1057.

Fan, J. and Q. Yao (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika 85*(3), pp. 645–660.

Fan, J. and Q. Yao (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer Series in Statistics. New York: Springer-Verlag.

Fan, J., Q. Yao, and Z. Cai (2003). Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society Series B 65*(1), 57–80.

Fan, J. and J. Zhang (2000). Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society Series B 62*(2), 303–322.

Fan, J. and W. Zhang (1999). Statistical estimation in varying-coefficient models. *The Annals of Statistics 27*, 1491–1518.

Fan, J. and W. Zhang (2008). Statistical methods with varying coefficient models. *Statistics and Its Interface 1*, 179–195.

Gao, J. and M. King (2004). Adaptive testing in continuous-time diffusion models. *Econometric Theory 20*(5), 844–882.

Gijbels, I. and A.-C. Goderniaux (2004). Bandwidth selection for change point estimation in nonparametric regression. *Technometrics 46*, 76–86.

Gijbels, I., A. Lambert, and P. Qiu (2007). Jump-preserving regression and smoothing using local linear fitting: A compromise. *Annals of the Institute of Statistical Mathematics 59*(2), 235–272.

Godtliebsen, F., E. Spjotvoll, and J. S. Marron (1997). A nonlinear gaussian filter applied to images with discontinuities. *Journal of Nonparametric Statistics 8*(1), 21–43.

Hall, P. and J. L. Horowitz (2005). Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics 33*(6), 2904–2929.

Han, S. (2014). Nonparametric estimation of triangular simultaneous equations models under weak identification. Department of Economics Working Papers 140414, The University of Texas at Austin, Department of Economics.

Hansen, B. E. (2000). Sample splitting and threshold estimation. *Econometrica 68*, 575–603.

Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory 24*(3), 726–748.

Hansen, B. E. (2011). Threshold autoregression in economics. *Statistics and Its Interface 4*, 123–127.

Hastie, T. J. and R. J. Tibshirani (1993). Varying-coefficient models. *Journal of the Royal Statistical Society Series B 55*, 757–796.

Hoover, D. R., J. A. Rice, C. O. Wu, and L.-P. Yang (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika 85*, 809–822.

Huang, J. Z. and H. Shen (2004). Functional coefficient regression models for non-linear time series: A polynomial spline approach. *Scandinavian Journal of Statistics 31*(4), 515–534.

Huang, J. Z., C. O. Wu, and L. Zhou (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika 89*, 111–128.

Huang, J. Z., C. O. Wu, and L. Zhou (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistics Sinica 14*, 763–788.

Hubrich, K. and T. Teräsvirta (2013). Thresholds and smooth transitions in vector autoregressive models. In T. B. Fomby, L. Kilian, and A. Murphy (Eds.), *VAR Models in Macroeconomics – New Developments and Applications: Essays in Honor of Christopher A. Sims*, Volume 32 of *Advances in Econometrics*, pp. 273–326. Emerald Group Publishing Limited.

Hurvich, C. M., J. S. Simonoff, and C.-L. Tsai (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society B 60*(2), 271–293.

Ichimura, H. and S. Lee (2010). Characterization of the asymptotic distribution of semiparametric m-estimators. *Journal of Econometrics 159*, 252–266.

Kang, K.-H., J.-Y. Koo, and C.-W. Park (2000). Kernel estimation of discontinuous regression functions. *Statistics & Probability Letters 47*, 277–285.

Lee, D. S. and T. Lemieux (2010). Regression discontinuity designs in economics. *Journal of Economic Literature 48*, 281–355.

Lee, E. R. and E. Mammen (2016). Local linear smoothing for sparse high dimensional varying coefficient models. *Electronic Journal of Statistics 10*(1), 855–894.

Leybourne, S., P. Newbold, and D. Vougas (1998). Unit roots and smooth transitions. *Journal of Time Series Analysis 19*, 83–97.

Lin, C.-F. J. and T. Teräsvirta (1994). Testing the constancy of regression parameters against continuous structural change. *Journal of Econometrics 62*(2), 211–228.

Lundbergh, S., T. Teräsvirta, and D. van Dijk (2003). Time-varying smooth transition autoregressive models. *Journal of Business & Economic Statistics 21*(1), 104–21.

Masry, E. (1996). Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *Journal of Time Series Analysis 17*(6), 571–599.

Medeiros, M. C. and A. Veiga (2003). Diagnostic checking in a flexible nonlinear time series model. *Journal of Time Series Analysis 24*(4), 461–482.

Medeiros, M. C. and A. Veiga (2005). A flexible coefficient smooth transition time series model. *IEEE Transactions on Neural Networks 16*, 97–113.

Meitz, M. and P. Saikkonen (2010). A note on the geometric ergodicity of a nonlinear ar-arch model. *Statistics & Probability Letters 80*(7-8), 631–638.

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

Newey, W. K. and J. L. Powell (2003). Instrumental variable estimation of nonparametric models. *Econometrica 71*(5), 1565–1578.

Newey, W. K., J. L. Powell, and F. Vella (1999). Nonparametric estimation of triangular simultaneous equations models. *Econometrica 67*(3), 565–603.

Park, B. U., E. Mammen, Y. K. Lee, and E. R. Lee (2015). Varying coefficient regression models: A review and new developments. *International Statistical Review 83*(1), 36–64.

Park, S. (2003). Semiparametric instrumental variables estimation. *Journal of Econometrics 112*(2), 381–399.

Patton, A., D. N. Politis, and H. White (2009). Correction to "automatic block-length selection for the dependent bootstrap". *Econometric Reviews 28*, 372–375.

Polzehl, J. and V. G. Spokoiny (2000). Adaptive weights smoothing with applications to image restoration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*(2), 335–354.

Porter, J. and P. Yu (2015). Regression discontinuity designs with unknown discontinuity points: Testing and estimation. *Journal of Econometrics 189*(1), 132–147.

Potter, S. M. (1995). A nonlinear approach to us gnp. *Journal of Applied Econometrics 10*(2), 109–125.

Rothman, P. (1998). Forecasting asymmetric unemployment rates. *Review of Economics and Statistics 80*, 164–168.

Sarantis, N. (1999). Modeling non-linearities in real effective exchange rates. *Journal of International Money and Finance 18*(1), 27–45.

Skalin, J. and T. Teräsvirta (2002). Modeling asymmetries and moving equilibria in unemployment rates. *Macroeconomic Dynamics 6*(2), 202–241.

Spokoiny, V. (1998). Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice. *Ann. Stat. 26*(4), 1356–1378.

Su, L. and T. Hoshino (2015). Sieve instrumental variable quantile regression estimation of functional coefficient models. Working paper, Singapore Management University, School of Economics.

Su, L., I. Murtazashvili, and A. Ullah (2013). Local linear GMM estimation of functional coefficient iv models with an application to estimating the rate of return to schooling. *Journal of Business & Economic Statistics 31*(2), 184–207.

Su, L. and A. Ullah (2008). Local polynomial estimation of nonparametric simultaneous equations models. *Journal of Econometrics 144*(1), 193–218.

Sun, S. and C.-Y. Chiang (1997). Limiting behavior of the perturbed empirical distribution functions evaluated at U-statistics for strongly mixing sequences of random variables. *Journal of Applied Mathematics and Stochastic Analysis 10*(1), 3–20.

Sun, Y., R. J. Carroll, and D. Li (2009). Semiparametric estimation of fixed-effects panel data varying coefficient models. In Q. Li and J. S. Racine (Eds.), *Nonparametric Econometric Methods*, Volume 25 of *Advances in Econometrics*, pp. 101–129. Emerald Group Publishing Limited.

Taylor, M. P., D. A. Peel, and L. Sarno (2001). Nonlinear mean-reversion in real exchange rates: Toward a solution to the purchasing power parity puzzles. *International Economic Review 42*(4), 1015–42.

Taylor, N., D. van Dijk, P. H. Franses, and A. Lucas (2000). Sets, arbitrage activity, and stock price dynamics. *Journal of Banking & Finance 24*(8), 1289–1306.

Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association 89*, 208–218.

Teräsvirta, T. and H. M. Anderson (1992). Characterizing nonlinearities in business cycles using smooth transition autoregressive models. *Journal of Applied Econometrics 7*(S), S119–36.

Tiao, G. C. and R. S. Tsay (1994). Some advances in non-linear and adaptive modelling in time-series. *Journal of Forecasting 13*, 109–131.

Tong, H. (1983). *Threshold Models in Non-Linear Time Series Analysis: Lecture Notes in Statistics*. Berlin: Springer.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes*. New York: Springer-Verlag.

van Dijk, D. and P. H. Franses (1999). Modeling multiple regimes in the business cycle. *Macroeconomic Dynamics 3*, 311–340.

van Dijk, D., T. Teräsvirta, and P. H. Franses (2002). Smooth transition autoregressive models – a survey of recent developments. *Econometric Reviews 21*, 1–47.

Wu, C. O., C. T. Chiang, and D. R. Hoover (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American Statistical Association 93*, 1388–1402.

Wu, J. S. and C. K. Chu (1993). Kernel-type estimators of jump points and values of a regression function. *The Annals of Statistics 21*, 1545–1566.

Xiao, Z., O. B. Linton, R. J. Carroll, and E. Mammen (2003). More efficient local polynomial estimation in nonparametric regression with autocorrelated errors. *Journal of the American Statistical Association 98*(464), 980–992.

Xue, L. and L. Yang (2006). Estimation of semi-parametric additive coefficient model. *Journal of Statistical Planning and Inference 136*(8), 2506–2534.

Yoshihara, K.-i. (1976). Limiting behavior of U-statistics for stationary, absolutely regular processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 35*(3), 237–252.

Zhang, F. (2005). *The Schur Complement and Its Applications*, Volume 4 of *Numerical Methods and Algorithms*. Springer.

Zhang, W., S.-Y. Lee, and X. Song (2002). Local polynomial fitting in semivarying coefficient model. *Journal of Multivariate Analysis 82*(1), 166–188.

Zhao, Y.-Y., J.-G. Lin, X.-F. Huang, and H.-X. Wang (2016). Adaptive jump-preserving estimates in varying-coefficient models. *Journal of Multivariate Analysis 149*, 65–80.

Zhao, Y.-Y., J.-G. Lin, H.-X. Wang, and X.-F. Huang (2017, Sep). Jump-detection-based estimation in time-varying coefficient models and empirical applications. *TEST 26*(3), 574–599.

Zhu, H., J. Fan, and L. Kong (2014). Spatially varying coefficient model for neuroimaging data with jump discontinuities. *Journal of the American Statistical Association 109*, 1084–1098.