A Varying-Coefficient Panel Data Model with Fixed Effects: Theory and an Application to U.S. Commercial Banks

Guohua Feng[§], Jiti Gao[⋆], Bin Peng[†] and Xiaohui Zhang[‡]

[§]University of North Texas, [⋆]Monash University, [†]University of Technology Sydney and [‡]Murdoch University

Abstract

In this paper, we propose a panel data semiparametric varying-coefficient model in which covariates (variables affecting the coefficients) are purely categorical. This model has two features: first, fixed effects are included to allow for correlation between individual unobserved heterogeneity and the regressors; second, it allows for cross-sectional dependence through a general spatial error dependence structure. We derive a semiparametric estimator for our model by using a modified within transformation, and then show the asymptotic and finite properties for this estimator under large N and T. Our Monte Carlo study suggests that our methodology works well for both large N and T, and large N and small T cases. Finally, we illustrate our model by analyzing the effects of state-level banking regulations on the returns to scale of commercial banks in the U.S.. Our empirical results suggest that returns to scale is higher in more regulated states than in less regulated states.

Keywords: Categorical variable; estimation theory; nonlinear panel data model; returns to scale.

JEL classification: C23, C51, D24, G21

- Guohua Feng, Department of Economics, University of North Texas, Denton, TX 76201, U.S.A. Email: guohua.feng@unt.edu.
- Jiti Gao, Department of Econometrics and Business Statistics, Monash University, VIC 3145, Australia. Email: Jiti.Gao@monash.edu. Jiti Gao acknowledges the Australian Research Council Discovery Grants Program support under Grant numbers: DP130104229 and DP150101012.
- Bin Peng, Economics Discipline Group, University of Technology Sydney, Ultimo, NSW 2007, Australia. Email: Bin.Peng@uts.edu.au.
- Xiaohui Zhang, School of Management and Governance, Murdoch University, Perth 6150, Australia. Email: Xiaohui.Zhang@murdoch.edu.au.

1 Introduction

Varying-coefficient models have attracted considerable attention in the past two decades. This is particularly true for both cross-sectional and time series varying-coefficient models. For instance, Li et al. (2002) propose a semiparametric varying-coefficient model in a cross-sectional setting, where covariates (i.e., variables affecting the coefficients) are assumed to be continuous in nature. Li and Racine (2010) extend Li et al. (2002) to a more general set-up, which admits both quantitative and qualitative covariates. More recently, Li et al. (2013) extend the cross-sectional varying-coefficient model literature further by proposing a semiparametric varying-coefficient with purely categorical covariates. Similarly, considerable work has also been done on time series varying-coefficient models. For example, Gao and Phillips (2013a) investigate the varying-coefficient model by allowing for the existence of nonstationarity. More references along this latter line can be found in Cai (2007) and Cai et al. (2009).

However, less progress has been made with panel data varying-coefficient models, primarily because of the difficulty involved in dealing with fixed effects. For example, Cai and Li (2008) propose a varying-coefficient dynamic panel data model, where they get around this difficulty by dropping fixed effects. Sun et al. (2009) propose a panel data varying-coefficient model, where they overcome the difficulty associated with fixed effects by imposing a widely-used identification restriction such that the sum of the fixed effects is zero (c.f. Su and Ullah (2011) and Chen et al. (2013)). Rodriguez-Poo and Soberon (2014) propose to use the first difference to remove the fixed effects by allowing N to increase to ∞ with fixed T. It is worth noting that in both of the latter two studies, covariates are assumed to be purely continuous and asymptotic theories are established accordingly.

The purpose of this paper is to contribute to this literature by extending Li et al. (2013)'s cross-sectional varying-coefficient model to a panel data context. To allow for unobserved individual heterogeneity, fixed effects are included in our model. As is well known, the inclusion of fixed effects has the advantage of allowing unobserved individual heterogeneity to be arbitrarily correlated with any other variables. With regards to the nature of the covariates, we follow Li et al. (2013) and only consider the case where all covariates are categorical. To remove fixed effects, we take advantage of the categorical nature of our covariates and implement a modified within transformation. The demeaned model can then be estimated using Li et al. (2013)'s semiparametric kernel estimation method. In addition, we establish the asymptotic properties of our estimator. It is worth noting that our asymptotic properties is established under large N and T, because it is much more challenging to establish asymptotic properties under $(N,T) \to (\infty,\infty)$ for panel data models. We further show in Section 2.4 that our modified within transformation is also valid for the case where T is fixed.

Another feature of our model is that it allows for cross-sectional dependence, an important

issue that has received considerable attention in the recent panel data literature (c.f. Andrews (2005), Pesaran (2006) and Bai (2009)). There are two well-known approaches to modeling cross-sectional dependence. The first approach, due to Pesaran (2006) and Bai (2009), is to use a factor structure to capture strong correlation between individuals. The second approach is to use a spatial error structure to model weak correlation between individuals. Excellent works adopting the second approach include, but are not limited to, Pesaran and Tosetti (2011), Chen et al. (2012a) and Chen et al. (2012b). In this paper, we adopt the second approach. Specifically, as shown in Assumption A.4 in Appendix A, we impose a general spatial correlation structure to link the cross-sectional dependence and stationary mixing condition together. The use of this structure enables our model to capture the type of cross-sectional dependence discussed by Chen et al. (2012b) and Dong et al. (2015).

We apply our panel data categorical varying-coefficient model by analyzing the effects of branch banking regimes on the returns to scale of commercial banks in the U.S. over the period 1986-2005. Until the middle of the 1970's banking in the U.S. was heavily regulated at the state level: in some states banks were prohibited from branching at all (unit banking regime), in some states they were restricted to branch within a portion of the state (limited branching banking regime), and in other states they were permitted to branch statewide (statewide branching banking regime). In the mid-1980s individual states began to remove restrictions on intrastate branching. This deregulation process culminated in the passage of the Riegle-Neal Interstate Banking and Branching Efficiency Act of 1994, which permitted nationwide branching as of June 1997 (nationwide branching banking regime). Since banking regime is an important factor in determining production technology, we use it as a categorical argument (covariate) of the varying coefficient. Specifically, we consider a categorical varying-coefficient translog cost function. Our results show that returns to scale is higher in more regulated states than in less regulated states. Our results also indicate that the majority of the banks face increasing returns to scale, a small percentage face decreasing returns to scale, and an even smaller percentage face constant returns to scale. This finding is potentially important as increasing returns to scale is often used to justify bank mergers and in policy debates on regulations limiting the size of banks.

The rest of this paper is organized as follows. Section 2 presents the panel data varying-coefficient model and derives the estimator of the model and the associated asymptotic results: (1) Sections 2.1 and 2.2 consider the relevant and irrelevant covariate cases, respectively; (2) then, based on these results, in Section 2.3 we propose a variable selection procedure to identify significant elements from regressors; (3) finally, Section 2.4 discusses some extensions. In Section 3, we conduct a Monte Carlo study investigating the finite sample properties of our methodology. Section 4 presents the application of our model and methodology to the U.S. commercial bank data. Section 5 concludes. Note that the assumptions and pertinent discussions needed for

deriving the asymptotic results are given in Appendix A at the end of this paper, while the proofs are provided in Appendix B in the supplementary document of this paper.

Before proceeding to Section 2, it is convenient to introduce some notations that will be used throughout this paper. $1(A)$ denotes an indicator function, i.e. $1(A) = 1$ if A is true, otherwise $1(A) = 0$; $\|\cdot\|$ denotes the Frobenius norm; $\to_P$ denotes converging in probability; $\to_D$ denotes converging in distribution.

2 Model Specification

We consider the following panel data model.

$$Y_{it} = X_{it}'\beta(Z_{it}) + w_i + u_{it}, \quad i = 1, \ldots, N \text{ and } t = 1, \ldots, T, \tag{2.1}$$

where u_{it} is a random error term; $X_{it} = (X_{it,1}, \ldots, X_{it,q})'$ is a q-dimensional vector of regressors; $\beta(\cdot)$ is a q-dimensional vector of unknown coefficient function; $Z_{it} = (Z_{it,1}, \ldots, Z_{it,r})'$ is an r-dimensional vector of discrete covariates; w_i is a fixed effect and can be arbitrarily correlated with any other variables. To distinguish between X_{it} and Z_{it}, they are respectively referred to as regressors and covariates hereafter. For an r-dimensional vector z, we use z_s to denote the s^{th} component of z, and assume that z_s takes c_s different values in $\{0, 1, \ldots, c_s - 1\}$ and $2 \leq c_s < \infty$ for $s = 1, \ldots, r$. When showing the asymptotic properties of our model and estimator below, we follow Li et al. (2013) and distinguish between the case where $\beta(z)$ is not a constant function with respect to z_s for $s = 1, 2, \ldots, r$, and the case where some elements of z_s do not have impacts on $\beta(\cdot)$ and are independent of all other variables. The former case is referred to as "relevant covariate case" and will be discussed in details in Section 2.1, while the latter one is referred to as "irrelevant covariate case" and will be discussed in details in Section 2.2.

The model (2.1) extends the cross-sectional varying-coefficient model of Li et al. (2013) to a panel data setting. As in Li et al. (2013), we focus on the case where Z_{it} is purely categorical. Therefore, we also adopt the kernel function of Aitchison and Aitken (1976) for unordered covariate below:

$$l(Z_{it,s}, z_s, \lambda_s) = \begin{cases} 1, & \text{if } Z_{it,s} = z_s \\ \lambda_s, & \text{otherwise} \end{cases}, \tag{2.2}$$

where the range of λ_s is $[0, 1]$ for $s = 1, \ldots, r$. It is easy to see that $\lambda_s = 0$ leads to an indicator function and $\lambda_s = 1$ gives a uniform weight function. Note that (2.2) allows one to extend the kernel density estimation technique to multivariate discrete spaces. With (2.2), we can construct a product kernel function as follows.

$$L(Z_{it}, z, \lambda) = \prod_{s=1}^{r} l(Z_{it,s}, z_s, \lambda_s) = \prod_{s=1}^{r} \lambda_s^{1(Z_{it,s} \neq z_s)}, \tag{2.3}$$

where $\lambda = (\lambda_1, \ldots, \lambda_r)'$.

We now discuss how to deal with the fixed effects in (2.1) (i.e., w_i) before proceeding further. To remove the impacts of fixed effects, some studies assume that $\sum_{i=1}^{N} w_i = 0$ (c.f. Sun et al. (2009), Su and Ullah (2011) and Chen et al. (2013)); some studies propose to take the first difference (c.f. Rodriguez-Poo and Soberon (2014)); and others assume that w_i has mean 0 and is uncorrelated with any other variables (c.f. Blundell and Bond (1998)). In this paper, we take a different approach by implementing a within transformation to remove the fixed effects.[2] However, we cannot follow the common practice of subtracting the simple average across t from both sides of (2.1), because $\beta(Z_{it})$ varies over t. To overcome this problem, we implement a modified within transformation that involves the use of the kernel function in (2.3). Our modified within transformation is very effective in that it enables us to deal with the fixed effects for both the case where both N and T are large and the case where N is large and T is small. Due to space limitations, we focus on the former case in what follows. For the latter case, it is easy to show that the estimator and associated asymptotic properties derived for the former case remain valid, by making some minor modifications to the proof for the former case.

Specifically, let $L_{js,it} = L(Z_{js}, Z_{it}, \lambda)$ for $1 \leq i, j \leq N$ and $1 \leq t, s \leq T$ and let $T_{it} = \sum_{s=1}^{T} L_{is,it}^p$, where $p \geq 2$ is a finite positive integer and chosen arbitrarily. In practice, the choice of $p = 2$ is enough. Let $\tilde{Y}_{it} = Y_{it} - \frac{1}{T_{it}} \sum_{s=1}^{T} Y_{is} L_{is,it}^p$, and $\tilde{X}_{it}$ and $\tilde{u}_{it}$ are defined in the same fashion. With these notations, our modified within transformation[3] can be written as

$$\begin{aligned}
\tilde{Y}_{it} &= X_{it}'\beta(Z_{it}) + w_i + u_{it} - \frac{1}{T_{it}} \sum_{s=1}^{T} \left(X_{is}'\beta(Z_{is}) + w_i + u_{is} \right) L_{is,it}^p \\
&= X_{it}'\beta(Z_{it}) - \frac{1}{T_{it}} \sum_{s=1}^{T} X_{is}' L_{is,it}^p \beta(Z_{it}) + \frac{1}{T_{it}} \sum_{s=1}^{T} X_{is}' L_{is,it}^p \beta(Z_{it}) - \frac{1}{T_{it}} \sum_{s=1}^{T} X_{is}' \beta(Z_{is}) L_{is,it}^p + \tilde{u}_{it} \\
&= \tilde{X}_{it}'\beta(Z_{it}) + \gamma_{it} + \tilde{u}_{it},
\end{aligned} \tag{2.4}$$

where $\gamma_{it} = \frac{1}{T_{it}} \sum_{s=1}^{T} X_{is}' \left(\beta(Z_{it}) - \beta(Z_{is}) \right) L_{is,it}^p$. Note that the kernel function (2.3) can also be

[2]The advantages of using within transformation have been well documented in Hsiao (2003).

[3]In an earlier version, we subtracted $\frac{1}{T_{it}} \sum_{s=1}^{T} Y_{it} 1(Z_{is} = Z_{it})$ with $T_{it} = \sum_{s=1}^{T} 1(Z_{is} = Z_{it})$ in the within transformation. However, it is very likely that some T_{it}'s will be zero when T is relatively small compared to the cardinality of the support of Z_{it}. The Associate Editor suggested subtracting $\frac{1}{T_{it}} \sum_{s=1}^{T} Y_{it} L_{js,it}$ with $T_{it} = \sum_{s=1}^{T} L_{is,it}$. Then, for (2.6) below, we would get $\left(\beta(Z_{it}) - \beta(Z_{is}) \right) L_{is,it} = O_P(\|\lambda\|)$ instead, which would affect the rate of convergence developed in Theorem 2.1.1. Motivated by this suggestion, we then consider (2.4). We gratefully thank the Associate Editor for this constructive suggestion.

expressed as

$$\begin{aligned} L(Z_{it}, z, \lambda) &= \prod_{s=1}^{r} \{1(Z_{it,s} = z_s) + \lambda_s 1(Z_{it,s} \neq z_s)\} \\ &= \prod_{s=1}^{r} 1(Z_{it,s} = z_s) + \sum_{s=1}^{r} \lambda_s 1_{s,Z_{it}=z} + \cdots + \prod_{s=1}^{r} \lambda_s 1(Z_{it,s} \neq z_s) \\ &= 1(Z_{it} = z) + \sum_{s=1}^{r} \lambda_s 1_{s,Z_{it}=z} + \cdots + \prod_{s=1}^{r} \lambda_s 1(Z_{it,s} \neq z_s), \end{aligned} \tag{2.5}$$

where $1_{s,Z_{it}=z} = 1(Z_{it,s} \neq z_s) \prod_{n=1,n\neq s}^{r} 1(Z_{it,n} = z_n)$ for simplicity. Due to the fact that $(\beta(Z_{it}) - \beta(Z_{is}))\, 1(Z_{it} = Z_{is}) = 0$, if λ is sufficiently small, then we obtain

$$(\beta(Z_{it}) - \beta(Z_{is}))\, L_{is,it}^{p} = O(\|\lambda\|^{p}) \tag{2.6}$$

uniformly. Hence, the truncation residual γ_{it} is controlled by the bandwidth λ only. In what follows, we will show that the optimal bandwidth selected below is indeed sufficiently small.

Using our modified within transformation in (2.4), we can estimate $\beta(z)$ for $\forall z \in \mathcal{D}$ as follows:

$$\hat{\beta}(z) = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{X}_{it} \tilde{X}_{it}' L(Z_{it}, z, \hat{\lambda}) \right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{X}_{it} \tilde{Y}_{it} L(Z_{it}, z, \hat{\lambda}), \tag{2.7}$$

where $\hat{\lambda}$ is obtained by minimizing the following cross-validation (CV) criterion function

$$CV(\lambda) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\tilde{Y}_{it} - \tilde{X}_{it}' \hat{\beta}_{-it}(Z_{it}) \right)^{2}; \tag{2.8}$$

and $\hat{\beta}_{-it}(Z_{it})$ is the leave-one-out estimator for $\beta(Z_{it})$

$$\hat{\beta}_{-it}(Z_{it}) = \left(\sum_{js, js\neq it} \tilde{X}_{js} \tilde{X}_{js}' L(Z_{js}, Z_{it}, \lambda) \right)^{-1} \sum_{js, js\neq it} \tilde{X}_{js} \tilde{Y}_{js} L(Z_{js}, Z_{it}, \lambda). \tag{2.9}$$

Having shown how to estimate our panel data categorical varying-coefficient model in (2.1), in what follows we will show the asymptotic properties for our estimator. As noted previously, we first discuss the asymptotic results for the relevant covariate case in Section 2.1 and then discuss the asymptotic results for the irrelevant covariate case in Section 2.2. In Section 2.3, we present a variable selection procedure for selecting significant variables from X_{it}, which completes our proofs of the asymptotic properties of our estimator. Due to space limitations, all assumptions needed for the proofs of the lemmas and theorems presented in Sections 2.1-2.3 are provided in Appendix A, while the proofs themselves are provided in Appendix B of the supplementary document of the paper.

2.1 Relevant Covariate Case

We start with the simple case where all elements of Z_{it} are assumed to be relevant. When deriving the asymptotic results for this case, we first show that minimizing the cross-validation criterion function ensures that $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_r)' = o_P(1)$ in Lemma 2.1.1, then use this property to further investigate $CV(\lambda)$ and show that the rate of convergence is $\hat{\lambda} = O_P\left(\frac{1}{NT}\right)$ in Theorem 2.1.1, and then show the asymptotic normality in Theorem 2.1.2 based on the result of Theorem 2.1.1.

Lemma 2.1.1. *Under Assumption A, as (N,T) go to (∞,∞) jointly, $\hat{\lambda} = o_P(1)$.*

This lemma states that $\hat{\lambda}$ converges to 0 as the sample size increases. Then it is reasonable to assume that λ, when deriving Theorem 2.1.1, is sufficiently small and close to $0_{r\times 1}$. Thus, the product kernel function (2.5) can be simplified as follows.

$$L(Z_{js}, Z_{it}, \lambda) = 1_{js,it} + \sum_{m=1}^{r} \lambda_m 1_{m,jsit} + O(\|\lambda\|^2),$$

where $1_{m,jsit} = 1(Z_{js,m} \neq Z_{it,m}) \prod_{n=1,n\neq m}^{r} 1(Z_{js,n} = Z_{it,n})$.

Theorem 2.1.1. *Under Assumption A, as (N,T) go to (∞,∞) jointly, $\hat{\lambda} = O_P\left(\frac{1}{NT}\right)$.*

Theorem 2.1.1 gives the rate of convergence for $\hat{\lambda}$, which is consistent with the rate shown by Li et al. (2013) for the cross-sectional case. This result is useful for establishing the asymptotic normality for $\hat{\beta}(z)$, because it significantly simplifies our proof by allowing us to use the frequency estimator (i.e., let $\lambda = 0_{r\times 1}$ in (2.7)). More details are given in the Appendix B.

Theorem 2.1.2. *Under Assumption A, as (N,T) go to (∞,∞) jointly, for $z \in \mathcal{D}$,*

$$\sqrt{NT}(\hat{\beta}(z) - \beta(z)) \to_D N(0, \Xi_1(z)^{-1}\Xi_0(z)\Xi_1(z)^{-1}),$$

where

$$\Xi_0(z) = \lim_{N,T\to\infty} \frac{1}{NT} \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T} E\left[u_{it}u_{js}(X_{it} - \mu_X(z))(X_{js} - \mu_X(z))' 1(Z_{it} = z)1(Z_{js} = z)\right],$$

$$\Xi_1(z) = p(z)\left(\Sigma_X(z) - \mu_X(z)\mu_X(z)'\right), \quad p(z) = \Pr(Z_{it} = z), \quad \Sigma_X(z) = E[X_{it}X_{it}'|Z_{it} = z],$$

$$\mu_X(z) = E[X_{it}|Z_{it} = z].$$

We now discuss how to conduct the hypothesis test based on Theorem 2.1.2. By (5) of Lemma B.2, it is easy to know

$$\hat{\Xi}_1(z) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{X}_{it}\tilde{X}_{it}' 1(Z_{it} = z) \to_P \Xi_1(z). \tag{2.10}$$

To consistently estimate $\Xi_0(z)$, we need to impose a stronger restriction, i.e. u_{it} is i.i.d. over i and t. This restriction is in line with the spirit of Corollary 3.1.ii and Theorem 3.3 of Gao and Phillips (2013*b*). Relevant discussions can also be found in Section 2.2.2 of Fan and Yao (2003). With this restriction, $\Xi_0(z)$ reduces to $\Xi_0(z) = p(z)\sigma_u^2\left(\Sigma_X(z) - \mu_X(z)\mu_X(z)'\right) = \sigma_u^2\Xi_1(z)$, so all we need is a consistent estimator for σ_u^2. For this purpose, we intuitively define

$$\hat{\sigma}_u^2 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{Y}_{it} - \tilde{X}_{it}'\hat{\beta}(Z_{it}))^2. \tag{2.11}$$

Then the next result follows immediately.

Corollary 2.1.1. *Under Assumption A, suppose further that u_{it} is i.i.d. over i and t. As (N,T) go to (∞,∞) jointly, for $z \in \mathcal{D}$,*

$$\sqrt{NT}\left(\hat{\sigma}_u^{-2}\hat{\Xi}_1(z)\right)^{1/2}(\hat{\beta}(z) - \beta(z)) \to_D N(0, I_q),$$

where $\hat{\sigma}_u^2$ and $\hat{\Xi}_1(z)$ are defined in (2.11) and (2.10) respectively.

It is worth noting that Corollary 2.1.1 can be used for testing if all variables in X_{it} are significant, when $\beta(z)$ is set to a vector of zeros. We note that the assumption on u_{it} (i.e., i.i.d. over i and t) is restrictive for situations where cross-dependence among u_{it}'s is present. In such situations, the variable selection procedure proposed in Section 2.3 can be used instead.

2.2 Irrelevant Covariate Case

In this subsection, we consider the case where some of the covariates are irrelevant in the sense that they are independent of all other variables in the model. Without losing generality, suppose the first r_1 $(1 \leq r_1 < r)$ elements of Z_{it} are relevant while the remaining $r_2 = r - r_1$ elements of Z_{it} are irrelevant. For notational simplicity, let $\bar{Z}_{it} = (Z_{it,1}, \ldots, Z_{it,r_1})'$ denote the r_1 relevant elements and let $\tilde{Z}_{it} = (Z_{it,r_1+1}, \ldots, Z_{it,r})'$ be the r_2 irrelevant elements. Conformably, we partition λ as follows $\lambda = (\bar{\lambda}', \tilde{\lambda}')'$, where $\bar{\lambda} = (\lambda_1, \ldots, \lambda_{r_1})'$ and $\tilde{\lambda} = (\lambda_{r_1+1}, \ldots, \lambda_r)'$. Let $\bar{\mathcal{D}}$ and $\tilde{\mathcal{D}}$ denote the sets that $\bar{\lambda}$ and $\tilde{\lambda}$ belong to respectively (i.e., $\mathcal{D} = \bar{\mathcal{D}} \times \tilde{\mathcal{D}}$).

As in Section 2.1, we start by stating our asymptotic results.

Lemma 2.2.1. *Under Assumptions A.1-A.4 and Assumption B, as (N,T) go to (∞,∞) jointly, $\hat{\lambda}_s = o_P(1)$ for $s = 1, \ldots, r_1$.*

Like Assumption 3 of Li et al. (2013), this lemma ensures that the $CV(\lambda)$ selected smoothing parameters associated with the relevant covariates will converge to 0. Using this lemma, we can further investigate $CV(\lambda)$ and rate of convergence, as follows.

Theorem 2.2.1. *Under Assumptions A.1-A.4 and Assumption B, as* (N,T) *go to* (∞,∞) *jointly,*

1. $\hat{\lambda}_s = O_P\left(\frac{1}{\sqrt{NT}}\right)$ *for* $s = 1,\ldots,r_1$;

2. $\Pr\left(\hat{\lambda}_{r_1+1} = 1,\ldots,\hat{\lambda}_r = 1\right) \geq \rho$ *for some* $\rho \in (0,1)$.

Note that the rate of convergence of $\hat{\lambda}$ for the irrelevant case is much slower compared to that given in Theorem 2.1.1, due to the presence of irrelevant covariates. The second result of Theorem 2.2.1 reveals that the estimates of $\hat{\lambda}_s$ for $s = r_1+1,\ldots,r$ are not always equal to 1. Due to cross-sectional dependence among the error terms and weak correlation between different time periods, the possible value of ρ becomes more complicated compared to that in Li et al. (2013). This theorem can be considered as a variable selection procedure for the covariates, but one cannot always remove all irrelevant covariates.

Theorem 2.2.2. *Under Assumptions A.1-A.4 and Assumption B, as* (N,T) *go to* (∞,∞) *jointly, for* $z \in \mathcal{D}$, $\hat{\beta}(z) - \beta(\bar{z}) = O_P\left(\frac{1}{\sqrt{NT}}\right)$.

Using Theorem 2.2.1, it is straightforward to show Theorem 2.2.2. However, we still cannot obtain the asymptotic distribution for the irrelevant covariate case. To deal with this problem, one can follow Li et al. (2013) and use bootstrapping techniques to obtain finite sample distributions for variables of interest.

2.3 Variable Selection on X_{it}

As is well-known, including spurious regressors can degrade estimation efficiency substantially (Wang and Xia, 2009). Unfortunately, this problem of spurious regressors may also happen to the panel data varying-coefficient model in (2.1). To avoid this potential problem, in this subsection we propose a variable selection procedure to identify significant regressors for the model. Compared to the significance test provided by Corollary 2.1.1, it is worth noting that this procedure does not require the assumption that u_{it} is i.i.d. over i and t.

To begin with, we assume that all detected irrelevant covariates (i.e., those with $\hat{\lambda}_s = 1$) have been removed and that the vector of remaining covariates is still denoted by $Z_{it} = (\bar{Z}'_{it}, \tilde{Z}'_{it})'$ as above (note here that $\tilde{Z}_{it}$ can be an empty vector). The purpose of this assumption is to reduce the total number of distinct realizations of z from our samples $\{Z_{it}, 1 \leq i \leq N, 1 \leq t \leq T\}$, denoted by m in this subsection. Note that m is always observable and converges to the cardinality of the support of Z_{it} in probability with non-degenerate probability imposed on Z_{it} as the sample size is sufficiently large. In addition, we relax the restriction on r_1 by assuming that $1 \leq r_1 \leq r$ with r_1 remaining unknown. This latter assumption ensures that both relevant and irrelevant cases are covered in what follows.

We further assume there exists an unknown set $\mathcal{A} \subseteq \{1, \ldots, q\}$ satisfying that $E|\beta_j(\bar{Z}_{it})|^2 = 0$ if and only if $j \in \mathcal{A}$, where $\beta_j(\bar{Z}_{it})$ denotes the j^{th} element of $\beta(\bar{Z}_{it})$. For notational simplicity, we assume that in the true model, $\mathcal{A} = \{q^* + 1, \ldots, q\}$ for some positive integer $1 \leq q^* \leq q$. In other words, only the first q^* variables in X_i have nonzero coefficients and our goal is to find this unknown $\mathcal{A}$.

Since m is observable, our parameters of interest can be denoted by an $m \times q$ matrix B. Correspondingly, its underlying, true coefficient function can also be denoted by an $m \times q$ matrix B_0. Formally,

$$\begin{aligned}
\underset{m\times q}{B} &= \{b_{js}\}_{m\times q} = (\beta_1, \ldots, \beta_m)' = (b_1, \ldots, b_q), \\
\underset{q\times 1}{\beta_j} &= (b_{j1}, \ldots, b_{jq})' \text{ for } j = 1, \ldots, m, \\
\underset{m\times 1}{b_s} &= (b_{1s}, \ldots, b_{ms})' \text{ for } s = 1, \ldots, q, \\
\underset{m\times q}{B_0} &= (\beta(\bar{z}^1), \ldots, \beta(\bar{z}^m))' = (b_{01}, \ldots, b_{0q^*}, 0, \ldots, 0), \\
\underset{m\times 1}{b_{0s}} &= (\beta_s(\bar{z}^1), \ldots, \beta_s(\bar{z}^m))' \text{ for } s = 1, \ldots, q^*,
\end{aligned} \tag{2.12}$$

where $\beta_s(\cdot)$ denotes the s^{th} element of $\beta(\cdot)$; $\bar{z}^j$ is an $r_1 \times 1$ vector including the first r_1 elements of z^j; and z^j denotes the j^{th} different realization by observing $\{Z_{it}, 1 \leq i \leq N, 1 \leq t \leq T\}$. It is easy to see that $\beta(\bar{z}^j)$ will reduce to $\beta(z^j)$ when $r_1 = r$. However, r_1 is unknown in general.

Note that the last $q - q^*$ columns of B_0 are zeros implying that B_0 has a group sparsity structure. In other words, entries in each column of B_0 form a group. Then selecting regressors becomes identifying those 0 columns in the matrix B_0. Following the spirit of Yuan and Lin (2006), we consider the following regularized least squares estimator:

$$\hat{B}_\tau = \{\hat{b}_{\tau,js}\}_{m\times q} = (\hat{\beta}_{\tau,1}, \ldots, \hat{\beta}_{\tau,m})' = (\hat{b}_{\tau,1}, \ldots, \hat{b}_{\tau,q}) = \underset{B\in\mathbb{R}^{m\times q}}{\operatorname{argmin}} Q_\tau(B) \tag{2.13}$$

and

$$Q_\tau(B) = \sum_{j=1}^{m}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{Y}_{it} - \tilde{X}_{it}'\beta_j\right)^2 L(Z_{it}, z^j, \hat{\lambda}) + \sum_{s=1}^{q}\tau_s\|b_s\|, \tag{2.14}$$

where $\hat{\lambda}$ is obtained by minimizing (2.8); the term $\sum_{s=1}^{q}\tau_s\|b_s\|$ is the group-wise regularizer and is defined as the weighted sum of the ℓ_2 norms of all the column vectors in B with the weight $\tau = (\tau_1, \ldots, \tau_q)'$ controlling the regularizer.[4]

Under the above setting, we present our first result on variable selection as follows:

[4]In the literature of group LASSO analysis, one usually allows both q and r to diverge to infinity (e.g. Lounici et al. (2011)). However, to our best knowledge, how to select optimal bandwidths for model (2.1) remains an unresolved issue for high dimensional cases. Given that the purpose of this study is to develop a panel data varying-coefficient model for the finite dimension case, we will not discuss the case where both q and r diverge to infinity.

Theorem 2.3.1. *Under Assumptions A.1-A.4, B and C, let* $1 \leq r_1 \leq r$. *As* $(N,T) \to (\infty,\infty)$,

1. *Let* $\tau^* = (\tau_1, \ldots, \tau_{q^*})'$ *and* $\frac{\|\tau^*\|}{\sqrt{NT}} \to \omega_1$, *where* ω_1 *is a constant satisfying that* $0 \leq \omega_1 < \infty$. *Then*

$$\left\| \hat{\beta}_{\tau,j} - \beta(\bar{z}^j) \right\| = O_P\left(\frac{1}{\sqrt{NT}}\right) \quad for \quad j = 1, \ldots, m,$$

   where $\bar{z}^j = (z_1^j, \ldots, z_{r_1}^j)'$.

2. *Let* $\frac{1}{\sqrt{NT}} \min_{s \in \{q^*+1,\ldots,q\}} \tau_s \geq \omega_2$, *where* ω_2 *is sufficiently large. Then*

$$\Pr(\|\hat{b}_{\tau,j}\| = 0) \to 1 \quad for \quad j = q^* + 1, \ldots, q.$$

The first result of Theorem 2.3.1 says if the regularizer weight is not too large, we always have optimal $\sqrt{NT}$ consistency for our estimator. The second result implies that when the regularizer weight is at level $\sqrt{NT}$, we can successfully get rid of those unimportant coefficients in our estimator and select a sub-model of the true model. A natural and simple choice of τ, which satisfies assumptions of both results, is that all the elements of τ are at level $\sqrt{NT}$. With a more careful data-driven choice of τ, we can further achieve the asymptotic normality whenever there is no irrelevant covariate by use of the following oracle[5] property for our estimator (2.13).

Theorem 2.3.2. *Under the conditions of Theorem 2.3.1,*

$$\left\| \hat{\beta}_{\tau,jU} - \hat{\beta}_{ora}(\bar{z}^j) \right\| = O_P\left(\frac{\|\tau^*\|}{NT}\right)$$

for $j = 1, \ldots, m$, *where* $\hat{\beta}_{ora}(\bar{z}^j)$ *is denoted by (2.7) with assuming that the true set* $\mathcal{A}$ *is known;* $\hat{\beta}_{\tau,jU} = (\hat{b}_{\tau,j1}, \ldots, \hat{b}_{\tau,jq^*})'$; $\hat{b}_{\tau,js}$ *for* $j = 1, \ldots, m$ *and* $s = 1, \ldots, q^*$ *are elements of* $\{\hat{b}_{\tau,js}\}_{m \times q}$ *denoted in (2.13); and* τ^* *is denoted in Theorem 2.3.1.*

In order to achieve asymptotic normality for the selected model (i.e., only using the regressors selected by Theorem 2.3.1), the rate of convergence of $\hat{\beta}_{\tau,jU}$ to $\hat{\beta}_{ora}(\bar{z}^j)$ should be much faster than $\frac{1}{\sqrt{NT}}$. The oracle property in Theorem 2.3.2 implies such a result as long as $\|\tau^*\|$ is much smaller than $\sqrt{NT}$. Therefore the simple choice of $\sqrt{NT}$ level for τ suggested above is not sufficient to achieve an asymptotic normality. Thus, in what follows we propose a data-driven procedure for choosing τ, which yields a much faster rate of convergence ($O_P\left(\frac{1}{NT}\right)$) to the oracle and then achieve the desired asymptotic normality property. From now on, we assume that whenever

[5]Notice that the word "oracle" refers to the same estimator as given in (2.7) but by assuming we know the true set $\mathcal{A}$. Here we completely ignore the inefficiency caused by the irrelevant covariates $\tilde{Z}_{it}$. The asymptotically efficient estimator is obtained when we know both the set $\mathcal{A}$ and all the irrelevant covariates. However, this can only be done at a certain probability based on Theorem 2.2.1.

$b_{0s} \neq 0$ for $s = 1, \ldots, q^*$, its ℓ_2 norm is larger than some universal constant $\|b_{0s}\| \geq \alpha_0 > 0$. This assumption is natural in the current fixed dimension setting.

As in Wang and Xia (2009), we use the following data-driven regularizer weight

$$\tau = \tilde{\tau} \left(\|\tilde{b}_1\|^{-1}, \ldots, \|\tilde{b}_q\|^{-1} \right)', \tag{2.15}$$

where $\tilde{\tau}$ is a scalar, $\tilde{b}_s$ is the s^{th} column of the unregularized estimator $\tilde{B}$, and $\tilde{B}$ is obtained from (2.14) by simply choosing $\tau_1 = \cdots = \tau_q = 0$. Using Assumption C and the first result of Theorem 2.3.1, it is easy to verify that $\|\tilde{b}_s\|^{-1} = O_P(1)$ for $s = 1, \ldots, q^*$ and $\|\tilde{b}_s\| = O_P\left(\frac{1}{\sqrt{NT}}\right)$ for $s = q^* + 1, \ldots, q$. In (2.15), the unregularized estimator $\tilde{B}$ is just the desired $(\sqrt{NT})$ consistent estimator. Given $\tilde{B}$, it is straightforward to tell which column of B_0 is likely to be zero or not. Specifically, a smaller $\|\tilde{b}_s\|$ implies that the s^{th} column is more likely to be zero and hence suggests a larger regularizer on $\|b_s\|$. Given the form of τ in (2.15), a selection on the vector τ becomes a selection on the scalar $\tilde{\tau}$. Note that the properties of $\|\tilde{b}_s\|^{-1}$ for $s = 1, \ldots, q$ imply that a large enough constant $\tilde{\tau}$ would satisfy all the technical conditions on τ needed for the above theorems with $\left\| \hat{\beta}_{\tau,jU} - \hat{\beta}_{ora}(\bar{z}^j) \right\| = O_P\left(\frac{1}{NT}\right)$. More specifically, we select the constant $\tilde{\tau}$ by the following modified BIC-type (MBIC) criterion.

$$BIC_{\tilde{\tau}} = \ln RSS_{\tilde{\tau}} + df_{\tilde{\tau}} \cdot \frac{\ln(NT)}{NT},$$

where $df_{\tilde{\tau}}$ is simply the number of nonzero coefficients identified by $\hat{B}_{\tilde{\tau}}$; $\hat{B}_{\tilde{\tau}}$ is obtained by using (2.13) and (2.15), i.e. $\hat{B}_{\tilde{\tau}} = (\hat{\beta}_{\tilde{\tau},1}, \ldots, \hat{\beta}_{\tilde{\tau},m})' = (\hat{b}_{\tilde{\tau},1}, \ldots, \hat{b}_{\tilde{\tau},q})$; and $RSS_{\tilde{\tau}}$ is defined as

$$RSS_{\tilde{\tau}} = \frac{1}{NT} \sum_{j=1}^{m} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\tilde{Y}_{it} - \tilde{X}_{it}' \hat{\beta}_{\tilde{\tau},j} \right)^2 L(Z_{it}, z^j, \hat{\lambda}).$$

The optimal weight parameter can then be obtained by

$$\hat{\tilde{\tau}} = \underset{\tilde{\tau}}{\operatorname{argmin}}\, BIC_{\tilde{\tau}}. \tag{2.16}$$

Recall that the true set of nonzero coefficients is denoted by $\mathcal{A}^c = \{1, \ldots, p^*\}$. Let $S_{\hat{\tilde{\tau}}} = \{j : \|\hat{\beta}_{\hat{\tilde{\tau}},j}\| > 0, 1 \leq j \leq q\}$ denote the variables selected using the regularized estimator $\hat{B}_{\hat{\tilde{\tau}}}$, where the tuning parameter is obtained using (2.16). With these notations, we present our next result as follows.

Theorem 2.3.3. *Under conditions of Theorem 2.3.1, as $(N, T) \to (\infty, \infty)$, the weight parameter selected by the modified BIC-type criterion (2.16) can:*

1. *Identify the true model consistently, i.e.* $\Pr(S_{\hat{\tilde{\tau}}} = \mathcal{A}^c) \to 1$;

2. *For the relevant covariate case, achieve the asymptotic normality, i.e.*

$$\sqrt{NT}(\hat{\beta}_{\hat{\tilde{\tau}},jU} - \beta_U(z^j)) \to_D N(0, \Xi_1^*(z^j)^{-1}\Xi_0^*(z^j)\Xi_1^*(z^j)^{-1}) \tag{2.17}$$

for $j = 1, \ldots, m$, where $\beta_U(z^j) = (\beta_1(z^j), \ldots, \beta_{q^*}(z^j))'$; $\Xi_0^*(z^j)$ and $\Xi_1^*(z^j)$ are the $q^* \times q^*$ principal sub-matrices of $\Xi_0(z^j)$ and $\Xi_1(z^j)$ denoted in Theorem 2.1.2 respectively; and $\beta_U(z^j)$ denotes the first q^* elements of $\beta(z^j)$.

3. *For the irrelevant covariate case,*

$$\hat{\beta}_{\hat{\tilde{\tau}},jU} - \beta_U(\bar{z}^j) = O_P\left(\frac{1}{\sqrt{NT}}\right) \tag{2.18}$$

for $j = 1, \ldots, m$, where $\beta_U(\bar{z}^j) = (\beta_1(\bar{z}^j), \ldots, \beta_{q^*}(\bar{z}^j))'$.

Having derived the asymptotic results for the finite dimension case in Sections 2.1-2.3, in the following subsection we will briefly discuss some extensions.

2.4 Extensions

In this subsection we will briefly show that our modified within transformation remains valid for the case where T is small, by using $\sum_{s=1}^T u_{is}L_{is,it}^p / \sum_{s=1}^T L_{is,it}^p$ as an example.

In (2.5), we have shown that

$$L(Z_{it}, z, \lambda) = 1(Z_{it} = z) + \sum_{m=1}^r \lambda_m 1_{m,Z_{it}=z} + \cdots + \prod_{m=1}^r \lambda_m 1(Z_{it,m} \neq z_m).$$

For sufficiently small λ,

- If $\sum_{s=1}^T 1(Z_{is} = Z_{it}) \neq 0$, it is obvious that $\lim_{\lambda \to 0_{r\times 1}} \sum_{s=1}^T u_{is}L_{is,it}^p / \sum_{s=1}^T L_{is,it}^p$ exists.
- If $\sum_{s=1}^T 1(Z_{is} = Z_{it}) = 0$, we just need to focus on the limit of $\lim_{\lambda \to 0_{r\times 1}} f(\lambda)/g(\lambda)$, where

$$f(\lambda) = \sum_{s=1}^T u_{is}\left(\sum_{m=1}^r \lambda_m 1_{m,Z_{is}=Z_{it}} + \cdots + \prod_{m=1}^r \lambda_m 1(Z_{is,m} \neq Z_{it,m})\right)^p,$$
$$g(\lambda) = \sum_{s=1}^T \left(\sum_{m=1}^r \lambda_m 1_{m,Z_{is}=Z_{it}} + \cdots + \prod_{m=1}^r \lambda_m 1(Z_{is,m} \neq Z_{it,m})\right)^p.$$

  Since both $f(\lambda)$ and $g(\lambda)$ are the polynomial functions of the elements of λ, it is easy to show that $\lim_{\lambda \to 0_{r\times 1}} f(\lambda)/g(\lambda)$ does exist.

Note that the existence of the above limit is uniform in i and t. Hence, for simplicity, one just needs to denote that $A_{u,it} = \lim_{\lambda \to 0_{r\times 1}} \sum_{s=1}^T u_{is}L_{is,it}^p / \sum_{s=1}^T L_{is,it}^p$. Then we know that the within

transformation does make sense for the small T case. The rest of the derivation follows the same lines as for the large N and T case. In our Monte Carlo study in the following section, we will demonstrate that our methodology works for the fixed T case as well.

For the cases where some of the discrete covariates are ordinal, the above kernel function (2.2) can be changed to

$$l(Z_{it,s}, z_s, \lambda_s) = \begin{cases} 1, & \text{if } Z_{it,s} = z_s \\ \lambda_s^{|Z_{it,s}-z_s|}, & \text{otherwise} \end{cases}, \tag{2.19}$$

which has been well documented in the literature (see Li and Racine (2010) and Li et al. (2013) for details). Then, it is straightforward to show that the asymptotic results established in Sections 2.1-2.3 remain valid.

3 Monte Carlo Study

In this section, we perform a Monte Carlo study to investigate the finite sample properties of our model and estimator.[6] The data generating process (DGP) is as follows.

$$Y_{it} = X_{it}'\beta(Z_{it}) + w_i + u_{it} \quad \text{and} \quad X_{it} = H_{it} + V_{it}. \tag{3.1}$$

Let $Z_{it} = (Z_{it,1}, \cdots, Z_{it,r})'$, where for $\forall j = 1, \ldots, r$, $Z_{it,j}$ is i.i.d. over i and t; and Z_{it} is chosen from $\{0, 1\}$ with the same probability every time, i.e. $\Pr(Z_{it,j} = 0) = \Pr(Z_{it,j} = 1) = 0.5$. V_{it} is i.i.d. over i and t and follows a normal distribution $N(Z_{it,1}/2 \cdot i_q, \sqrt{Z_{it,1} + 1} \cdot I_q)$, where i_q is a $q \times 1$ one vector and I_q is a q-dimensional identity matrix. $H_{it} = (H_{it,1}, \ldots, H_{it,q})'$. For $\forall j = 1, \ldots, q$, $H_{it,j}$ is generated as $H_{it,j} = \rho(j)H_{it-1,j} + i.i.d.\ N(0, 1)$ and $\rho(j) = 0.1 * \lfloor 9 \cdot U(0, 1) \rfloor$, where $U(0, 1)$ denotes the uniform distribution; $\lfloor a \rfloor$ denotes rounding the element of a to the nearest integer greater than or equal to that element, i.e. $a \leq \lfloor a \rfloor$. Thus, for $\forall j = 1, \ldots, q$, $H_{it,j}$ is independent in the cross-sectional dimension and a stationary AR(1) process in the time-series dimension with the coefficient $\rho(j)$ being randomly chosen from the set $\{0.1, 0.2, \ldots, 0.9\}$.

The fixed effects are generated using $w_i = \frac{1}{Tq} \sum_{t=1}^{T} \sum_{j=1}^{q} X_{it,j}$ to ensure that it is correlated with the regressors and covariates. To introduce cross-sectional dependence, the error terms (denoted by $u_t = (u_{1t}, \ldots, u_{Nt})$) are generated using $u_t = 0.5u_{t-1} + \varepsilon_t$, where $\varepsilon_t \sim i.i.d.\ N(0_{N\times 1}, \Sigma_u)$ and for $i, j = 1, \ldots, N$ the $(i, j)^{th}$ element of Σ_u is $0.5^{|i-j|}$.

When conducting Monte Carlo simulation, we consider both relevant and irrelevant cases. Formally, these two cases are generated as follows:

- Relevant covariate case: $\beta_j(Z_{it}) = j/2 \cdot \sum_{k=1}^{r} Z_{it,k} + 1$,

[6] The Matlab codes are available upon request, and will be published on authors' website soon.

- Irrelevant covariate case: $\beta_j(Z_{it}) = j/2 \cdot Z_{it,1} + 1$,

where $\beta_j(Z_{it})$ denotes the j^{th} element of the coefficient function $\beta(z)$ for $\forall j = 1, \ldots, q$.

More specifically, we consider the following four sub-cases:

1. Relevant covariate case with $q = 3$, $r = 2$,
2. Irrelevant covariate case with $q = 3$, $r = 2$,
3. Relevant covariate case with $q = 5$, $r = 2$, $q^* = 2$ (i.e., $\beta_j(z) = 0$ for $j \geq 3$),
4. Irrelevant covariate case with $q = 5$, $r = 2$, $q^* = 2$ (i.e., $\beta_j(z) = 0$ for $j \geq 3$),

where the variable p used for implementing the within transformation is always chosen as 5.

For sub-cases 1 and 2, we estimate the model in (3.1) using (2.7) for each generated data set.[7] For notational convenience, this method is referred to as the "DMK" model, where DM stands for demeaned variables (i.e., variables formed using the modified within transformation) and K means that the estimates are obtained using the the kernel function. For comparison purpose, we also estimate a variant of (2.7), where every kernel function is replaced with the indicator function. This method is referred to as "DMI".

For each generated data set and the corresponding estimate on $\beta(z)$, we calculate the squared error (SE) as follows.

$$\text{SE} = \sum_{z \in \mathcal{D}} p(z) \left(\hat{\beta}_j(z) - \beta_j(z) \right)^2, \tag{3.2}$$

where, for $j = 1, \ldots, q$, $\hat{\beta}_j(z)$ denotes the j^{th} element of $\hat{\beta}(z)$. We then replicate the above procedure 1000 times and report mean squared errors (MSE) for sub-cases 1 and 2 respectively in Table 1, where NA indicates the value can not be calculated, because the denominator (T_{it}) becomes 0. As can be seen from Table 1, when T is small (i.e., $T = 5$ or 7) relative to the cardinality of the support of Z_{it}, the use of the DMI model results in many NAs in both the relevant and irrelevant covariate cases. This is because the denominator of the DMI model (i.e., T_{it}) tends to be zero when T is small. When N and T are large, both DMK and DMI yield very small MSEs regardless of the nature of the covariates. However, we note that the DMK model outperforms the DMI model in the irrelevant covariate case in that the former model yields smaller MSEs.

For sub-cases 3 and 4, our estimates of $\beta(z)$ are expected to have three columns of zero. For each generated data set, we estimate $\hat{B}_\tau$ by (2.13).[8] To evaluate alternative estimators, we

[7] As explained previously, $p = 2$ is enough in (2.4) in practice. We choose $p = 2$ for the simulated and real data studies in this paper. We have experimented a variety of choices on p, where the results are almost identical and the differences happen after the fourth decimal for both Monte Carlo study and the application to U.S. commercial banks provided in the next section.

[8] The algorithm is provided in Appendix B.

Table 1: MSEs for Sub-cases 1 and 2 ($q = 3$ and $r = 2$)

			DMK			DMI		
		$T \setminus N$	50	100	200	50	100	200
Relevant	$\hat{\beta}_1(z)$	5	0.02174	0.00986	0.00457	NA	NA	NA
		7	0.01001	0.00488	0.00241	NA	NA	NA
		20	0.00188	0.00094	0.00045	0.00189	0.00094	0.00045
		40	0.00079	0.00039	0.00019	0.00079	0.00039	0.00019
	$\hat{\beta}_2(z)$	5	0.02286	0.00972	0.00464	NA	NA	NA
		7	0.01021	0.00516	0.00235	NA	NA	NA
		20	0.00182	0.00092	0.00047	0.00183	0.00093	0.00047
		40	0.00079	0.00041	0.00020	0.00079	0.00041	0.00020
	$\hat{\beta}_3(z)$	5	0.02290	0.00966	0.00482	NA	NA	NA
		7	0.01051	0.00504	0.00245	NA	NA	NA
		20	0.00183	0.00089	0.00044	0.00182	0.00089	0.00044
		40	0.00081	0.00040	0.00020	0.00081	0.00040	0.00020
Irrelevant	$\hat{\beta}_1(z)$	5	0.01407	0.00645	0.00308	NA	NA	NA
		7	0.00628	0.00318	0.00156	NA	NA	NA
		20	0.00116	0.00060	0.00028	0.00189	0.00093	0.00045
		40	0.00049	0.00024	0.00012	0.00079	0.00039	0.00019
	$\hat{\beta}_2(z)$	5	0.01426	0.00645	0.00318	NA	NA	NA
		7	0.00639	0.00336	0.00152	NA	NA	NA
		20	0.00113	0.00057	0.00030	0.00182	0.00093	0.00048
		40	0.00046	0.00025	0.00012	0.00079	0.00041	0.00020
	$\hat{\beta}_3(z)$	5	0.01479	0.00637	0.00318	NA	NA	NA
		7	0.00660	0.00322	0.00159	NA	NA	NA
		20	0.00115	0.00056	0.00029	0.00184	0.00089	0.00044
		40	0.00050	0.00025	0.00012	0.00081	0.00040	0.00020

1. $\hat{\beta}_j(z)$ denotes the j^{th} element of $\hat{\beta}(z)$.
2. NA indicates the value can not be calculated, because the denominator (T_{it}) becomes 0.

compute a modified measure of squared error (SE_1). Specifically, we calculate the conventional squared error for each element of $\hat{B}_\tau$ in each replication, store them in matrix MB, and then sum up the elements of MB as follows to get SE_1:

$$\text{SE}_1 = \frac{1}{q}\sum_{s=1}^{q}\sum_{j=1}^{m} p(z^j)\text{MB}_{js}, \tag{3.3}$$

where MB_{js} represent the $(j,s)^{th}$ element of MB; m and z^j are denoted in (2.14). We then replicate the above procedure 1000 times and report the mean of SE_1 (MSE_1). For comparison, we also estimate the model in (3.1) using the unregularized estimator and the oracle estimator respectively. For each of these two estimators, we report its associated MSE_1's as defined in (3.3). The results are summarized in Table 2. As can be seen, the oracle estimator has smaller MES_1's compared with the regularized and unregularized estimators. This is not surprising, because oracle estimator uses full information when implementing the regression. In addition, we note that the regularized estimator produces lower MES_1's than the unregularized estimator. As N and T are sufficiently large, the MSE_1's from the regularized estimator are very close to those from the oracle estimator.

In sum, our Monte Carlo study suggests that our methodology works well for large N and small T, and large N and T cases. To further show the usefulness of our methodology in solving

Table 2: MSE_1 of Sub-cases 3 and 4 (with $q = 5$, $r = 2$, $q^* = 2$)

		Relevant			Irrelevant		
	$T \setminus N$	50	100	200	50	100	200
Regularized	5	0.01576	0.00518	0.00220	0.01239	0.00360	0.00149
	7	0.00583	0.00236	0.00102	0.00395	0.00157	0.00067
	20	0.00081	0.00038	0.00019	0.00049	0.00023	0.00011
	40	0.00031	0.00016	0.00008	0.00018	0.00009	0.00005
Unregularized	5	0.02284	0.00975	0.00465	0.01527	0.00629	0.00300
	7	0.01045	0.00498	0.00237	0.00649	0.00317	0.00149
	20	0.00189	0.00091	0.00045	0.00112	0.00056	0.00028
	40	0.00076	0.00039	0.00020	0.00044	0.00023	0.00012
Oracle	5	0.00825	0.00378	0.00185	0.00562	0.00246	0.00121
	7	0.00422	0.00200	0.00094	0.00259	0.00127	0.00060
	20	0.00075	0.00037	0.00018	0.00045	0.00022	0.00011
	40	0.00030	0.00015	0.00008	0.00018	0.00009	0.00005

real-world problems, in the following section we provide an application to commercial banks in the U.S..

4 An Application to U.S. Commercial Banks

In this section we provide an application of the varying-coefficient model proposed in Section 2 to the analysis of the effects of geographical deregulation on the returns to scale of commercial banks in the U.S.. Until the middle of the 1970's banking in the U.S. was heavily regulated at the state level. Generally, there were three different types of state regulation on bank branching: "unit banking", where banks were only permitted to operate in one location; "limited branching", where the branching abilities of individual banks were limited to a portion of the state; and "statewide banking" where individual banks were permitted to branch statewide. In the mid-1980s individual states began to loosen regulations on intrastate branching, often moving from unit banking to limited branching and then to statewide banking. It is worth noting that different states changed their regulatory restrictions on expansion at different times. This deregulation process eventually culminated in the passage of the Riegle-Neal Interstate Banking and Branching Efficiency of 1994, which permitted nationwide branching as of June 1997 (Jayaratne and Strahan, 1997). In sum, commercial banks in the U.S. undergone four branch banking regimes in the 1980s and 1990s: (1) unit banking, (2) limited branching, (3) statewide banking, and (4) full interstate branching, thus offering researchers a unique opportunity to study the effects of geographical deregulation on the returns to scale of commercial banks in the U.S..

The data used in this application are obtained from the Reports of Income and Condition (Call Reports) published by the Federal Reserve Bank of Chicago. The sample covers the period 1986-2005, a period that includes the four policy regimes. We examine only continuously operating large banks with assets of at least \$1 billion (in 1986 dollars) to avoid the impact of entry and exit and to focus on the performance of a core of healthy, surviving institutions.

This gives a total of 466 banks over 20 years (i.e. 80 quarters, so $N = 466$ and $T = 80$). To select the relevant variables, we follow the commonly-accepted intermediation approach (Sealey and Lindley, 1977). On the input side, three inputs are included: (1) the quantity of labor; (2) the quantity of purchased funds and deposits; and (3) the quantity of physical capital, which includes premises and other fixed assets. On the output side, three outputs are specified: (1) consumer loans; (2) securities, which includes all non-loan financial assets; and (3) non-consumer loans, which is composed of industrial, commercial, and real estate loans. All the quantities are constructed as in Berger and Mester (2003). These quantities are also deflated by the GDP deflator to the base year 1986, except for the quantity of labor.

4.1 The Varying-Coefficient Translog Cost Function

We use a varying-coefficient translog cost function, which has the standard form of the varying-coefficient model described in Section 2, to represent the production technology of commercial banks in the U.S.. A primary feature of this function is that its coefficients are allowed to vary depending on the banking regime under which a bank operates, because there is considerable evidence that branch banking regime affects production technology (Mason, 2013; Mester, 2005). Specifically, this function is written as[9]

$$\begin{aligned}\ln C = \alpha_0(Z) + \sum_{j=1}^{\bar{N}} \alpha_j(Z) \ln W_j + \sum_{m=1}^{\bar{M}} \gamma_m(Z) \ln Y_m + \tau(Z)t + \frac{1}{2}\delta(Z)t^2 \\ + \frac{1}{2}\sum_{j=1}^{\bar{N}}\sum_{k=1}^{\bar{N}} \beta_{jk}(Z) \ln W_j \ln W_k + \frac{1}{2}\sum_{m=1}^{\bar{M}}\sum_{n=1}^{\bar{M}} \rho_{mn}(Z) \ln Y_m \ln Y_n \\ + \sum_{j=1}^{\bar{N}}\sum_{m=1}^{\bar{M}} \psi_{jm}(Z) \ln W_j \ln Y_m + \sum_{j=1}^{\bar{N}} \phi_j(Z) t \ln W_j + \sum_{m=1}^{\bar{M}} \varphi_m(Z) t \ln Y_m, \end{aligned} \tag{4.1}$$

where C is total cost; t is a time trend; Y_m for $m = 1, \ldots, \bar{M}$ is a variable representing output; and W_j for $j = 1, \ldots, \bar{N}$ is a variable representing input price. In our case, $\bar{N} = \bar{M} = 3$. Z is specified to be a four-category variable indicating different branch banking regimes that existed during our sample period. Specifically, we set $Z = 0$ for banks operating in unit banking states, $Z = 1$ for banks operating in limited branching states, $Z = 2$ for banks operating in statewide banking states, and $Z = 3$ for banks operating in nationwide branching states. As previously noted, different states changed their regulatory restrictions on expansion at different times, indicating that Z varies in both the cross-sectional and time series dimensions.

The usual symmetry restrictions require $\beta_{jk}(Z) = \beta_{kj}(Z)$ for $j, k = 1, \ldots, \bar{N}$ and $\rho_{mn}(Z) =$

[9]The variable selection method outlined in Section 2.3 is not needed here, because microeconomic theory provides clear guidance on what variables should be included in cost functions (see, for example, Diewert and Wales (1987)). In addition, the translog functional form is commonly used in the literature, since it provides a second order approximation to the underlying true cost function (Christensen et al., 1975).

$\rho_{nm}(Z)$ for $m, n = 1, \ldots, \bar{M}$. Moreover, to ensure linear homogeneity of the cost function in input prices, the following restrictions are imposed

$$\sum_{j=1}^{\bar{N}} \alpha_j(Z) = 1, \quad \sum_{j=1}^{\bar{N}} \beta_{jk}(Z) = \sum_{j=1}^{\bar{N}} \psi_{jm}(Z) = \sum_{j=1}^{\bar{N}} \phi_j(Z) = 0. \tag{4.2}$$

To impose the linear homogeneity restrictions in (4.2), we follow Griffiths et al. (2000) and normalize the cost and input prices in (4.1) by one of the input prices (say, $W_{\bar{N}}$)

$$\begin{aligned} \ln \frac{C}{W_{\bar{N}}} = {} & \alpha_0(Z) + \sum_{j=1}^{\bar{N}-1} \alpha_j(Z) \ln \frac{W_j}{W_{\bar{N}}} + \sum_{m=1}^{\bar{M}} \gamma_m(Z) \ln Y_m + \tau(Z)t + \frac{1}{2}\delta(Z)t^2 \\ & + \frac{1}{2} \sum_{j=1}^{\bar{N}-1} \sum_{k=1}^{\bar{N}-1} \beta_{jk}(Z) \ln \frac{W_j}{W_{\bar{N}}} \ln \frac{W_k}{W_{\bar{N}}} + \frac{1}{2} \sum_{m=1}^{\bar{M}} \sum_{n=1}^{\bar{M}} \rho_{mn}(Z) \ln Y_m \ln Y_n \\ & + \sum_{j=1}^{\bar{N}-1} \sum_{m=1}^{\bar{M}} \psi_{jm}(Z) \ln \frac{W_j}{W_{\bar{N}}} \ln Y_m + \sum_{j=1}^{\bar{N}-1} \phi_j(Z)t \ln \frac{W_j}{W_{\bar{N}}} + \sum_{m=1}^{\bar{M}} \varphi_m(Z)t \ln Y_m. \end{aligned} \tag{4.3}$$

In matrix notations, the normalized varying-coefficient translog cost function in (4.3), after appending a fixed effect term and a random error term, can be written as (2.1), where the dependent variable is $\ln \frac{C}{W_{\bar{N}}}$; the regressors are a vector comprising all the variables which appear on the right hand side of (4.3); and $\beta(\cdot)$ is the corresponding vector of coefficients of the translog function. Note that after the within transformation $\alpha_0(Z)$ will disappear along with the fixed effect. However, this does not affect our empirical results.

Given the estimated parameters of (4.3)[10], it is possible to compute returns to scale as $\text{RTS} = \left(\sum_{m=1}^{\bar{M}} \epsilon_{cY_m}\right)^{-1}$, where for $m = 1, \ldots, \bar{M}$

$$\epsilon_{cY_m} = \frac{\partial \ln C}{\partial \ln Y_m} = \gamma_m(Z) + \sum_{n=1}^{\bar{M}} \rho_{mn}(Z) \ln Y_n + \sum_{j=1}^{\bar{N}} \psi_{jm}(Z) \ln W_j + \varphi_m(Z)t$$

is the cost elasticity of the j^{th} output.

For comparison purposes, we also consider a fully parametric translog cost function, in which three binary variables are used to control for the different branch banking regimes. Specifically, (i) UNIT equals to 1 for banks operating in unit banking states (0 otherwise); (ii) LIMITED equals to 1 for banks operating in limited branching states (0 otherwise); and (iii) STATEWIDE equals to 1 for banks operating in statewide banking states (0 otherwise). Specifically, the normalized fully parametric translog cost function is written as

[10]There are two methods to estimate this cost function: one is to estimate it directly and the other is to estimate it together with its share equations. From an economic theoretical perspective, both methods are correct although the second one has better statistical efficiency (see, for example, Feng and Serletis (2008)). However, to better illustrate our single equation panel data varying-coefficient model, we use the first method in this paper.

$$\ln \frac{C}{W_{\bar{N}}} = \alpha_0 + \sum_{j=1}^{\bar{N}-1} \alpha_j \ln \frac{W_j}{W_{\bar{N}}} + \sum_{m=1}^{\bar{M}} \gamma_m \ln Y_m + \tau t + \frac{1}{2}\delta t^2 + \frac{1}{2}\sum_{j=1}^{\bar{N}-1}\sum_{k=1}^{\bar{N}-1} \beta_{jk} \ln \frac{W_j}{W_{\bar{N}}} \ln \frac{W_k}{W_{\bar{N}}}$$
$$+\frac{1}{2}\sum_{m=1}^{\bar{M}}\sum_{n=1}^{\bar{M}} \rho_{mn} \ln Y_m \ln Y_n + \sum_{j=1}^{\bar{N}-1}\sum_{m=1}^{\bar{M}} \psi_{jm} \ln \frac{W_j}{W_{\bar{N}}} \ln Y_m + \sum_{j=1}^{\bar{N}-1} \phi_j t \ln \frac{W_j}{W_{\bar{N}}}$$
$$+ \sum_{m=1}^{\bar{M}} \varphi_m t \ln Y_m + \xi_1 \text{UNIT} + \xi_2 \text{LIMITED} + \xi_3 \text{STATEWIDE}, \tag{4.4}$$

where symmetry requires $\beta_{jk} = \beta_{kj}$ and $\rho_{mn} = \rho_{nm}$. In matrix notations, (4.4), after appending a fixed effect term and a random error term, can be written as

$$Y_{it} = X_{it}'\beta_0 + w_i + u_{it}, \tag{4.5}$$

where X_{it} is a vector comprising all the variables which appear on the right hand side of (4.4); and β_0 is the corresponding vector of coefficients of the translog function (including the intercept).

4.2 Empirical Results

We estimate the normalized varying-coefficient translog cost function in (4.3), using the panel data varying-coefficient estimator in (2.7). Parameter estimates and standard errors associated with this function are reported in Panel A of Table 3. We also estimate the normalized fully translog cost function in (4.4) and report its parameter estimates and standard errors in Panel B of Table 3. To compare the performance of these two competing models, we perform a test using the procedure proposed by Li et al. (2013). If we treat $(\alpha_0 + \xi_1 \text{UNIT} + \xi_2 \text{LIMITED} + \xi_3 \text{STATEWIDE})$ in the fully parametric translog cost function as the coefficient for the constant term, it is easy to see that the fully parametric translog cost function in (4.4) is a special case of the varying-coefficient translog cost function in (4.3). With this in mind, then, testing if the varying-coefficient translog cost function outperforms the fully parametric translog cost function is equivalent to testing if the latter model has the same specification as the former model, or more specifically, if the latter model has the same set of coefficients as the former model. To test parameter constancy, we extend the bootstrap-based procedure outlined in Li et al. (2013) to a panel data setting. Detailed description of the procedure can be found therein. For our case, the test statistic is 0.4968, well above the critical value of 0.0876 at 1% level of significance, suggesting strongly that the null hypothesis is rejected. In other words, the varying-coefficient translog cost function is preferred to the fully parametric translog cost function.

It is also of interest to compare results from the varying-coefficient translog cost function where the bandwidth (λ) is optimally selected using (2.8) with results from the same cost function but with λ set to zero *a priori*. The latter function can be obtained by replacing the kernel

Table 3: Estimates of Different Methods

	Panel A: (2.7) with bandwidth $\hat{\lambda} = 0.0003$ selected by (2.8)								Panel B		Panel C: (2.7) with bandwidth $\lambda = 0$ for all kernels							
	UNIT		LIMITED		STATEWIDE		NATIONWIDE		Fully parametric		UNIT		LIMITED		STATEWIDE		NATIONWIDE	
	Est	std	Est	std	Est	std	Est	std	Est	std	Est	std	Est	std	Est	std	Est	std
α_1	0.3912	0.0295	0.3343	0.0129	0.3442	0.0055	0.4199	0.0091	0.3554	0.0045	0.4984	0.0566	0.3329	0.0131	0.3444	0.0053	0.4208	0.0092
α_2	0.0556	0.0121	0.0411	0.0056	0.0539	0.0031	0.0574	0.0033	0.0633	0.0019	0.0568	0.0136	0.0412	0.0056	0.0539	0.0031	0.0571	0.0032
α_3	0.5532	0.0371	0.6246	0.0140	0.6019	0.0059	0.5227	0.0097	0.5814	0.0047	0.4448	0.0631	0.6259	0.0143	0.6017	0.0059	0.5221	0.0098
γ_1	0.1844	0.0507	0.2823	0.0129	0.3079	0.0064	0.3636	0.0062	0.3213	0.0034	0.1686	0.0609	0.2819	0.0126	0.3079	0.0063	0.3637	0.0061
γ_2	0.0514	0.0169	0.1381	0.0088	0.1218	0.0043	0.1163	0.0048	0.1304	0.0023	0.0236	0.0240	0.1382	0.0088	0.1217	0.0042	0.1163	0.0045
γ_3	0.6368	0.0252	0.4759	0.0133	0.5068	0.0072	0.4877	0.0082	0.5022	0.0040	0.6632	0.0344	0.4756	0.0132	0.5069	0.0069	0.4878	0.0079
β_{11}	0.1055	0.0440	0.0612	0.0204	0.1758	0.0096	0.2159	0.0124	0.1797	0.0128	0.1550	0.0621	0.0602	0.0200	0.1759	0.0098	0.2162	0.0128
β_{12}	-0.0001	0.0071	0.0045	0.0033	-0.0276	0.0027	-0.0173	0.0029	-0.0190	0.0021	-0.0007	0.0083	0.0046	0.0032	-0.0276	0.0025	-0.0174	0.0029
β_{13}	-0.1054	0.0477	-0.0657	0.0213	-0.1482	0.0091	-0.1986	0.0115	-0.1607	0.0120	-0.1543	0.0667	-0.0648	0.0206	-0.1482	0.0093	-0.1989	0.0118
β_{21}	-0.0001	0.0071	0.0045	0.0033	-0.0276	0.0027	-0.0173	0.0029	-0.0190	0.0021	-0.0007	0.0083	0.0046	0.0032	-0.0276	0.0025	-0.0174	0.0029
β_{22}	0.0222	0.0031	0.0155	0.0010	0.0170	0.0014	0.0178	0.0011	0.0162	0.0007	0.0221	0.0030	0.0155	0.0010	0.0170	0.0014	0.0178	0.0011
β_{23}	-0.0221	0.0070	-0.0200	0.0033	0.0106	0.0026	-0.0005	0.0031	0.0028	0.0021	-0.0214	0.0077	-0.0201	0.0034	0.0106	0.0025	-0.0004	0.0031
β_{31}	-0.1054	0.0477	-0.0657	0.0213	-0.1482	0.0091	-0.1986	0.0115	-0.1607	0.0120	-0.1543	0.0667	-0.0648	0.0206	-0.1482	0.0093	-0.1989	0.0118
β_{32}	-0.0221	0.0070	-0.0200	0.0033	0.0106	0.0026	-0.0005	0.0031	0.0028	0.0021	-0.0214	0.0077	-0.0201	0.0034	0.0106	0.0025	-0.0004	0.0031
β_{33}	0.1274	0.0520	0.0857	0.0223	0.1376	0.0092	0.1992	0.0113	0.1579	0.0115	0.1757	0.0714	0.0849	0.0214	0.1376	0.0094	0.1993	0.0116
ρ_{11}	0.1126	0.0220	0.1196	0.0119	0.1464	0.0045	0.1567	0.0049	0.1463	0.0036	0.1105	0.0228	0.1195	0.0117	0.1465	0.0046	0.1567	0.0047
ρ_{12}	0.0170	0.0158	0.0082	0.0069	-0.0099	0.0024	-0.0095	0.0027	-0.0028	0.0020	0.0162	0.0170	0.0083	0.0068	-0.0099	0.0024	-0.0095	0.0026
ρ_{13}	-0.1727	0.0108	-0.1521	0.0063	-0.1530	0.0029	-0.1493	0.0052	-0.1527	0.0026	-0.1762	0.0117	-0.1521	0.0064	-0.1530	0.0030	-0.1493	0.0048
ρ_{21}	0.0170	0.0158	0.0082	0.0069	-0.0099	0.0024	-0.0095	0.0027	-0.0028	0.0020	0.0162	0.0170	0.0083	0.0068	-0.0099	0.0024	-0.0095	0.0026
ρ_{22}	0.0145	0.0186	0.0601	0.0038	0.0351	0.0025	0.0190	0.0027	0.0280	0.0012	0.0150	0.0204	0.0602	0.0039	0.0351	0.0024	0.0190	0.0026
ρ_{23}	-0.0341	0.0110	-0.0568	0.0055	-0.0158	0.0032	0.0015	0.0034	-0.0141	0.0021	-0.0354	0.0117	-0.0570	0.0058	-0.0158	0.0032	0.0015	0.0033
ρ_{31}	-0.1727	0.0108	-0.1521	0.0063	-0.1530	0.0029	-0.1493	0.0052	-0.1527	0.0026	-0.1762	0.0117	-0.1521	0.0064	-0.1530	0.0030	-0.1493	0.0048
ρ_{32}	-0.0341	0.0110	-0.0568	0.0055	-0.0158	0.0032	0.0015	0.0034	-0.0141	0.0021	-0.0354	0.0117	-0.0570	0.0058	-0.0158	0.0032	0.0015	0.0033
ρ_{33}	0.2414	0.0137	0.2011	0.0060	0.1758	0.0045	0.1460	0.0076	0.1669	0.0033	0.2480	0.0141	0.2011	0.0061	0.1758	0.0044	0.1459	0.0072
τ	-0.0034	0.0009	-0.0035	0.0004	-0.0031	0.0002	-0.0071	0.0006	-0.0031	0.0001	0.0006	0.0030	-0.0035	0.0004	-0.0031	0.0002	-0.0073	0.0006
δ	-0.0002	0.0000	-0.0001	0.0000	0.0000	0.0000	0.0002	0.0000	0.0000	0.0000	-0.0001	0.0001	-0.0001	0.0000	0.0000	0.0000	0.0002	0.0000
ψ_{11}	0.0353	0.0182	0.0333	0.0137	0.0026	0.0039	0.0153	0.0040	0.0235	0.0036	0.0327	0.0208	0.0330	0.0130	0.0026	0.0041	0.0154	0.0038
ψ_{12}	0.0072	0.0112	-0.0169	0.0097	-0.0128	0.0030	0.0012	0.0029	-0.0059	0.0025	0.0037	0.0141	-0.0166	0.0093	-0.0128	0.0029	0.0012	0.0029
ψ_{13}	-0.0244	0.0139	-0.0090	0.0082	0.0174	0.0046	-0.0146	0.0043	-0.0103	0.0036	-0.0159	0.0175	-0.0092	0.0083	0.0174	0.0046	-0.0147	0.0042
ψ_{21}	-0.0040	0.0043	0.0036	0.0020	0.0149	0.0016	-0.0032	0.0017	0.0088	0.0011	-0.0033	0.0046	0.0035	0.0020	0.0149	0.0016	-0.0032	0.0016
ψ_{22}	-0.0004	0.0040	-0.0016	0.0019	-0.0069	0.0014	-0.0022	0.0011	-0.0066	0.0008	0.0000	0.0040	-0.0016	0.0019	-0.0069	0.0013	-0.0023	0.0011
ψ_{23}	0.0068	0.0039	0.0031	0.0019	-0.0010	0.0017	0.0102	0.0019	0.0058	0.0011	0.0061	0.0043	0.0031	0.0019	-0.0010	0.0017	0.0102	0.0019
ψ_{31}	-0.0313	0.0199	-0.0368	0.0144	-0.0175	0.0042	-0.0121	0.0043	-0.0324	0.0038	-0.0294	0.0229	-0.0365	0.0138	-0.0175	0.0045	-0.0122	0.0040
ψ_{32}	-0.0069	0.0124	0.0185	0.0106	0.0197	0.0029	0.0010	0.0031	0.0125	0.0025	-0.0037	0.0154	0.0181	0.0102	0.0197	0.0028	0.0011	0.0031
ψ_{33}	0.0176	0.0139	0.0059	0.0084	-0.0164	0.0048	0.0044	0.0045	0.0045	0.0038	0.0098	0.0173	0.0061	0.0083	-0.0164	0.0049	0.0044	0.0044
ϕ_1	0.0029	0.0008	0.0018	0.0004	0.0001	0.0002	-0.0045	0.0006	-0.0007	0.0003	0.0049	0.0012	0.0017	0.0004	0.0002	0.0002	-0.0046	0.0007
ϕ_2	0.0002	0.0002	-0.0004	0.0001	0.0007	0.0001	0.0006	0.0001	0.0004	0.0000	0.0002	0.0002	-0.0004	0.0001	0.0006	0.0001	0.0006	0.0001
ϕ_3	-0.0031	0.0008	-0.0014	0.0004	-0.0008	0.0002	0.0039	0.0006	0.0003	0.0003	-0.0051	0.0011	-0.0014	0.0004	-0.0008	0.0002	0.0040	0.0006
φ_1	-0.0005	0.0004	-0.0011	0.0002	-0.0002	0.0001	-0.0017	0.0002	-0.0005	0.0001	-0.0002	0.0005	-0.0011	0.0002	-0.0002	0.0001	-0.0017	0.0002
φ_2	-0.0008	0.0005	0.0009	0.0002	-0.0002	0.0001	-0.0002	0.0002	-0.0002	0.0001	-0.0015	0.0005	0.0009	0.0002	-0.0002	0.0001	-0.0002	0.0002
φ_3	0.0009	0.0005	0.0001	0.0002	-0.0002	0.0001	0.0015	0.0003	0.0003	0.0001	0.0013	0.0006	0.0001	0.0002	-0.0002	0.0001	0.0015	0.0003
UNIT									-0.0102	0.0035								
LIMITED									0.0003	0.0027								
STATEWIDE									0.0071	0.0023								

functions in (2.7) by indicator functions. This comparison is interesting because the estimation of the latter function is equivalent to estimating four separate fixed-coefficient translog cost functions with one for each branch bank regime. Parameter estimates and standard errors associated with the former function are reported in Panel A of Table 3 (as discussed previously), while those associated with the later function are reported in panel C of the same table. A comparison of these two panels reveals that parameter estimates from both functions are rather close for all four banking regimes with the exception of unit banking regime, further confirming that branch banking regime has a strong impact on the production technology of the commercial banks. Besides, we also find that standard errors from the case where λ is optimally selected are generally smaller than their counterparts from the case where $\lambda = 0$, because the former case allows borrowing information across branch banking regimes.

Table 4: Results on Return to Scales (RTS)

	Panel A		Panel B: Average RTS under Different Banking Regimes							
	Overall Average RTS		UNIT		LIMITED		STATEWIDE		NATIONWIDE	
Year	RTS	std	RTS	std	RTS	std	RTS	std	RTS	std
1986	1.0526	0.0060	1.0995	0.0228	1.0407	0.0055	1.0361	0.0050	NA	NA
1987	1.0528	0.0059	1.0995	0.0226	1.0405	0.0055	1.0377	0.0050	NA	NA
1988	1.0458	0.0043	1.0962	0.0205	1.0410	0.0053	1.0413	0.0050	NA	NA
1989	1.0469	0.0038	1.0986	0.0198	1.0383	0.0052	1.0492	0.0050	NA	NA
1990	1.0503	0.0036	1.1022	0.0197	1.0405	0.0052	1.0522	0.0050	NA	NA
1991	1.0508	0.0038	1.0981	0.0218	1.0400	0.0053	1.0573	0.0052	NA	NA
1992	1.0531	0.0040	NA	NA	1.0395	0.0054	1.0594	0.0052	NA	NA
1993	1.0533	0.0040	NA	NA	1.0380	0.0055	1.0605	0.0053	NA	NA
1994	1.0559	0.0043	NA	NA	1.0332	0.0056	1.0621	0.0053	NA	NA
1995	1.0563	0.0042	NA	NA	1.0323	0.0054	1.0629	0.0052	NA	NA
1996	1.0616	0.0043	NA	NA	1.0365	0.0054	1.0685	0.0052	NA	NA
1997	1.0649	0.0044	NA	NA	1.0391	0.0054	1.0709	0.0052	NA	NA
1998	1.0564	0.0065	NA	NA	NA	NA	1.0818	0.0059	1.0550	0.0069
1999	1.0585	0.0064	NA	NA	NA	NA	1.0854	0.0059	1.0569	0.0068
2000	1.0590	0.0064	NA	NA	NA	NA	1.0872	0.0058	1.0577	0.0067
2001	1.0621	0.0064	NA	NA	NA	NA	1.0912	0.0058	1.0607	0.0067
2002	1.0644	0.0067	NA	NA	NA	NA	NA	NA	1.0644	0.0067
2003	1.0667	0.0067	NA	NA	NA	NA	NA	NA	1.0667	0.0067
2004	1.0682	0.0066	NA	NA	NA	NA	NA	NA	1.0682	0.0066
2005	1.0688	0.0066	NA	NA	NA	NA	NA	NA	1.0688	0.0066
Average	1.0576	0.0034	1.0995	0.0213	1.0390	0.0052	1.0605	0.0051	1.0625	0.0067

Having established the superority of the varying-coefficient translog cost function over the fully parametric translog cost function, in what follows we focus on empirical results from the former function. Panel A of Table 4 presents the annual average returns to scale (RTS) estimate for each year, obtained by averaging over all sampled banks in that year. As can be seen, it is greater than one for all years, ranging from 1.037 to 1.056, suggesting that on average the commercial banks exhibit increasing returns to scale. This finding is consistent with Wheelock and Wilson (2012), who, using a non-parametric local-linear estimator to estimate the cost relationship for commercial banks in the U.S. over the period 1984-2006, find that U.S. banks operated under increasing returns to scale.

It is also of interest to compare the estimates of RTS across different regimes. For this purpose, we calculate the average RTS for each banking regime in each year by averaging within each regime in that year. The results are reported in Panel B, Table 4, where "NA" indicates that the corresponding policy regime doesn't exist or expires in that year. We see that average RTS is generally higher in more regulated states than in less regulated states for a given year. Taking 1986 for example, average RTS is 1.0995 for unit banking states, as compared to 1.0407 for limited branching states and 1.0361 for statewide branching states. This result suggest that banks in more regulated states are forced to operate at scales further below their optimal scales than those in less regulated states. It is worth noting at this point that optimal scales in less regulated states are much higher than those in more regulated states. To illustrate this point, we calculate the optimal scale for each banking regime in 1986 by averaging total assets across banks under that regime that face constant returns to scale. Our result shows that the optimal scale for statewide branching states is $1.177 billion, as compared to $1 million for unit banking states and 4 million for limited branching states. This result suggests that geographical deregulation greatly changes banking production technology in the U.S. Another interesting finding that emerges from Table 4 is that average RTS have increased over time for both statewide and national branching regimes. A possible explanation is that as banks grow bigger under less regulated regimes, they are more likely to afford new technologies. The adoption of new technologies further increases the banks' optimal scales over time, which results in higher RTS for given bundles of inputs.

Table 5: Returns To Scale at Individual Bank Level

Year	DRS	CRS	IRS
1986	13.52%	11.59%	74.89%
1987	11.59%	12.23%	76.18%
1988	13.09%	7.94%	78.97%
1989	13.09%	5.36%	81.55%
1990	9.23%	3.86%	86.91%
1991	5.79%	3.65%	90.56%
1992	5.36%	3.00%	91.63%
1993	5.79%	2.58%	91.63%
1994	4.51%	2.15%	93.35%
1995	4.72%	1.50%	93.78%
1996	3.65%	1.50%	94.85%
1997	3.65%	0.43%	95.92%
1998	2.58%	2.58%	94.85%
1999	2.58%	2.36%	95.06%
2000	2.79%	1.50%	95.71%
2001	1.93%	2.58%	95.49%
2002	1.93%	1.29%	96.78%
2003	1.93%	0.86%	97.21%
2004	2.15%	0.86%	97.00%
2005	2.15%	1.29%	96.57%
Average	5.34%	3.36%	91.30%

DRS: decreasing returns to scale
CRS: constant returns to scale
IRS: increasing returns to scale

In addition to the annual average RTS estimates, we are also interested in RTS estimates

at individual bank level. We compute the percentage of banks facing increasing, constant, or decreasing returns to scale for each year. This computation is performed by counting the number of cases where the 95% credible intervals are strictly less than 1.0 (indicating decreasing returns to scale, i.e., DRS), contain 1.0 (indicating constant returns to scale, i.e., CRS), or strictly greater than 1.0 (indicating increasing returns to scale, i.e., IRS). The results are presented in Table 5. Two findings emerge from this table. First, on average the majority (91.30%) of the banks face increasing returns to scale, a small percentage (5.34%) face decreasing returns to scale, and an even smaller percentage (3.36%) face constant returns to scale. Second, the percentage of banks facing increasing returns to scale shows a "first increase and then stabilize" pattern, the percentage of banks facing decreasing returns to scale shows a "first decrease and then stabilize" pattern, and the percentage of banks facing constant returns to scale also shows a "first decrease and then stabilize" pattern. Specifically, the percentage of banks facing increasing returns to scale increases markedly from 74.89% in 1986 to 96.76% in 2002 and then stabilizes at around that level for the rest of the sample period; the percentage of banks facing decreasing returns to scale decreases noticeably from 13.52% in 1986 to 1.93% in 2001 and then stabilizes at around that level afterwards (with the exception of the last year when the percentage goes up to 8.22%); and the percentage of banks facing constant returns to scale falls consistently from 11.59% in 1986 to 1.29% in 2002 stabilizes at around that level afterwards. This result is consistent with our previous discussion that both geographical deregulation and subsequent technological adoptions increase the bank's optimal scales over time, leaving more and more banks operating under increasing returns to scale.

5 Conclusion

In this paper, we extend Li et al. (2013)'s cross-sectional varying-coefficient model to a panel data context, where fixed effects are included to allow for correlation between individual unobserved heterogeneity and the regressors. In dealing with the fixed effects, we do not impose any identification restriction as done in previous studies. Instead, we take advantage of the fact that our covariates are categorical, and use a modified within transformation. We show the exact asymptotic properties of our estimator for the relevant covariate case and the irrelevant covariate case. To avoid including spurious regressors in our panel data varying-coefficient model, we also provide a variable selection procedure for selecting significant regressors. We further conduct a Monte Carlo study to investigate the finite sample properties of our estimator.

Finally, we show how our model and methodology can be used by analyzing the effects of state-level banking regulations on the returns to scale of commercial banks in the U.S. over the period 1986-2005. Specifically, we estimate a varying-coefficient translog cost function, where

branch banking regime is used as a covariate of the varying coefficient. We compare this cost function with a fully parametric cost function where branch banking regimes are treated as binary variable. Our tests reject the latter cost function in favor of the former one. Our empirical results from the varying-coefficient translog cost function show that returns to scale is higher in more regulated states than in less regulated states. Our results also indicate that the majority of the banks face increasing returns to scale, a small percentage face decreasing returns to scale, and an even smaller percentage face constant returns to scale.

Appendix A: Assumptions with Discussions

Assumption A:

1. $\beta(z)$ is not a constant function with respect to z and uniformly bounded on the support $\mathcal{D}$ of z, i.e. $\max_{z\in\mathcal{D}} \|\beta(z)\| < \infty$. For $z = (z_1, \ldots, z_r)' \in \mathcal{D}$, z_s takes c_s different integer values in $\{0, 1, \ldots, c_s - 1\}$ and $c_s \geq 2$ for $s = 1, \ldots, r$. Moreover, r is finite and $\max_{1\leq s\leq r} c_s < \infty$. Let $p(z) = \Pr(Z_{it} = z) > 0$ for $\forall z \in \mathcal{D}$.

2. Suppose that Z_{it} is independent and identically distributed (i.i.d.) over i and t. Moreover, $\{X_{i1}, \ldots, X_{iT}\}$ is independent across i.

3. $\forall z \in \mathcal{D}$, $i = 1, \ldots, N$ and $t = 1, \ldots, T$, $E[X_{it}|Z_{it} = z] = \mu_X(z)$, $E[X_{it}X_{it}'|Z_{it} = z] = \Sigma_X(z)$, where $\|\mu_X(z)\|$ and $\|\Sigma_X(z)\|$ are uniformly bounded in z. X_{it} is independent of Z_{js} for $(i,t) \neq (j,s)$. $X_t = (X_{1t}, \ldots, X_{Nt})'$ is strictly stationary and α-mixing with $E\|X_{it}\|^4 < \infty$. Suppose the following results hold:

$$\max_{1\leq i\leq N} \max_{z\in\mathcal{D},\ \lambda\in[0,1]^r} \left| \frac{1}{T}\sum_{s=1}^{T} X_{is}L^p(Z_{is}, z, \lambda) - \Delta_2(z,\lambda) \right| \to_P 0,$$
$$\max_{1\leq i\leq N} \max_{z\in\mathcal{D},\ \lambda\in[0,1]^r} \left| \frac{1}{T}\sum_{s=1}^{T} X_{is}'\beta(Z_{is})L^p(Z_{is}, z, \lambda) - \Delta_{2\beta}(z,\lambda) \right| \to_P 0,$$

   where $\Delta_2(z,\lambda) = E[X_{it}L^p(Z_{it}, z, \lambda)|z, \lambda]$ and $\Delta_{2\beta}(z,\lambda) = E[X_{it}\beta(Z_{it})L^p(Z_{it}, z, \lambda)|z, \lambda]$.

4. $u_t = (u_{1t}, \ldots, u_{Nt})'$ is strictly stationary and α-mixing. Denote $\mathcal{X} = \{(X_{js}, Z_{js}), 1 \leq j \leq N, 1 \leq s \leq T\}$. $E[u_{it}|\mathcal{X}] = 0$ and $E[u_{it}^2|\mathcal{X}] = \sigma_u^2$ for $1 \leq i \leq N$ and $1 \leq t \leq T$. Conditional on $\mathcal{X}$, let $\alpha_{u,ij}(|t-s|)$ denote the α-mixing coefficient between u_{it} and u_{js}, such that for a $\delta_2 > 0$, $\sum_{i=1}^N \sum_{j=1}^N \sum_{t=1}^T \sum_{s=1}^T (\alpha_{u,ij}(|t-s|))^{\frac{\delta_2}{4+\delta_2}} = O(NT)$. For the same δ_2, $E[|u_{it}|^{4+\delta_2}|\mathcal{X}] \leq c_1 < \infty$ uniformly, where c_1 is a constant. For the time dimension, let $\max_{1\leq i\leq N} \sum_{t=1}^T \sum_{s=1}^T |E[u_{it}u_{is}|\mathcal{X}]| = O(T)$.

5. $\lambda_s \in [0,1]$ for $s = 1, \ldots, r$. Define

$$CV_0(\lambda) = \sum_{z\in\mathcal{D}} p(z)(\beta(z) - \eta(z,\lambda))'\Omega(z,\lambda)(\beta(z) - \eta(z,\lambda)),$$

$$+\sum_{z\in\mathcal{D}} p(z)\left(\Delta_{3\beta}(z,\lambda)-\Delta_3(z,\lambda)'\beta(z)\right)^2$$
$$+2\sum_{z\in\mathcal{D}} p(z)(\mu_X(z)-\Delta_3(z,\lambda))'(\beta(z)-\eta(z,\lambda))\left(\Delta_{3\beta}(z,\lambda)-\Delta_3(z,\lambda)'\beta(z)\right),$$

where

$$\begin{aligned}
\Delta_1(z,\lambda) &= E[L^p(Z_{it},z,\lambda)|z,\lambda],\\
\Delta_3(z,\lambda) &= \Delta_2(z,\lambda)/\Delta_1(z,\lambda),\\
\Delta_{3\beta}(z,\lambda) &= \Delta_{2\beta}(z,\lambda)/\Delta_1(z,\lambda),\\
\Omega(z,\lambda) &= \Sigma_X(z)+\Delta_3(z,\lambda)\Delta_3(z,\lambda)'-\Delta_3(z,\lambda)\mu_X(z)'-\mu_X(z)\Delta_3(z,\lambda),\\
\Sigma_{XX}(z,\lambda) &= E\left[\Omega(Z_{it},\lambda)L(Z_{it},z,\lambda)|z,\lambda\right],\\
\Sigma_{XX\beta}(z,\lambda) &= E\left[\Omega(Z_{it},\lambda)\beta(Z_{it})L(Z_{it},z,\lambda)|z,\lambda\right],\\
\eta(z,\lambda) &= \Sigma_{XX}^{-1}(z,\lambda)\Sigma_{XX\beta}(z,\lambda).
\end{aligned}$$

$CV_0(\lambda)=0$ holds only when $\lambda=(\lambda_1,\ldots,\lambda_r)'=0_{r\times 1}$.

Assumption A.1 is standard and the same as Assumption 1.1 of Li et al. (2013). In order to deal with the case where the cardinality of $\mathcal{D}$ is infinite, one workaround is as follows.

- Suppose that $r=1$. $Z_{it}\in\{0,1,2,\ldots,\nu(N,T)-1\}$, where $\nu(N,T)\to\infty$ and $\nu(N,T)/(NT)\to c$ for $0\le c<\infty$ as $(N,T)\to(\infty,\infty)$. In this case, the following model can be considered

$$Y_{it}=X_{it}'\beta(Z_{it}/\nu(N,T))+w_i+u_{it},\quad i=1,\ldots,N \text{ and } t=1,\ldots,T \tag{A.1}$$

  Here we can treat $\beta(\cdot)$ as a function with continuous covariates. (A.1) then becomes the model proposed by Sun et al. (2009). This normalization technique is similar to the one employed by Cai (2007) and Chen et al. (2012*b*) in dealing with time varying-coefficient models.

Although optimal bandwidth selection has been fully investigated in an i.i.d. cross-sectional setting in the literature (see Li and Racine (2010) and Li et al. (2013) for details), little work has been done for panel data models (c.f. Sun et al. (2009) and Chen et al. (2012*b*)). For example, optimal bandwidth selection remains an unresolved issue for the panel data model considered in Sun et al. (2009). This issue is even more daunting for varying-coefficient panel data models with mixed covariates.

Assumption A.2 is standard in the literature (c.f. Assumption A1 of Cai and Li (2008); Assumption 1 of Sun et al. (2009); Assumption A1 of Chen et al. (2013); Assumption 1.1 of Li et al. (2013) and Assumption 3.1 of Rodriguez-Poo and Soberon (2014)). Due to the use of conditional expectation, we are not able to impose certain weak cross-sectional dependence on X_{it} and Z_{it} as we do for u_{it}. When all elements of Z_{it} are continuous, one certainly can allow Z_{it} to be α-mixing in the same way as Z_{it} can be assumed to be alpha-maxing as in Sun et al. (2009) and Rodriguez-Poo and Soberon (2014). However, since Z_{it} is purely discrete in this study, we assume that Z_{it} is independent over i and t. In the literature of time series, Andrews (1984) has shown that even the process $x_{t+1}=0.5x_t+\varepsilon_t$ is not

α-mixing when ε_t has a binomial distribution. More details and relevant discussions can be found in Fan and Yao (2003). Thus, we believe Assumption 2 is reasonable.

Alternatively, we can assume that $\mathcal{Z} = \{Z_{it}, 1 \leq i \leq T, 1 \leq t \leq T\}$ are pre-determined. Therefore, conditional on $\mathcal{Z}$, we can impose certain weak cross-sectional dependence on X_{it}. Accordingly, we need to adjust our assumptions and analysis, but the consistency and asymptotic normality remain valid. In an even more extreme case, we can assume that all Z_{it}'s are pre-determined to be z and $\lambda = 0$ (or 1). Then the model (2.1) will reduce to the classic panel data model with fixed effects. In this extreme case, the assumptions and analysis can be significantly simplified.

By construction of $L^p(Z_{it}, z, \lambda)$, it is easy to show

$$\begin{aligned}
&\max_{z\in\mathcal{D},\ \lambda\in[0,1]^r} \left| \frac{1}{T}\sum_{s=1}^{T} X_{is} L^p(Z_{is}, z, \lambda) - \Delta_2(z,\lambda) \right| \to_P 0, \\
&\max_{z\in\mathcal{D},\ \lambda\in[0,1]^r} \left| \frac{1}{T}\sum_{s=1}^{T} X_{is}'\beta(Z_{is}) L^p(Z_{is}, z, \lambda) - \Delta_{2\beta}(z,\lambda) \right| \to_P 0.
\end{aligned} \tag{A.2}$$

Due to the within transformation, we have to assume that (A.2) holds uniformly across i in Assumption A.3, which is in the same spirit of Assumption A1 of Su et al. (2014), Assumption A1 of Chen et al. (2013) and Assumption C of Bai (2009). Below we provide an example to demonstrate why this assumption is reasonable.

- For simplicity, suppose that all variables are scalars and consider the data generating process as $X_{it} = H_{it} + \varepsilon_{it}$ and $\varepsilon_{it} \sim N(Z_{it}, Z_{it} + 1)$, where $H_{it} = 0.5 \cdot H_{i,t-1} + v_{it}$ is an AR(1) process; $v_{it} \sim N(0,1)$ is i.i.d. over i and t; $Z_{it} = 0$ with probability 0.4 and $Z_{it} = 1$ with probability 0.6. In this example, the requirements of Assumption A.3 are certainly satisfied. Moreover, this example particularly implies that the choice of Z_{it} affects only the value of X_{it}, but does not affect the value of X_{js} for $(j,s) \neq (i,t)$.

Assumption A.4 is the same as that in Arellano (1987) and in the same spirit as Assumption C of Bai (2009), Assumptions A2 and A4 of Chen et al. (2012*b*) and Assumption 1 of Dong et al. (2015). Two examples are given below to demonstrate this assumption is reasonable:

- It can be easily seen that Assumption A.4 holds if u_{it} is i.i.d. over i and t.

- We now use a factor model structure as an example to show that Assumption A.4 is verifiable. Suppose that $u_{it} = \gamma_i f_t + \varepsilon_{it}$, where all variables are scalars and ε_{it} is i.i.d. over i and t with mean zero. Simple algebra shows that the coefficient $\alpha_{u,ij}(|t-s|)$ reduces to $\alpha_{ij} \cdot b(|t-s|)$, in which $\alpha_{ij} = E[\gamma_i \gamma_j]$ and $b(|t-s|)$ is the α-mixing coefficient of the factor time series $\{f_1, \ldots, f_T\}$. If f_t is a strictly stationary α-mixing process and α_{ij} converges to 0 at a certain rate as $|i-j|$ increases, Assumption A.4 can easily be verified. More details and useful empirical examples can be found in Chen et al. (2012*b*).

Moreover, if we assume that every variable is i.i.d. over i and t (alternatively, we can employ a random effects setting without using the within transformation), we can allow for heteroskedasticity

by assuming $E[u_{it}^2|X_{it}, Z_{it}] = \sigma_u(X_{it}, Z_{it})$ (c.f. Li et al. (2013)). However, when deriving asymptotic results in a panel data setting with serial coloration and cross-sectional dependence, one normally deals with $E[u_{it}u_{js}X_{it}X_{js}'|Z_{it}, Z_{js}, X_{it}, X_{js}]$. In this case, we could assume that $\nu(X_{it}, X_{js}, Z_{it}, Z_{js}) = E[u_{it}u_{js}X_{it}X_{js}'|Z_{it}, Z_{js}, X_{it}, X_{js}]$ and further impose restrictions on $\nu(X_{it}, X_{js}, Z_{it}, Z_{js})$. However, this would make our analysis much more complicated. In addition, heteroskedasticity is not the main focus of this paper. We would like to point out that one way of imposing both heteroskedasticity and cross-sectional dependence is to follow Robinson (2011) and Lee and Robinson (2013). More details are given as follows.

- Assume that $u_{it} = \sigma(X_{it}, Z_{it})e_{it}$ and $e_{it} = \sum_{h=1}^{\infty}\sum_{l=0}^{\infty} a_{ihl}\varepsilon_{h,t-l}$, where $\varepsilon_{i,j}$ is i.i.d. with mean 0 and variance 1 over (i, j) and a_{ihl}'s are constants. Simple algebra shows $E[u_{it}^2|X_{it}, Z_{it}] = \sigma^2(X_{it}, Z_{it})\sum_{h=1}^{\infty}\sum_{l=0}^{\infty} a_{ihl}$. When (X_{it}, Z_{it}) is i.i.d. across i and $\sum_{h=1}^{\infty}\sum_{l=0}^{\infty} a_{ihl}$ is the same for all $1 \leq i \leq N$, we can show that the error terms are i.i.d. across i. Otherwise, heteroskedasticity will occur. With this setting, more restrictions are needed for developing asymptotic results. Robinson (2011) and Lee and Robinson (2013) have used this technique to revisit some cross-sectional data models. However, more work will be needed to extend this technique to panel data models.

Assumption A.5 is a panel data version of Assumption 2 of Li et al. (2013) and ensures that $CV_0(\lambda)$ is uniquely minimized at 0. By Theorem 2.1 of Newey and McFadden (1994), this assumption implies that $\hat{\lambda}$ obtained by minimizing (2.8) converges to $0_{r\times 1}$. In order to further explain why this assumption is reasonable, we expand the product form of $L(Z_{it}, z, \lambda)$ as a summation form:

$$\begin{aligned} L(Z_{it}, z, \lambda) &= \prod_{s=1}^{r}\{1(Z_{it,s} = z_s) + \lambda_s 1(Z_{it,s} \neq z_s)\} \\ &= \prod_{s=1}^{r} 1(Z_{it,s} = z_s) + \sum_{s=1}^{r}\lambda_s 1_{s,Z_{it}=z} + \cdots + \prod_{s=1}^{r}\lambda_s 1(Z_{it,s} \neq z_s) \\ &= 1(Z_{it} = z) + \sum_{s=1}^{r}\lambda_s 1_{s,Z_{it}=z} + \cdots + \prod_{s=1}^{r}\lambda_s 1(Z_{it,s} \neq z_s), \end{aligned}$$

where $1_{s,Z_{it}=z} = 1(Z_{it,s} \neq z_s)\prod_{n=1,n\neq s}^{r} 1(Z_{it,n} = z_n)$ for simplicity. Then, we can further rewrite the following expectations:

$$\begin{aligned} \Delta_1(z, \lambda) &= E[L^p(Z_{it}, z, \lambda)|z, \lambda] = p(z) + \delta_1(z, \lambda), \\ \Delta_2(z, \lambda) &= E[X_{it}L^p(Z_{it}, z, \lambda)|z, \lambda] = p(z)\mu_X(z) + \delta_2(z, \lambda), \\ \Delta_{2\beta}(z, \lambda) &= E[X_{it}\beta(Z_{it})L^p(Z_{is}, z, \lambda)|z, \lambda] = p(z)\mu_X(z)'\beta(z) + \delta_3(z, \lambda), \end{aligned} \tag{A.3}$$

where $\delta_1(z, \lambda)$, $\delta_2(z, \lambda)$ and $\delta_{2\beta}(z, \lambda)$ can be expressed as

$$\delta_1(z, \lambda) = \lambda\delta_1^*(z, \lambda), \quad \delta_2(z, \lambda) = \lambda\delta_2^*(z, \lambda), \quad \delta_3(z, \lambda) = \lambda\delta_3^*(z, \lambda).$$

Thus, it is easy to know that $\delta_1(z,0) = \delta_2(z,0) = \delta_{2\beta}(z,0) = 0$. Moreover, when $\lambda = 0$, $\Delta_3(z,\lambda)$ and $\Delta_{3\beta}(z,\lambda)$ will reduce to $\mu_X(z)$ and $\mu_X(z)'\beta(z)$ respectively.

Before proceeding to Assumption B, denote

$$\begin{aligned}
&p(z) = p(\bar{z}) \cdot p(\tilde{z}), \quad p(\bar{z}) = \Pr(\bar{Z}_{it} = \bar{z}), \quad p(\tilde{z}) = \Pr(\tilde{Z}_{it} = \tilde{z}),\\
&L(Z_{it}, z, \lambda) = L(\bar{Z}_{it}, \bar{z}, \bar{\lambda}) \cdot L(\tilde{Z}_{it}, \tilde{z}, \tilde{\lambda}),\\
&L(\bar{Z}_{it}, \bar{z}, \bar{\lambda}) = \prod_{s=1}^{r_1} \lambda_s^{1(Z_{it,s} \neq z_s)}, \quad L(\tilde{Z}_{it}, \tilde{z}, \tilde{\lambda}) = \prod_{s=r_1+1}^{r} \lambda_s^{1(Z_{it,s} \neq z_s)},
\end{aligned}$$

where $\bar{z} = (z_1, \ldots, z_{r_1})'$ and $\tilde{z} = (z_{r_1+1}, \ldots, z_r)'$. Also, $\beta(z)$, $\mu_X(z)$, $\Sigma_X(z)$, $\eta(z,\lambda)$, $\Delta_3(z,\lambda)$, $\Delta_{3\beta}(z,\lambda)$ and $\Omega(z,\lambda)$ denoted in Assumption A.5 will respectively reduce to $\beta(\bar{z})$, $\mu_X(\bar{z})$, $\Sigma_X(\bar{z})$, $\eta(\bar{z},\bar{\lambda})$, $\Delta_3(\bar{z},\bar{\lambda})$, $\Delta_{3\beta}(\bar{z},\bar{\lambda})$ and $\Omega(\bar{z},\bar{\lambda})$ with $\bar{z} \in \bar{\mathcal{D}}$ for the irrelevant covariate case.

Assumption B:

1. The irrelevant covariates $\tilde{Z}_{it}$'s for $i = 1, \ldots, N$ and $t = 1, \ldots, T$ are independent of all the other variables.

2. $\lambda_s \in [0,1]$ for $s = 1, \ldots, r_1, r_1+1, \ldots, r$. Define

$$\begin{aligned}
CV_0^*(\bar{\lambda}) = &\sum_{\bar{z} \in \bar{\mathcal{D}}} p(\bar{z})(\beta(\bar{z}) - \eta(\bar{z},\bar{\lambda}))'\Omega(\bar{z},\bar{\lambda})(\beta(\bar{z}) - \eta(\bar{z},\bar{\lambda})),\\
&+ \sum_{\bar{z} \in \bar{\mathcal{D}}} p(\bar{z}) \left(\Delta_{3\beta}(\bar{z},\bar{\lambda}) - \Delta_3(\bar{z},\bar{\lambda})'\beta(\bar{z})\right)^2\\
&+2 \sum_{\bar{z} \in \bar{\mathcal{D}}} p(\bar{z})(\mu_X(\bar{z}) - \Delta_3(\bar{z},\bar{\lambda}))'(\beta(\bar{z}) - \eta(\bar{z},\bar{\lambda})) \left(\Delta_{3\beta}(\bar{z},\bar{\lambda}) - \Delta_3(\bar{z},\bar{\lambda})'\beta(\bar{z})\right).
\end{aligned}$$

   $CV_0^*(\bar{\lambda}) = 0$ holds only when $\bar{\lambda} = (\lambda_1, \ldots, \lambda_{r_1})' = 0_{r_1 \times 1}$.

Assumption B is a panel data version of Assumption 3 of Li et al. (2013). Ideally, one can assume conditional independence instead of independence in Assumption B.1. However, the former is troublesome even for i.i.d. data (Li et al., 2013). In view of this, we adopt the assumption of unconditional independence in this paper. All discussions for Assumption A.5 also apply to Assumption B.2.

Assumption C:

1. For a random variable $\bar{Z}_{it} \in \bar{\mathcal{D}}$ and $\beta(\bar{Z}_{it}) = (\beta_1(\bar{Z}_{it}), \ldots, \beta_q(\bar{Z}_{it}))'$, suppose there exists a positive integer $1 \leq q^* \leq q$ such that $0 < E|\beta_j(\bar{Z}_{it})|^2 < \infty$ for $j = 1, \ldots, q^*$ and $E|\beta_j(\bar{Z}_{it})|^2 = 0$ for $j = q^*+1, \ldots, q$.

2. For $\bar{z} \in \bar{\mathcal{D}}$, let $\Sigma_1(\bar{z}) = \Sigma_X(\bar{z}) - \mu_X(\bar{z})\mu_X(\bar{z})'$. Suppose that

$$0 < \rho_1 \leq \min_{\bar{z} \in \bar{\mathcal{D}}} \rho_{min}(\Sigma_1(\bar{z})) \leq \max_{\bar{z} \in \bar{\mathcal{D}}} \rho_{max}(\Sigma_1(\bar{z})) \leq \rho_2 < \infty,$$

   where $\rho_{min}(\Sigma_1(\bar{z}))$ and $\rho_{max}(\Sigma_1(\bar{z}))$ denote the minimum and maximum eigenvalues of $\Sigma_1(\bar{z})$ respectively.

Assumption C.1 defines the sparsity structure for the coefficient function. It indicates that one element of the coefficient function is removed only when it does not have an impact on all $\beta(\bar{z}^1), \ldots, \beta(\bar{z}^m)$. Note that $\Sigma_1(\bar{z})$ is essentially a covariance matrix, implying Assumption C.2 is reasonable.

References

Aitchison, J. and Aitken, C. (1976), 'Multivariate binary discrimination by the kernel method', *Biometrika* **63**(3), 413–420.

Andrews, D. W. K. (1984), 'Nonstrong mixing autoregressive processes', *Journal of Applied Probability* **21**(4), 930–934.

Andrews, D. W. K. (2005), 'Cross–section regression with common shocks', *Econometrica* **73**(5), 1551–1585.

Arellano, M. (1987), 'Computing robust standard errors for within-groups estimators', *Oxford bulletin of economics and statistics* **49**(4), 432–434.

Bai, J. (2009), 'Panel data models with interactive fixed effects', *Econometrica* **77**(4), 1229–1279.

Berger, A. and Mester, L. (2003), 'Explaining the dramatic changes in performance of u.s. banks: technical change, deregulation, and dynamic changes in competition', *Journal of Financial Intermediation* pp. 57–95.

Blundell, R. and Bond, S. (1998), 'Initial conditions and moment restrictions in dynamic panel data models', *Journal of Econometrics* **87**(1), 115–143.

Cai, Z. (2007), 'Trending time-varying coefficient time series models with serially correlated errors', *Journal of Econometrics* **136**(1), 163–188.

Cai, Z. and Li, Q. (2008), 'Nonparametric estimation of varying coefficient dynamic panel data models', *Econometric Theory* **24**(5), 1321–1342.

Cai, Z., Li, Q. and Park, J. (2009), 'Functional–coefficient models for nonstationary time series data', *Journal of Econometrics* **148**(2), 101–113.

Chen, J., Gao, J. and Li, D. (2012*a*), 'A new diagnostic test for cross–section uncorrelatedness in nonparametric panel data models', *Econometric Theory* **28**(5), 1–20.

Chen, J., Gao, J. and Li, D. (2012*b*), 'Semiparametric trending panel data models with cross–sectional dependence', *Journal of Econometrics* **171**(1), 71–85.

Chen, J., Gao, J. and Li, D. (2013), 'Estimation in partially linear single–index panel data models with fixed effects', *Journal of Business and Economic Statistics* **31**(3), 315–330.

Christensen, L. R., Jorgenson, D. W. and Lau, L. J. (1975), 'Transcendental logarithmic utility functions', *The American Economic Review* **65**(3), 367–383.

Diewert, W. E. and Wales, T. J. (1987), 'Flexible functional forms and global curvature conditions', *Econometrica* **55**(1), 43–68.

Dong, C., Gao, J. and Peng, B. (2015), 'Semiparametric single-index panel data model with cross-sectional dependence', *Journal of Econometrics* **188**(1), 301–312.

Fan, J. and Yao, Q. (2003), *Nonlinear Time Series: Nonparametric and Parametric Methods*, Springer-Verlag.

Feng, G. and Serletis, A. (2008), 'Productivity trends in u.s. manufacturing: Evidence from the nq and aim cost functions', *Journal of Econometrics* **142**(1), 281–311.

Gao, J. and Phillips, P. C. B. (2013*a*), 'Functional coefficient nonstationary regression', *Cowles Foundation Discussion Paper NO. 1911* .

Gao, J. and Phillips, P. C. B. (2013*b*), 'Semiparametric estimation in triangular system equations with nonstationarity', *Journal of Econometrics* **176**(1), 59–79.

Griffiths, W. E., O'Donnell, C. J. and Cruz, A. T. (2000), 'Imposing regularity conditions on a system of cost and cost–share equations: A bayesian approach', *Australian Journal of Agricultural and Resource Economics* **44**(1), 107–127.

Hsiao, C. (2003), *Analysis of Panel Data*, Cambridge University Press.

Jayaratne, J. and Strahan, P. E. (1997), 'The benefits of branching deregulation', *Economic Policy Review* **3**(4), 13–29.

Lee, J. and Robinson, P. (2013), 'Series estimation under cross-sectional dependence', *http://sticerd.lse.ac.uk/dps/em/em570.pdf* .

Li, Q., Huang, C. J., Li, D. and F, T. (2002), 'Semiparametric smooth coefficient models', *Journal of Business and Economic Statistics* **20**(3), 412–422.

Li, Q., Ouyang, D. and Racine, J. S. (2013), 'Categorical semiparametrics varying–coefficient models', *Journal of Applied Econometrics* **28**(4), 551–579.

Li, Q. and Racine, J. S. (2010), 'Smooth varying-coefficient estimation and inference for qualitative and quantitative data', *Econometric Theory* **26**(6), 1607–1637.

Lounici, K., Pontil, M., Van De Geer, S. and Tsybakov, A. B. (2011), 'Oracle inequalities and optimal inference under group sparsity', *Annals of Statistics* **39**(4), 2164–2204.

Mason, J. E. (2013), *The Transformation of Commercial Banking in the United States 1956–1991*, United Kingdom: Taylor and Francis.

Mester, L. J. (2005), 'Optimal industrial structure in banking', *Working Papers 08-2, Federal Reserve Bank of Philadelphia* .

Newey, W. K. and McFadden, D. (1994), 'Large sample estimation and hypothesis testing', *Handbook of Econometrics* pp. 2113–2245.

Pesaran, M. H. (2006), 'Estimation and inference in large heterogeneous panels with a multifactor error structure', *Econometrica* **74**(4), 967–1012.

Pesaran, M. H. and Tosetti, E. (2011), 'Large panels with common factors and spatial correlation', *Journal of Econometrics* **161**(2), 182–202.

Robinson, P. (2011), 'Asymptotic theory for nonparametric regression with spatial data', *Journal of Econometrics* **165**(1), 5–19.

Rodriguez-Poo, J. M. and Soberon, A. (2014), 'Direct semi–parametric estimation of fixed effects panel data varying coefficient models', *Econometrics Journal* **17**(1), 107–138.

Sealey, C. W. and Lindley, J. T. (1977), 'Inputs, outpus, and a theory of production and cost at depository financial institutions', *Journal of Finance* **32**(4), 1251–1266.

Su, L., Shi, Z. and Phillips, P. C. B. (2014), 'Identifying latent structures in panel data', *Working paper* .

Su, L. and Ullah, A. (2011), *Nonparametric and semiparametric panel econometric models: estimation and testing*, Handbook of Empirical Economics and Finance, New York: Taylor and Francis Group.

Sun, Y., Carroll, R. J. and Li, D. (2009), 'Semiparametric estimation of fixed effects panel data varying coefficient models', *Advances in Econometrics* **25**, 101–130.

Wang, H. and Xia, Y. (2009), 'Shrinkage estimation of the varying coefficient', *Journal of the American Statistical Association* **104**(486), 747–757.

Wheelock, D. C. and Wilson, P. W. (2012), 'Do large banks have lower costs? new estimates of returns to scale for u.s. banks', *Journal of Money, Credit and Banking* **44**(1), 171–199.

Yuan, M. and Lin, Y. (2006), 'Model selection and estimation in regression with grouped variables', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(1), 49–67.

Supplementary Document for the submission of

"A Varying-Coefficient Panel Data Model with Fixed Effects: Theory and an Application to U.S. Commercial Banks"

Guohua Feng[§], Jiti Gao[⋆], Bin Peng[†] and Xiaohui Zhang[‡]

[§]University of North Texas, [⋆]Monash University, [†]University of Technology Sydney and [‡]Murdoch University

Appendix B

In this file, we provide the algorithm for the variable selection procedure and the proofs of the asymptotic results.

B.1 Algorithm

The procedure for obtaining regularized estimates is described as follows:

1. Minimize the cross-validation criterion function (2.8) in order to choose $\hat{\lambda}$.

2. Select $\tilde{\tau}$ defined in (2.15) from a sufficient large set, say $[1, \sqrt[4]{NT}]$, by using a grid search. For each choice of $\tilde{\tau}$, estimate (2.13) using a similar procedure as proposed in Hunter and Li (2005) and Wang and Xia (2009). Define

$$\hat{B}_{\tilde{\tau}}^{(n)} = (\hat{\beta}_{\tilde{\tau},1}^{(n)}, \ldots, \hat{\beta}_{\tilde{\tau},m}^{(n)})' = (\hat{b}_{\tilde{\tau},1}^{(n)}, \ldots, \hat{b}_{\tilde{\tau},q}^{(n)}) \tag{B.1}$$

to be the estimate obtained in the n^{th} iteration. Then the loss function given above can be locally approximated by

$$\begin{aligned} &\sum_{j=1}^{m}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{Y}_{it} - \tilde{X}_{it}'\beta_j\right)^2 L(Z_i, z^j, \hat{\lambda}) + \sum_{s=1}^{q}\tilde{\tau}_s\frac{\|b_s\|^2}{\|\hat{b}_{\tilde{\tau},s}^{(n)}\|} \\ &= \sum_{j=1}^{m}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{Y}_{it} - \tilde{X}_{it}'\beta_j\right)^2 L(Z_i, z^j, \hat{\lambda}) + \sum_{s=1}^{q}\tilde{\tau}_s\frac{\beta_{j,s}^2}{\|\hat{b}_{\tilde{\tau},s}^{(n)}\|}\right). \end{aligned} \tag{B.2}$$

The minimizer of (B.2) is given by $\hat{B}_{\tilde{\tau}}^{(n+1)} = (\hat{\beta}_{\tilde{\tau},1}^{(n+1)}, \ldots, \hat{\beta}_{\tilde{\tau},m}^{(n+1)})'$, where for $j = 1, \ldots, m$

$$\hat{\beta}_{\tilde{\tau},j}^{(n+1)} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{it}\tilde{X}_{it}'L(Z_i, z^j, \hat{\lambda}) + D^{(n)}\right)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{it}\tilde{Y}_{it}L(Z_i, z^j, \hat{\lambda}), \tag{B.3}$$

and $D^{(n)} = \text{diag}\left(\|\hat{b}_{\tilde{\tau},1}^{(n)}\|^{-1}\tilde{\tau}_1, \ldots, \|\hat{b}_{\tau,q}^{(n)}\|^{-1}\tilde{\tau}_q\right)$. Repeat this procedure until $\|\hat{B}_{\tilde{\tau}}^{(n+1)} - \hat{B}_{\tilde{\tau}}^{(n)}\| <$ *tolerance*, where *tolerance* is a sufficiently small number (say, 10^{-8}).

3. Select the optimal estimator based on the modified BIC-type criterion.

Computer codes for implementing this procedure are available upon request and will be available in authors' website soon for general use.

B.2 Proofs

For notational simplicity, let $\hat{\beta}_{it} = \hat{\beta}_{-it}(Z_{it})$, $\beta_{it} = \beta(Z_{it})$, $1_{js,it} = 1(Z_{js} = Z_{it})$ and $\mathcal{Z} = \{Z_{it} : 1 \leq i \leq N, 1 \leq t \leq T\}$. Recall that we have defined $CV_0(\lambda)$, $\eta(z)$, $\Sigma_{XX}(z)$ and $\Sigma_{XX\beta}(z)$ in Assumption A.5. In the following proof, we will use these notations without defining them again. Also, in this note $O(1)$'s are some constants which may be different at each appearance.

Lemma B.1. *For two square matrices A and B with the same dimensions, suppose that A is non-singular and $\left\|A^{-1}B\right\| < 1$. Then we have the following expansion:*

$$(A+B)^{-1} = A^{-1} - A^{-1}BA^{-1} + A^{-1}BA^{-1}BA^{-1} - A^{-1}BA^{-1}BA^{-1}BA^{-1} + \cdots$$

The proof of Lemma B.1 is straightforward and thus omitted.

Lemma B.2. *Under Assumption A, as $(N,T) \to (\infty,\infty)$ jointly*

1. $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}^2 \to_P \sigma_u^2$;
2. $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{it}\tilde{X}_{it}'L(Z_{it},z,\lambda) - \Sigma_{XX}(z) \to_P 0$;
3. $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{it}\tilde{X}_{it}'\beta(Z_{it})L(Z_{it},z,\lambda) - \Sigma_{XX\beta}(z) \to_P 0$;
4. $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{it}\tilde{X}_{it}' - E[\Omega(Z_{it},\lambda)|\lambda] \to_P 0$;
5. $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{it}\tilde{X}_{it}'1(Z_{it}=z) - p(z)\Omega(z,\lambda) \to_P 0$;
6. $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{it}\tilde{X}_{it}'\beta(Z_{it}) - E[\Omega(Z_{it},\lambda)\beta(Z_{it})] \to_P 0$;
7. $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{it}\tilde{u}_{it}1(Z_{it}=z) = O_P\left(\frac{1}{\sqrt{NT}}\right)$;
8. $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{it}\tilde{u}_{it}L(Z_{it},z,\lambda) = O_P\left(\frac{1}{\sqrt{NT}}\right)$;
9. $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{it}\tilde{u}_{it} = O_P\left(\frac{1}{\sqrt{NT}}\right)$.

Proof of Lemma B.2:

1). We begin by expanding $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}^2$ as follows:

$$\begin{aligned}
&\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}^2 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(u_{it} - \frac{1}{T_{it}}\sum_{s=1}^{T}u_{is}L_{is,it}^p\right)\left(u_{it} - \frac{1}{T_{it}}\sum_{s=1}^{T}u_{is}L_{is,it}^p\right) \\
&= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}u_{it}^2 + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{T_{it}}\sum_{s_1=1}^{T}u_{is_1}L_{is_1,it}^p\frac{1}{T_{it}}\sum_{s_2=1}^{T}u_{is_2}L_{is_2,it}^p \\
&\quad - \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{T_{it}}\sum_{s=1}^{T}u_{it}u_{is}L_{is,it}^p
\end{aligned} \tag{B.4}$$

For the first term on RHS of (B.4), write

$$
\begin{aligned}
&E\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}u_{it}^2-\sigma_u^2\right|^2\\
&\leq \frac{1}{N^2T^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}c_{\delta_2}\left(\alpha_{u,ij}(|t-s|)\right)^{\delta_2/(4+\delta_2)}\left(E[u_{it}^{4+\delta_2}|\mathcal{X}]\cdot E[u_{js}^{4+\delta_2}|\mathcal{X}]\right)^{2/(4+\delta_2)}\\
&\leq O(1)\frac{1}{N^2T^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\left(\alpha_{u,ij}(|t-s|)\right)^{\delta_2/(4+\delta_2)}=O\left(\frac{1}{NT}\right),
\end{aligned}
$$

where $c_{\delta_2}=2^{(4+2\delta_2)/(4+\delta_2)}\cdot(4+\delta_2)/\delta_2$; the first inequality is by the Davydov inequality (c.f. pages 19-20 in Bosq (1996) and the supplement of Su and Jin (2012)); and the last line follows from Assumption A.4.

For the third term on right hand side of (B.4),

$$
\begin{aligned}
&\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{T_{it}}\sum_{s=1}^{T}u_{it}u_{is}L_{is,it}^2\leq\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}|u_{it}|\left|\frac{T}{T_{it}}\right|\left|\frac{1}{T}\sum_{s=1}^{T}u_{is}L_{is,it}^p\right|\\
&\leq O_P\left(\frac{1}{\sqrt{T}}\right)\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}|u_{it}|=O_P\left(\frac{1}{\sqrt{T}}\right),
\end{aligned}
$$

where the second inequality follows from Assumption A.4.

Similarly, $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{T_{it}}\sum_{s_1=1}^{T}u_{is_1}L_{is_1,it}^2\frac{1}{T_{it}}\sum_{s_2=1}^{T}u_{is_2}L_{is_2,it}^2=O_P\left(\frac{1}{T}\right)$, which completes the proof of the first result of this lemma. ■

2). We start by rewriting (2) of Lemma B.2 as follows:

$$
\begin{aligned}
&\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{it}\tilde{X}_{it}'L(Z_{it},z,\lambda)\\
&=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it}-\frac{1}{T_{it}}\sum_{s=1}^{T}X_{is}L_{is,it}^p\right)\left(X_{it}-\frac{1}{T_{it}}\sum_{s=1}^{T}X_{is}L_{is,it}^p\right)'L(Z_{it},z,\lambda)\\
&=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}X_{it}X_{it}'L(Z_{it},z,\lambda)+\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{T_{it}^2}\sum_{s_1=1}^{T}\sum_{s_2=1}^{T}X_{is_1}L_{is_1,it}^pX_{is_2}'L_{is_2,it}^pL(Z_{it},z,\lambda)\\
&\quad-\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{T_{it}}\sum_{s=1}^{T}X_{is}L_{is,it}^pX_{it}'L(Z_{it},z,\lambda)\\
&\quad-\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}X_{it}\frac{1}{T_{it}}\sum_{s=1}^{T}X_{is}'L_{is,it}^pL(Z_{it},z,\lambda).
\end{aligned}
\tag{B.5}
$$

We now consider each term on RHS of (B.5) respectively. We start with the first term on RHS of (B.5) as follows:

$$
\begin{aligned}
&E\left[\left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}X_{it}X_{it}'L(Z_{it},z,\lambda)-E[\Sigma_X(Z_{it})L(Z_{it},z,\lambda)]\right\|^2\right]\\
&=\frac{1}{N^2T^2}\sum_{m=1}^{q}\sum_{n=1}^{q}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}E\Big[\left(X_{it,m}X_{it,n}L(Z_{it},z,\lambda)-E[\Sigma_{X,mn}(Z_{lk})L(Z_{lk},z,\lambda)]\right)
\end{aligned}
$$

$$\cdot \left(X_{is,m}X_{is,n}L(Z_{is},z,\lambda) - E[\Sigma_{X,mn}(Z_{lk})L(Z_{lk},z,\lambda)]\right)\Big]$$
$$\leq \frac{1}{N^2T^2}\sum_{m=1}^{q}\sum_{n=1}^{q}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\left\{E\left[\left|X_{it,m}X_{it,n}L(Z_{it},z,\lambda)\right|^2\right]E\left[\left|X_{is,m}X_{is,n}L(Z_{it},z,\lambda)\right|^2\right]\right\}^{1/2}$$
$$\leq O(1)\frac{1}{N^2T^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\left\{E\left[\|X_{it}\|^4\right]E\left[\|X_{is}\|^4\right]\right\}^{1/2} = O\left(\frac{1}{N}\right), \tag{B.6}$$

where $X_{it,m}$ denotes the m^{th} element of X_{it} for $m = 1,\ldots,q$; $\Sigma_{X,mn}(z)$ denotes the $(m,n)^{th}$ element of $\Sigma_X(z)$ for $m,n = 1,\ldots,q$; the first inequality follows from Cauchy-Schwarz inequality; and the second inequality follows from $L(Z_{it},z,\lambda)$ being bounded uniformly. It thus implies that

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}X_{it}X_{it}'L(Z_{it},z,\lambda) - E[\Sigma_X(Z_{it})L(Z_{it},z,\lambda)] \to_P 0.$$

For the second term on RHS of (B.5), by Assumption A.3, we can write

$$\begin{aligned}
&\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{T_{it}^2}\sum_{s_1=1}^{T}\sum_{s_2=1}^{T}X_{is_1}L^p_{is_1,it}X_{is_2}'L^p_{is_2,it}L(Z_{it},z,\lambda)\\
=\ &\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Delta_3(Z_{it},\lambda)\Delta_3(Z_{it},\lambda)'L(Z_{it},z,\lambda) + o_P(1)\\
\to_P\ &E[\Delta_3(Z_{it},\lambda)\Delta_3(Z_{it},\lambda)'L(Z_{it},z,\lambda)],
\end{aligned} \tag{B.7}$$

where the last line follows from the same procedure as used in (B.6).

Similarly, for the last two terms on RHS of (B.5),

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{T_{it}}\sum_{s=1}^{T}X_{is}L^p_{is,it}X_{it}'L(Z_{it},z,\lambda) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Delta_3(Z_{it},\lambda)X_{it}'L(Z_{it},z,\lambda) + o_P(1)$$
$$\to_P E[\Delta_3(Z_{it},\lambda)X_{it}'L(Z_{it},z,\lambda)] = E[\Delta_3(Z_{it},\lambda)\mu_X(Z_{it})'L(Z_{it},z,\lambda)].$$

With the above discussions, the result follows. ■

3)-6). These four results follow by applying a similar procedure as used for proving the second result of this lemma. ■

7). We begin by expanding the left hand term of (7) of this lemma:

$$\begin{aligned}
&\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{it}\tilde{u}_{it}1(Z_{it}=z)\\
=\ &\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it} - \frac{1}{T_{it}}\sum_{s=1}^{T}X_{is}L^p_{is,it}\right)\left(u_{it} - \frac{1}{T_{it}}\sum_{s=1}^{T}u_{is}L^p_{is,it}\right)1(Z_{it}=z)\\
=\ &\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it} - \Delta_3(Z_{it},\lambda)\right)\left(u_{it} - \frac{1}{T_{it}}\sum_{s=1}^{T}u_{is}L^p_{is,it}\right)1(Z_{it}=z)\\
&-\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\frac{1}{T_{it}}\sum_{s=1}^{T}X_{is}L^p_{is,it} - \Delta_3(Z_{it},\lambda)\right)\left(u_{it} - \frac{1}{T_{it}}\sum_{s=1}^{T}u_{is}L^p_{is,it}\right)1(Z_{it}=z).
\end{aligned} \tag{B.8}$$

Firstly, we consider the second term on RHS of (B.8).

$$
\begin{aligned}
&E\left[\left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\frac{1}{T_{it}}\sum_{s=1}^{T}X_{is}L_{is,it}^{p}-\Delta_3(Z_{it},\lambda)\right)u_{it}1(Z_{it}=z)\right\|^2\right]\\
&\leq o(1)\frac{1}{N^2T^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t_1=1}^{T}\sum_{t_2=1}^{T}\left|E[u_{it_1}u_{jt_2}|\mathcal{X}]\right|\\
&\leq o(1)\frac{1}{N^2T^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t_1=1}^{T}\sum_{t_2=1}^{T}c_{\delta_2}\left(\alpha_{u,ij}(|t_1-t_2|)\right)^{\delta_2/(4+\delta_2)}\\
&\quad\cdot\left(E[|u_{it_1}|^{2+\delta_2/2}|\mathcal{X}]\right)^{2/(4+\delta_2)}\left(E[|u_{jt_2}|^{2+\delta_2/2}|\mathcal{X}]\right)^{2/(4+\delta_2)}\\
&\leq o(1)\frac{1}{N^2T^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t_1=1}^{T}\sum_{t_2=1}^{T}\left(\alpha_{u,ij}(|t_1-t_2|)\right)^{\delta_2/(4+\delta_2)}=o_P\left(\frac{1}{NT}\right),
\end{aligned}
\tag{B.9}
$$

where $c_{\delta_2}=2^{(4+2\delta_2)/(4+\delta_2)}\cdot(4+\delta_2)/\delta_2$; the first equality follows from Assumptions A.3-A.4; the second inequality follows from Davydov inequality; and the last line follows from Assumption A.4.

Similarly, we can obtain

$$
\begin{aligned}
&E\left[\left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\frac{1}{T_{it}}\sum_{s=1}^{T}X_{is}L_{is,it}^{p}-\Delta_3(Z_{it},\lambda)\right)\frac{1}{T_{it}}\sum_{s=1}^{T}u_{is}L_{is,it}^{2}1(Z_{it}=z)\right\|^2\right]=o_P\left(\frac{1}{NT}\right),\\
&E\left[\left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it}-\Delta_3(Z_{it},\lambda)\right)\frac{1}{T_{it}}\sum_{s=1}^{T}u_{is}L_{is,it}^{p}1(Z_{it}=z)\right\|^2\right]=o_P\left(\frac{1}{NT}\right).
\end{aligned}
$$

We therefore can further write

$$
\begin{aligned}
\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{it}\tilde{u}_{it}1(Z_{it}=z)&=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it}-\Delta_3(Z_{it,\lambda})\,u_{it}1(Z_{it}=z)\right.\\
&=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it}-\mu_X(z)\right)u_{it}1(Z_{it}=z)+o_P\left(\frac{1}{\sqrt{NT}}\right).
\end{aligned}
\tag{B.10}
$$

For the term on RHS of (B.10), write

$$
\begin{aligned}
&E\left[\left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it}-\mu_X(z)\right)u_{it}1(Z_{it}=z)\right\|^2\right]\\
&\leq\frac{1}{N^2T^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\left|E[\left(X_{it}-\mu_X(z)\right)'\left(X_{js}-\mu_X(z)\right)u_{it}u_{js}1(Z_{it}=z)1(Z_{js}=z)]\right|\\
&\leq O(1)\frac{1}{N^2T^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\left|E[u_{it}u_{js}|\mathcal{X}]\right|=O\left(\frac{1}{NT}\right),
\end{aligned}
$$

where the last equality follows from the proof of (B.9).

With the above discussions, the result follows. ■

8)-9) These two results follow by applying a similar procedure used for proving the seventh result of this lemma. ■

Note that the finite sample property of the leave-one-out estimator is different from the estimator in (2.7) provided in the main file which uses the whole sample, but they are interchangeable in the following analysis due to the assumption that both N and T are sufficiently large. Therefore, we express $\hat{\beta}_{it}$ as the estimator which uses the whole sample in what follows. A similar technique is also used in Li et al. (2013, p. 569).

$$\begin{aligned}\hat{\beta}_{it} - \beta_{it} = &\left(\sum_{j=1}^{N}\sum_{s=1}^{T} \tilde{X}_{js}\tilde{X}_{js}' L(Z_{js}, Z_{it}, \lambda)\right)^{-1} \sum_{j=1}^{N}\sum_{s=1}^{T} \tilde{X}_{js}\tilde{X}_{js}'(\beta_{js} - \beta_{it}) L(Z_{js}, Z_{it}, \lambda) \\ &+ \left(\sum_{j=1}^{N}\sum_{s=1}^{T} \tilde{X}_{js}\tilde{X}_{js}' L(Z_{js}, Z_{it}, \lambda)\right)^{-1} \sum_{j=1}^{N}\sum_{s=1}^{T} \tilde{X}_{js}\tilde{u}_{js} L(Z_{js}, Z_{it}, \lambda), \qquad \text{(B.11)}\end{aligned}$$

where we define $\beta_{it} = \beta(Z_{it})$ for notational simplicity.

Proof of Lemma 2.1.1:

We use Theorem 2.1 of Newey and McFadden (1994) to verify that $\hat{\lambda} = o_P(1)$. By Assumption A.5, $CV_0(\lambda)$ is uniquely minimized at $\lambda = (\lambda_1, \ldots, \lambda_r)' = 0$. Here, λ belongs to a compact set $[0,1]^r$, and $CV_0(\lambda)$ is continuous on $[0,1]^r$. Then we need only to show that $CV(\lambda)$ converges uniformly in probability to $CV_0(\lambda) + c$ below, where c is a positive constant uniformly in λ. For this purpose, write

$$\begin{aligned}CV(\lambda) &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{X}_{it}'(\beta_{it} - \hat{\beta}_{it}) + \gamma_{it}\right)^2 + \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{X}_{it}'(\beta_{it} - \hat{\beta}_{it}) + \gamma_{it}\right)\tilde{u}_{it} + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}^2 \\ &\equiv CV_1(\lambda) + CV_2(\lambda) + CV_3, \qquad \text{(B.12)}\end{aligned}$$

where $\gamma_{it} = \frac{1}{T_{it}}\sum_{s=1}^{T} X_{is}' \left(\beta(Z_{is}) - \beta(Z_{it})\right) L_{is,it}^{p}$.

The result (1) of Lemma B.2 implies $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}^2 \to_P \sigma_u^2$ uniformly in λ. Thus, we just need to focus on $CV_1(\lambda)$ and $CV_2(\lambda)$ below. Before proceeding further, we first investigate $\hat{\beta}_{it} - \beta_{it}$ and γ_{it}. By results (2), (3) and (8) of Lemma B.2, we can further write

$$\begin{aligned}\hat{\beta}_{it} - \beta_{it} &= \Sigma_{XX}^{-1}(Z_{it}, \lambda)\Sigma_{XX\beta}(Z_{it}, \lambda) - \beta(Z_{it}) + o_P(1) = \eta(Z_{it}, \lambda) - \beta(Z_{it}) + o_P(1), \\ \gamma_{it} &= \Delta_{3\beta}(Z_{it}, \lambda) - \Delta_3(Z_{it}, \lambda)'\beta(Z_{it}) + o_P(1). \qquad \text{(B.13)}\end{aligned}$$

By (B.13), $CV_1(\lambda)$ can be rewritten as

$$CV_1(\lambda) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left((X_{it} - \Delta_3(Z_{it}, \lambda))'(\beta(Z_{it}) - \eta(Z_{it}, \lambda)) + \Delta_{3\beta}(Z_{it}, \lambda) - \Delta_3(Z_{it}, \lambda)'\beta(Z_{it})\right)^2 + o_P(1).$$

Then by Assumptions A.2-A.3, it is easy to show that $CV_1(\lambda) \to_P CV_0(\lambda)$. Similarly, we can show that $CV_2(\lambda) = o_P(1)$ uniformly in λ.

Therefore, we have shown that $CV(\lambda) \to_P CV_0(\lambda) + \sigma_u^2$. Thus, all the conditions needed for Theorem 2.1 of Newey and McFadden (1994) are satisfied. Then the result follows. ■

Proof of Theorem 2.1.1:

In Lemma 2.1.1, we have shown $\hat{\lambda} = o_P(1)$, so it is reasonable to assume that λ, in proving this theorem, is sufficiently small and close to $0_{r\times 1}$. We now investigate the cross-validation criterion function and write

$$\begin{aligned} CV(\lambda) &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{X}_{it}'(\beta_{it}-\hat{\beta}_{it})\right)^2 + \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{it}'(\beta_{it}-\hat{\beta}_{it})\tilde{u}_{it} + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}^2 \\ &\quad + \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{X}_{it}'(\beta_{it}-\hat{\beta}_{it}) + \tilde{u}_{it}\right)\gamma_{it} + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\gamma_{it}^2 \\ &\equiv CV_1(\lambda) + CV_2(\lambda) + CV_3 + CV_4(\lambda) + CV_5(\lambda), \end{aligned} \tag{B.14}$$

where $\gamma_{it} = \frac{1}{T_{it}}\sum_{s=1}^{T} X_{is}'\left(\beta(Z_{it}) - \beta(Z_{is})\right) L_{is,it}^{p}$.

In (2.6), we have shown that $\gamma_{it} = O\left(\|\lambda\|^p\right)$ uniformly when λ is sufficiently small. In connection with the construction of $CV_4(\lambda)$ and $CV_5(\lambda)$, and Lemma B.2, we are able to obtain that $CV_4(\lambda) = O_P\left(\|\lambda\|^p\right)$ and $CV_5(\lambda) = O\left(\|\lambda\|^{2p}\right)$. By (1) of Lemma B.2, $CV_3 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}^2 \to_P \sigma_u^2$ and is independent of λ, so we focus on $CV_1(\lambda)$ and $CV_2(\lambda)$ below. To facilitate our analysis, we need to further consider $\hat{\beta}_{it} - \beta_{it}$. By Lemma 2.1.1, we can express the the kernel function as

$$L(Z_{js}, Z_{it}, \lambda) = 1_{js,it} + \sum_{m=1}^{r}\lambda_m 1_{m,jsit} + O(\|\lambda\|^2), \tag{B.15}$$

where $1_{m,jsit} = 1(Z_{js,m} \neq Z_{it,m})\prod_{n=1,n\neq m}^{r} 1(Z_{js,n} = Z_{it,n})$.

In what follows, we substitute (B.15) into each term on RHS of (B.11). Firstly,

$$\begin{aligned} &\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\tilde{X}_{js}\tilde{X}_{js}'L(Z_{js}, Z_{it}, \lambda) \\ &= \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\tilde{X}_{js}\tilde{X}_{js}'1_{js,it} + \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\tilde{X}_{js}\tilde{X}_{js}'\sum_{m=1}^{r}\lambda_m 1_{m,jsit} + O_P(\|\lambda\|^2) \\ &\equiv A_{1it} + A_{2it\lambda} + O_P(\|\lambda\|^2), \end{aligned} \tag{B.16}$$

where the first equality is due to result (4) of Lemma B.2.

Secondly,

$$\begin{aligned} &\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\tilde{X}_{js}\tilde{X}_{js}'(\beta_{js}-\beta_{it})L(Z_{js}, Z_{it}, \lambda) \\ &= \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\tilde{X}_{js}\tilde{X}_{js}'(\beta_{js}-\beta_{it})1_{js,it} + \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\tilde{X}_{js}\tilde{X}_{js}'(\beta_{js}-\beta_{it})\sum_{m=1}^{r}\lambda_m 1_{m,jsit} + O_P(\|\lambda\|^2) \\ &\equiv 0 + B_{2it\lambda} + O_P(\|\lambda\|^2), \end{aligned} \tag{B.17}$$

where the first equality is due to (4) and (6) of Lemma B.2 and the uniform bound on $\beta(z)$; and the zero term of the last line is due to $(\beta_{js}-\beta_{it})1_{js,it} = 0$.

Thirdly,

$$\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\tilde{X}_{js}\tilde{u}_{js}L(Z_{js}, Z_{it}, \lambda)$$

$$= \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\tilde{X}_{js}\tilde{u}_{js}1_{js,it} + \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\tilde{X}_{js}\tilde{u}_{js}\sum_{m=1}^{r}\lambda_m 1_{m,jsit} + O_P\left(\frac{\|\lambda\|^2}{\sqrt{NT}}\right)$$
$$\equiv C_{1it} + C_{2it\lambda} + O_P\left(\frac{\|\lambda\|^2}{\sqrt{NT}}\right), \tag{B.18}$$

where the first equality is due to (9) of Lemma B.2.

For the terms on RHS of (B.16)-(B.18), by Lemma B.2, it is straightforward to obtain

$$A_{1it}^{-1} = O_P(1), \quad A_{2it\lambda} = O_P\left(\|\lambda\|\right), \quad B_{2it\lambda} = O_P(\|\lambda\|),$$
$$C_{1it} = O_P\left(\frac{1}{\sqrt{NT}}\right), \quad C_{2it\lambda} = O_P\left(\frac{\|\lambda\|}{\sqrt{NT}}\right). \tag{B.19}$$

By (B.16), using Lemma B.1 twice gives the following expression.

$$\left(\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\tilde{X}_{js}\tilde{X}_{js}'L(Z_{js},Z_{it},\lambda)\right)^{-1} = \left(A_{1it} + A_{2it\lambda} + O_P(\|\lambda\|^2)\right)^{-1}$$
$$= (A_{1it} + A_{2it\lambda})^{-1} + O_P(\|\lambda\|^2) = A_{1it}^{-1} - A_{1it}^{-1}A_{2it\lambda}A_{1it}^{-1} + O_P(\|\lambda\|^2) \tag{B.20}$$

We then use (B.19) and (B.20) to further simplify (B.11) as follows.

$$\hat{\beta}_{it} = \beta_{it} + \left(A_{1it}^{-1} - A_{1it}^{-1}A_{2it\lambda}A_{1it}^{-1}\right)\left(B_{2it\lambda} + C_{1it} + C_{2it\lambda}\right) + O_P\left(\frac{\|\lambda\|^2}{\sqrt{NT}}\right) + O_P\left(\|\lambda\|^3\right) \tag{B.21}$$

We are now ready to further analyze $CV_1(\lambda)$ and $CV_2(\lambda)$ by using (B.19) and (B.21).

$$CV_1(\lambda) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{X}_{it}'(\beta_{it} - \hat{\beta}_{it})\right)^2$$
$$= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\tilde{X}_{it}'\left(A_{1it}^{-1}A_{2it\lambda}A_{1it}^{-1} - A_{1it}^{-1}\right)\left(B_{2it\lambda} + C_{1it} + C_{2it\lambda}\right)\right\}^2 + O_P\left(\frac{\|\lambda\|^2}{\sqrt{NT}}\right) + O_P\left(\|\lambda\|^3\right)$$
$$= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(D_{3it}^2 - 2D_{1it}D_{2it} + 2D_{2it}D_{3it}\right)$$
$$+O_P\left(\frac{\|\lambda\|^2}{\sqrt{NT}}\right) + O_P\left(\|\lambda\|^3\right) + \text{ terms independent of } \lambda,$$

where $D_{1it} = \tilde{X}_{it}'A_{1it}^{-1}\left(A_{2it\lambda}A_{1it}^{-1}C_{1it} - C_{2it\lambda}\right)$, $D_{2it} = \tilde{X}_{it}'A_{1it}^{-1}C_{1it}$ and $D_{3it} = \tilde{X}_{it}'A_{1it}^{-1}B_{2it\lambda}$.

$$CV_2(\lambda) = \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}\tilde{X}_{it}'(\beta_{it} - \hat{\beta}_{it})$$
$$= \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}\tilde{X}_{it}'A_{1it}^{-1}A_{2it\lambda}A_{1it}^{-1}\left(B_{2it\lambda} + C_{1it} + C_{2it\lambda}\right)$$
$$-\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}\tilde{X}_{it}'A_{1it}^{-1}\left(B_{2it\lambda} + C_{1it} + C_{2it\lambda}\right) + O_P\left(\frac{\|\lambda\|^2}{NT}\right) + O_P\left(\frac{\|\lambda\|^3}{\sqrt{NT}}\right)$$
$$= \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}\tilde{X}_{it}'A_{1it}^{-1}A_{2it\lambda}A_{1it}^{-1}C_{1,it} - \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}\tilde{X}_{it}'A_{1it}^{-1}B_{2it\lambda} - \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}\tilde{X}_{it}'A_{1it}^{-1}C_{2it\lambda}$$

$$+O_P\left(\frac{\|\lambda\|^2}{\sqrt{NT}}\right)+ \text{ terms independent of } \lambda,$$

where the first equality follows from (9) of Lemma B.2 and (B.21); and the second equality follows from (9) of Lemma B.2 and (B.19).

Note that

$$\begin{aligned}
&\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{2it}D_{3it} = \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{X}_{it}'A_{1it}^{-1}C_{1it}\tilde{X}_{it}'A_{1it}^{-1}B_{2it\lambda} \\
&= \frac{2}{N^3T^3}\sum_{m=1}^{r}\lambda_m \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{N}\sum_{s=1}^{T}\sum_{k=1}^{N}\sum_{r=1}^{T} \tilde{X}_{it}'A_{1it}^{-1}\tilde{X}_{js}\tilde{u}_{js}1_{js,it}\tilde{X}_{it}'A_{1it}^{-1}\tilde{X}_{kr}\tilde{X}_{kr}'(\beta_{kr}-\beta_{it})1_{m,krit} \\
&= \frac{2}{N^3T^3}\sum_{m=1}^{r}\lambda_m \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{N}\sum_{s=1}^{T}\sum_{k=1}^{N}\sum_{r=1}^{T} \tilde{X}_{kr}'A_{1kr}^{-1}\tilde{X}_{it}\tilde{u}_{it}1_{it,kr}\tilde{X}_{kr}'A_{1kr}^{-1}\tilde{X}_{js}\tilde{X}_{js}'(\beta_{js}-\beta_{kr})1_{m,jskr} \\
&= \frac{2}{N^3T^3}\sum_{m=1}^{r}\lambda_m \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{N}\sum_{s=1}^{T}\sum_{k=1}^{N}\sum_{r=1}^{T} \tilde{X}_{it}'A_{1it}^{-1}\tilde{X}_{kr}\tilde{u}_{it}1_{it,kr}\tilde{X}_{kr}'A_{1it}^{-1}\tilde{X}_{js}\tilde{X}_{js}'(\beta_{js}-\beta_{it})1_{m,jsit} \\
&= \frac{2}{N^2T^2}\sum_{m=1}^{r}\lambda_m \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{N}\sum_{s=1}^{T} \tilde{X}_{it}'A_{1it}^{-1}\tilde{u}_{it}\tilde{X}_{js}\tilde{X}_{js}'(\beta_{js}-\beta_{it})1_{m,jsit} \\
&= \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}\tilde{X}_{it}'A_{1it}^{-1}B_{2it\lambda},
\end{aligned}$$

where the third equality follows from changing the index (it, js, kr) to (kr, it, js); the fourth equality follows from the definition of $1_{it,kr}$; and the fifth equality follows from the definition of A_{1it}. Note that the term on RHS of the above equation can be canceled out by the leading term of $CV_2(\lambda)$.

Thus, we are now able to further write

$$\begin{aligned}
CV(\lambda) = &\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(D_{3it}^2 - 2D_{1it}D_{2it}\right) + \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}\tilde{X}_{it}'A_{1it}^{-1}A_{2it\lambda}A_{1it}^{-1}C_{1,it} \\
&-\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}\tilde{X}_{it}'A_{1it}^{-1}C_{2it\lambda} + O_P\left(\frac{\|\lambda\|^2}{\sqrt{NT}}\right) + O_P\left(\|\lambda\|^3\right) \\
&+ \text{ terms independent of } \lambda. \qquad \text{(B.22)}
\end{aligned}$$

Moreover, by (B.19) and some tedious algebra, we can show

$$\begin{aligned}
&\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}D_{3it} = O_P\left(\|\lambda\|^2\right), \quad \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}D_{1it}D_{2it} = O_P\left(\frac{\|\lambda\|}{NT}\right), \\
&\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}\tilde{X}_{it}'A_{1it}^{-1}A_{2it\lambda}A_{1it}^{-1}C_{1,it} = O_P\left(\frac{\|\lambda\|}{NT}\right), \\
&\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}\tilde{X}_{it}'A_{1it}^{-1}C_{2it\lambda} = O_P\left(\frac{\|\lambda\|}{NT}\right).
\end{aligned}$$

Based on the above discussions, (B.22) can be further simplified as follows.

$$CV(\lambda) = O_P\left(\frac{\|\lambda\|}{NT}\right) + O_P(\|\lambda\|^2) + \text{ terms independent of } \lambda, \qquad \text{(B.23)}$$

which immediately implies that $\hat{\lambda} = O_P\left(\frac{1}{NT}\right)$. ■

Proof of Theorem 2.1.2:

Denote

$$\check{T}_{it} = \sum_{s=1}^{T} 1(Z_{is} = Z_{it}), \quad \check{Y}_{it} = Y_{it} - \frac{1}{\check{T}_{it}} \sum_{s=1}^{T} Y_{is} 1_{is,it},$$
$$\check{X}_{it} = X_{it} - \frac{1}{\check{T}_{it}} \sum_{s=1}^{T} X_{is} 1_{is,it}, \quad \check{u}_{it} = u_{it} - \frac{1}{\check{T}_{it}} \sum_{s=1}^{T} u_{is} 1_{is,it}.$$

Note that for the large N and small T case, $\frac{1}{\check{T}_{it}} \sum_{s=1}^{T} u_{is} 1_{is,it}$ should be replaced by

$$A_{u,it} = \lim_{\lambda \to 0_{r\times 1}} \sum_{s=1}^{T} u_{is} L_{is,it}^{p} / \sum_{s=1}^{T} L_{is,it}^{p}$$

as discussed in Section 2.4. $\frac{1}{\check{T}_{it}} \sum_{s=1}^{T} Y_{is} 1_{is,it}$ and $\frac{1}{\check{T}_{it}} \sum_{s=1}^{T} X_{is} 1_{is,it}$ should be changed in a similar fashion.

Expanding all kernel functions in (2.7) by using (B.15) easily leads to $\hat{\beta}(z) = \check{\beta}(z) + O_P\left(\frac{1}{NT}\right)$ by Lemma B.1 and Theorem 2.1.1, where $\check{\beta}(z)$

$$\check{\beta}(z) = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \check{X}_{it} \check{X}_{it}' 1(Z_{it} = z) \right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \check{X}_{it} \check{Y}_{it}' 1(Z_{it} = z).$$

Thus, it is straightforward to obtain $\sqrt{NT}(\hat{\beta}(z) - \beta(z)) = \sqrt{NT}(\check{\beta}(z) - \beta(z)) + O_P\left(\frac{1}{\sqrt{NT}}\right)$. Below we just need to focus on $\sqrt{NT}(\check{\beta}(z) - \beta(z))$, so write

$$\begin{aligned}
&\sqrt{NT}(\check{\beta}(z) - \beta(z)) \\
&= \sqrt{NT} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \check{X}_{it} \check{X}_{it}' 1(Z_{it} = z) \right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \check{X}_{it} \left(\check{X}_{it}' (\beta(Z_{it}) - \beta(z)) + \check{u}_{it} \right) 1(Z_{it} = z) \\
&= \sqrt{NT} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \check{X}_{it} \check{X}_{it}' 1(Z_{it} = z) \right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \check{X}_{it} \check{u}_{it} 1(Z_{it} = z),
\end{aligned}$$

where the second equality is due to $(\beta(Z_{it}) - \beta(z)) 1(Z_{it} = z) = 0$.

As with (5) of Lemma B.2, it is easy to show that

$$\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \check{X}_{it} \check{X}_{it}' 1(Z_{it} = z) \to_P p(z) \left(\Sigma_X(z) - \mu_X(z) \mu_X(z)' \right) = \Xi_1(z).$$

Therefore, we need only to focus on $\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} \check{X}_{it} \check{u}_{it} 1(Z_{it} = z)$. As with the proof for (7) of Lemma B.2, we can show that

$$\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} \check{X}_{it} \check{u}_{it} 1(Z_{it} = z) = \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(X_{it} - \mu_X(z) \right) u_{it} 1(Z_{it} = z) + o_P(1).$$

Thus, we focus on $\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} X_{it} u_{it} 1(Z_{it} = z)$ below. For notational simplicity, denote that

$$\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it}-\mu_X(z)\right)u_{it}1(Z_{it}=z)=\sum_{t=1}^{T}V_{T,N}(t),$$

where $V_{T,N}(t)=\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\left(X_{it}-\mu_X(z)\right)u_{it}1(Z_{it}=z)$. By the construction of $V_{T,N}(t)$ and Assumptions A.2-A.4, $V_{T,N}(t)$ is stationary and α-mixing. We can then apply the large-block and small-block technique to show the normality below (c.f. Theorem 2.21 in Fan and Yao (2003); Lemma A.1 in Gao (2007); Lemma A.1 in Chen et al. (2012)). For this purpose, we partition the set $\{1,\ldots,T\}$ into $2k_T+1$ subsets with a large block of size l_T, a small block of size s_T and the remaining set of size $T-k_T(l_T+s_T)$, where, for any $\lambda>2$, $l_T=\lfloor T^{(\lambda-1)/\lambda}\rfloor$, $s_T=\lfloor T^{1/\lambda}\rfloor$ and $k_T=\lfloor T/(l_T+s_T)\rfloor$. Denote that for $n=1,\ldots,k_T$

$$\tilde{V}_n=\sum_{t=(n-1)(l_T+s_T)+1}^{nl_T+(n-1)s_T}V_{T,N}(t),\quad \bar{V}_n=\sum_{t=nl_T+(n-1)s_T+1}^{n(l_T+s_T)}V_{T,N}(t)\text{ and }\hat{V}=\sum_{t=k_T(l_T+s_T)+1}^{T}V_{T,N}(t).$$

By the properties of α-mixing process and a procedure similar to A.6 and A.7 in Chen et al. (2012), we obtain that $E\left\|\sum_{n=1}^{k_T}\bar{V}_n\right\|^2=O\left(\frac{k_Ts_T}{T}\right)=o(1)$ and $E\left\|\hat{V}\right\|^2=O\left(\frac{T-k_Tl_T}{T}\right)=o(1)$. Thus, we just need to focus on $\sum_{n=1}^{k_T}\tilde{V}_n$ below. Using Proposition 2.6 in Fan and Yao (2003) and the condition on the α-mixing coefficient, we have

$$\left|E\left[\exp\left\{\sum_{n=1}^{k_T}\|\tilde{V}_n\|\right\}\right]-\prod_{n=1}^{k_T}E\left[\exp\left\{\|\tilde{V}_n\|\right\}\right]\right|\leq C(k_T-1)\alpha(s_T)\to 0,$$

where C is a constant; $\alpha(\cdot)$ denotes the upper bound of the α-mixing coefficients provided in Assumption A and is achievable in the same way as Assumption A.4 of Chen et al. (2012). Then we obtain that $\tilde{V}_n$ for $n=1,\ldots,k_T$ are asymptotically independent. Furthermore, as in the proof of Theorem 2.21.(ii) in Fan and Yao (2003), we have $\mathrm{Cov}\left[\tilde{V}_1\right]=\frac{l_T}{T}\Xi_0(z)(I_q+o(1))$, where

$$\Xi_0(z)=\lim_{N,T\to\infty}\frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}E\left[u_{it}u_{js}(X_{it}-\mu_X(z))(X_{js}-\mu_X(z))'1(Z_{it}=z)1(Z_{js}=z)\right].$$

It further implies that

$$\sum_{n=1}^{k_T}\mathrm{Cov}\left[\tilde{V}_n\right]=k_T\cdot\mathrm{Cov}\left[\tilde{V}_1\right]=\frac{k_Tl_T}{T}\Xi_0\left(I_q+o\left(1\right)\right)\to\Xi_0,$$

which indicates the Feller condition is satisfied.

Moreover, by Cauchy-Schwarz inequality, we have

$$E\left[\left\|\tilde{V}_n\right\|^2\cdot I\left\{\|V_n\|\geq\varepsilon\right\}\right]\leq\left\{E\left\|\tilde{V}_n\right\|^3\right\}^{2/3}\cdot\left\{P\left(\left\|\tilde{V}_n\right\|\geq\varepsilon\right)\right\}^{1/3}\leq C\left\{E\left\|\tilde{V}_n\right\|^3\right\}^{2/3}\cdot\left\{E\left\|\tilde{V}_n\right\|^2\right\}^{1/3}$$

and by Lemma B.2 in Chen et al. (2012)

$$E\left\|\tilde{V}_n\right\|^3\leq\left(\frac{l_T}{T}\right)^{3/2}\left\{E\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(X_{i1}-\mu_X(z)\right)u_{i1}1(Z_{i1}=z)\right\|^4\right\}^{3/4}<\infty.$$

Therefore, $E\left\|\tilde{V}_n\right\|^3 = O\left(\left(\frac{l_T}{T}\right)^{3/2}\right)$, which implies that

$$E\left[\left\|\tilde{V}_n\right\|^2 \cdot I\left\{\|V_n\| \geq \varepsilon\right\}\right] \leq O\left(\left(\frac{l_T}{T}\right)^{4/3}\right) = o\left(\frac{l_T}{T}\right).$$

Consequently, $\sum_{n=1}^{k_T} E\left[\left\|\tilde{V}_n\right\|^2 \cdot I\left\{\|V_n\| \geq \varepsilon\right\}\right] = o\left(\frac{k_T l_T}{T}\right) = o(1)$. Therefore, the Lindeberg condition is satisfied. Based on the above discussions, $\sqrt{NT}(\check{\beta}(z) - \beta(z)) \to_D N(0, \Xi_1(z)^{-1}\Xi_0(z)\Xi_1(z)^{-1})$, which completes the proof. ■

Proof of Corollary 2.1.1:

All we need to show is that $\hat{\sigma}_u^2 \to_P \sigma_u^2$. We start by writing

$$\hat{\sigma}_u^2 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{X}_{it}'(\beta(Z_{it}) - \hat{\beta}(Z_{it})) + \tilde{u}_{it} + \gamma_{it})^2 = A_1 + A_2 + 2A_3 + 2A_4 + A_5,$$

where

$$A_1 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{X}_{it}'(\beta(Z_{it}) - \hat{\beta}(Z_{it})))^2, \quad A_2 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}^2,$$
$$A_3 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{it}'(\beta(Z_{it}) - \hat{\beta}(Z_{it}))\tilde{u}_{it}, \quad A_4 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{it}'(\beta(Z_{it}) - \hat{\beta}(Z_{it}) + \tilde{u}_{it})\gamma_{it},$$
$$A_5 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\gamma_{it}^2.$$

For A_1, we have

$$|A_1| \leq \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|\tilde{X}_{it}\right\|^2\left\|\beta(Z_{it}) - \hat{\beta}(Z_{it})\right\|^2 \leq O_P\left(\frac{1}{NT}\right)\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|\tilde{X}_{it}\right\|^2 = O_P\left(\frac{1}{NT}\right),$$

where the second inequality follows from Theorem 2.1.2. Thus, $A_1 \to_P 0$. Similarly, we can show that $A_3 \to_P 0$. By (1) of Lemma B.2, $A_2 \to_P \sigma_u^2$. Moreover, we have shown $A_4 = O_P(\|\hat{\lambda}\|^p)$ and $A_5 = O_P(\|\hat{\lambda}\|^{2p})$ in proving Theorem 2.1.1. Therefore, the result follows. ■

Note that if we replace Assumption A.5 with Assumption B, we can still show that Lemma B.2 holds by making some slight modifications to the proof. Specifically, for (2)-(3) of Lemma B.2, $\Sigma_{XX}(z)$ and $\Sigma_{XX\beta}(z)$ become $\Sigma_{XX}(\bar{z}) \cdot E[L(\tilde{Z}_{it}, \tilde{z}, \tilde{\lambda})]$ and $\Sigma_{XX\beta}(\bar{z}) \cdot E[L(\tilde{Z}_{it}, \tilde{z}, \tilde{\lambda})]$ for $\forall z \in \mathcal{D}$, respectvely; for (4)-(6) of Lemma B.2, $\Omega(z, \lambda)$, $p(z)$ and $\beta(z)$ reduce to $\Omega(\bar{z}, \bar{\lambda})$, $p(\bar{z})$ and $\beta(\bar{z})$, respectively; (1) and (7)-(9) of Lemma B.2 hold without requiring any modification. Thus, when establishing asymptotic results for the irrelevant case in what follows, we will still use the basic results proved in Lemma B.2.

Proof of Lemma 2.2.1:

By Assumption B, $CV_0^*(\bar{\lambda})$ is uniquely minimized at $\bar{\lambda} = (\lambda_1, \ldots, \lambda_{r_1})' = 0$ and $\bar{\lambda}$ belongs to a compact set $[0,1]^{r_1}$. Also, $CV_0^*(\bar{\lambda})$ is continuous on $[0,1]^{r_1}$. Then we need only to show that $CV(\lambda)$ converges uniformly in probability to $CV_0^*(\bar{\lambda}) + c$ below, where c is a positive constant. Note that λ_s

for $s = r_1 + 1, \ldots, r$ associated with the irrelevant covariates get canceled in the asymptotic results, so they do not play a role when we minimize the cross-validation criterion function. Without loss of generality, λ_s for $s = r_1 + 1, \ldots, r$ can be considered as arbitrary constants. The following procedure holds uniformly in λ_s for $s = r_1 + 1, \ldots, r$.

Note also that for the irrelevant case the coefficient function reduces to $\beta(\bar{z})$. Thus, write

$$\begin{aligned} CV(\lambda) &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{X}_{it}'(\bar{\beta}_{it} - \hat{\beta}_{it}) + \gamma_{it}\right)^2 + \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{X}_{it}'(\bar{\beta}_{it} - \hat{\beta}_{it}) + \gamma_{it}\right)\tilde{u}_{it} + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}^2 \\ &\equiv CV_1(\lambda) + CV_2(\lambda) + CV_3, \end{aligned}$$

where $\bar{\beta}_{it} = \beta(\bar{Z}_{it})$ and $\gamma_{it} = \frac{1}{T_{it}}\sum_{s=1}^{T} X_{is}'\left(\beta(\bar{Z}_{is}) - \beta(\bar{Z}_{it})\right)L_{is,it}^p$.

By result (1) of Lemma B.2, $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}^2 \to_P \sigma_u^2$ uniformly in λ. Thus, we just need to focus on $CV_1(\lambda)$ and $CV_2(\lambda)$ below. Recall that $L(Z_{js}, z, \lambda) = L(\bar{Z}_{js}, \bar{z}, \bar{\lambda})L(\tilde{Z}_{js}, \tilde{z}, \tilde{\lambda})$. As discussed before, Lemma B.2 holds if Assumption A.5 is replaced with Assumption B. Thus, it is easy to know that $\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\tilde{X}_{js}\tilde{u}_{js}L(Z_{js}, z, \lambda) \to_P 0$. Moreover, for $\forall z \in \mathcal{D}$,

$$\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\tilde{X}_{js}\tilde{X}_{js}'L(Z_{js}, z, \lambda) \to_P \Sigma_{XX}(\bar{z}, \bar{\lambda}) \cdot E[L(\tilde{Z}_{js}, \tilde{z}, \tilde{\lambda})] \tag{B.24}$$

and

$$\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\tilde{X}_{js}\tilde{X}_{js}'\beta(\bar{Z}_{js})L(Z_{js}, z, \lambda) \to_P \Sigma_{XX\beta}(\bar{z}, \bar{\lambda}) \cdot E[L(\tilde{Z}_{js}, \tilde{z}, \tilde{\lambda})]. \tag{B.25}$$

Note that $E[L(\tilde{Z}_{js}, \tilde{z}, \tilde{\lambda})]$ gets canceled after we substitute (B.24) and (B.25) into (B.11). We thus write

$$\begin{aligned} \hat{\beta}_{it} - \bar{\beta}_{it} &= \Sigma_{XX}^{-1}(\bar{Z}_{it}, \bar{\lambda})\Sigma_{XX\beta}(\bar{Z}_{it}, \bar{\lambda}) - \beta(\bar{Z}_{it}) + o_P(1) = \eta(\bar{Z}_{it}, \bar{\lambda}) - \beta(\bar{Z}_{it}) + o_P(1), \\ \gamma_{it} &= \Delta_{3\beta}(\bar{Z}_{it}, \bar{\lambda}) - \Delta_3(\bar{Z}_{it}, \bar{\lambda})'\beta(\bar{Z}_{it}) + o_P(1). \end{aligned} \tag{B.26}$$

By (B.26), $CV_1(\lambda)$ can be rewritten as

$$CV_1(\lambda) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left((X_{it} - \Delta_3(\bar{Z}_{it}, \bar{\lambda}))'(\beta(\bar{Z}_{it}) - \eta(\bar{Z}_{it}, \bar{\lambda})) + \Delta_{3\beta}(\bar{Z}_{it}, \bar{\lambda}) - \Delta_3(\bar{Z}_{it}, \bar{\lambda})'\beta(\bar{Z}_{it})\right)^2 + o_P(1).$$

Then by Assumptions A.2-A.3, it is easy to know that $CV_1(\lambda) \to_P CV_0^*(\bar{\lambda})$.

Similarly, we can show that $CV_2(\lambda) = o_P(1)$. With the above discussions, it is easy to see $CV(\lambda) \to_P CV_0^*(\bar{\lambda}) + \sigma_u^2$ uniformly in $\tilde{\lambda} \in \tilde{\mathcal{D}}$. Thus, all the conditions needed for Theorem 2.1 of Newey and McFadden (1994) are satisfied. Then the result follows. ■

Proof of Theorem 2.2.1:

1). Note that we have shown that $\hat{\lambda}_s = o_P(1)$ for $s = 1, \ldots, r_1$ in Lemma 2.2.1. Therefore, it is reasonable to assume that $\bar{\lambda}$ used in proving this theorem is sufficiently small and close to $0_{r_1 \times 1}$. For simplicity, define $\bar{1}_{itjs} = 1(\bar{Z}_{it} = \bar{Z}_{js})$ and $\bar{1}_{n,itjs} = 1(Z_{it,n} \neq Z_{js,n})\prod_{m=1,m\neq n}^{r_1} 1(Z_{it,m} = Z_{js,m})$ for

$n = 1, \ldots, r_1$. Let $\bar{L}_{jsit,\bar{\lambda}} = L(\bar{Z}_{js}, \bar{Z}_{it}, \bar{\lambda})$ and $\tilde{L}_{jsit,\tilde{\lambda}} = L(\tilde{Z}_{js}, \tilde{Z}_{it}, \tilde{\lambda})$. Using the kernel function of Aitchison and Aitken (1976) and the expansion technique used in (B.15), we can write

$$L(Z_{js}, Z_{it}, \lambda) = \bar{L}_{jsit,\bar{\lambda}} \tilde{L}_{jsit,\tilde{\lambda}} = \left(\bar{1}_{jsit} + \sum_{n=1}^{r_1} \lambda_n \bar{1}_{n,jsit} + O(\|\bar{\lambda}\|^2) \right) \tilde{L}_{jsit,\tilde{\lambda}}. \tag{B.27}$$

Before investigating the cross-validation criterion function, we further simplify $\hat{\beta}_{it} - \bar{\beta}_{it}$. Write

$$\begin{aligned}
\hat{\beta}_{it} - \bar{\beta}_{it} &= \left(\frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} \tilde{X}_{js} \tilde{X}'_{js} \bar{L}_{jsit,\bar{\lambda}} \tilde{L}_{jsit,\tilde{\lambda}} \right)^{-1} \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} \tilde{X}_{js} \tilde{X}'_{js} (\bar{\beta}_{js} - \bar{\beta}_{it}) \bar{L}_{jsit,\bar{\lambda}} \tilde{L}_{jsit,\tilde{\lambda}} \\
&\quad + \left(\frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} \tilde{X}_{js} \tilde{X}'_{js} \bar{L}_{jsit,\bar{\lambda}} \tilde{L}_{jsit,\tilde{\lambda}} \right)^{-1} \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} \tilde{X}_{js} \tilde{u}_{js} \bar{L}_{jsit,\bar{\lambda}} \tilde{L}_{jsit,\tilde{\lambda}} \\
&= \left(A_{1it} + A_{2it\lambda} + O_P(\|\bar{\lambda}\|^2) \right)^{-1} (B_{it} + C_{it}),
\end{aligned} \tag{B.28}$$

where the term $O_P(\|\bar{\lambda}\|^2)$ in the last line follows from (B.27) and (4) of Lemma B.2; and

$$\begin{aligned}
A_{1it} &= \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} \tilde{X}_{js} \tilde{X}'_{js} \bar{1}_{jsit} \tilde{L}_{jsit,\tilde{\lambda}} \\
A_{2it\lambda} &= \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} \tilde{X}_{js} \tilde{X}'_{js} \sum_{n=1}^{r_1} \lambda_n \bar{1}_{n,jsit} \tilde{L}_{jsit,\tilde{\lambda}} \\
B_{it} &= \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} \tilde{X}_{js} \tilde{X}'_{js} (\bar{\beta}_{js} - \bar{\beta}_{it}) \bar{L}_{jsit,\bar{\lambda}} \tilde{L}_{jsit,\tilde{\lambda}} \\
C_{it} &= \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} \tilde{X}_{js} \tilde{u}_{js} \bar{L}_{jsit,\bar{\lambda}} \tilde{L}_{jsit,\tilde{\lambda}}.
\end{aligned}$$

Applying a similar procedure as used for proving (2) of Lemma B.2 to A_{1it} and $A_{2it\lambda}$, we obtain $A_{1it} = O_P(1)$ and $A_{2it\lambda} = O_P(\|\bar{\lambda}\|)$. Applying the same procedure to B_{it}, we have

$$\begin{aligned}
B_{it} &= \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} \tilde{X}_{js} \tilde{X}'_{js} (\bar{\beta}_{js} - \bar{\beta}_{it}) \left(\bar{1}_{jsit} + \sum_{n=1}^{r_1} \lambda_n \bar{1}_{n,jsit} + O(\|\bar{\lambda}\|^2) \right) \tilde{L}_{jsit,\tilde{\lambda}} \\
&= 0 + B_{2it\lambda} + O_P(\|\bar{\lambda}\|^2),
\end{aligned}$$

where the zero term follows from $(\bar{\beta}_{js} - \bar{\beta}_{it}) \bar{1}_{jsit} = 0$ and

$$\begin{aligned}
B_{2it\lambda} &= \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} \tilde{X}_{js} \tilde{X}'_{js} (\bar{\beta}_{js} - \bar{\beta}_{it}) \sum_{n=1}^{r_1} \lambda_n \bar{1}_{n,jsit} \tilde{L}_{jsit,\tilde{\lambda}} \\
&= \sum_{n=1}^{r_1} \lambda_n \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} \tilde{X}_{js} \tilde{X}'_{js} (\bar{\beta}_{js} - \bar{\beta}_{it}) \bar{1}_{n,jsit} \tilde{L}_{jsit,\tilde{\lambda}} = O_P\left(\|\bar{\lambda}\| \right).
\end{aligned}$$

Using a similar procedure as used for proving (7) of Lemma B.2 to C_{it}, we obtain

$$C_{it} = \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} \tilde{X}_{js} \tilde{u}_{js} \left(\bar{1}_{jsit} + \sum_{n=1}^{r_1} \lambda_n \bar{1}_{n,jsit} + O(\|\bar{\lambda}\|^2) \right) \tilde{L}_{jsit,\tilde{\lambda}}$$

$$= C_{1it} + C_{2it\lambda} + O_P\left(\frac{\|\bar{\lambda}\|^2}{\sqrt{NT}}\right),$$

where

$$C_{1it} = \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\tilde{X}_{js}\tilde{u}_{js}\bar{1}_{jsit}\tilde{L}_{jsit,\tilde{\lambda}} = O_P\left(\frac{1}{\sqrt{NT}}\right),$$
$$C_{2it\lambda} = \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\tilde{X}_{js}\tilde{u}_{js}\sum_{n=1}^{r_1}\lambda_n\bar{1}_{n,jsit}\tilde{L}_{jsit,\tilde{\lambda}} = O_P\left(\frac{\|\bar{\lambda}\|}{\sqrt{NT}}\right).$$

Based on the above discussions, applying Lemma B.1 twice to the term on RHS of (B.28) gives

$$\hat{\beta}_{it} - \bar{\beta}_{it} = \left(A_{1it}^{-1} - A_{1it}^{-1}A_{2it\lambda}A_{1it}^{-1}\right)\left(B_{2it\lambda} + C_{1it} + C_{2it\lambda}\right) + O_P\left(\frac{\|\bar{\lambda}\|^2}{\sqrt{NT}}\right) + O_P\left(\|\bar{\lambda}\|^3\right). \quad \text{(B.29)}$$

Write

$$\begin{aligned}
CV(\lambda) &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{X}_{it}'(\bar{\beta}_{it} - \hat{\beta}_{it})\right)^2 + \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{it}'(\bar{\beta}_{it} - \hat{\beta}_{it})\tilde{u}_{it} + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}^2 \\
&\quad + \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{X}_{it}'(\bar{\beta}_{it} - \hat{\beta}_{it}) + \tilde{u}_{it}\right)\gamma_{it} + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\gamma_{it}^2 \\
&\equiv CV_1(\lambda) + CV_2(\lambda) + CV_3 + CV_4(\lambda) + CV_5(\lambda),
\end{aligned}$$

where $\gamma_{it} = \frac{1}{T_{it}}\sum_{s=1}^{T} X_{is}'\left(\beta(Z_{is}) - \beta(Z_{it})\right)L_{is,it}^{p}$. In connection with the construction of γ_{it}, we are able to obtain that $CV_4(\lambda) = O_P\left(\|\bar{\lambda}\|^p\right)$ and $CV_5(\lambda) = O\left(\|\bar{\lambda}\|^{2p}\right)$. Replacing $\hat{\beta}_{it} - \bar{\beta}_{it}$ with (B.29) in $CV_1(\lambda)$ and $CV_2(\lambda)$ gives

$$\begin{aligned}
CV_1(\lambda) &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{X}_{it}'(\hat{\beta}_{it} - \bar{\beta}_{it})\right)^2 \\
&= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{\tilde{X}_{it}'\left(A_{1it}^{-1}A_{2it\lambda}A_{1it}^{-1} - A_{1it}^{-1}\right)\left(B_{2it\lambda} + C_{1it} + C_{2it\lambda}\right)\right\}^2 + O_P\left(\frac{\|\bar{\lambda}\|^2}{\sqrt{NT}}\right) + O_P\left(\|\bar{\lambda}\|^3\right) \\
&= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(D_{3it}^2 - 2D_{1it}D_{2it} + 2D_{2it}D_{3it}\right) \\
&\quad + O_P\left(\frac{\|\bar{\lambda}\|^2}{\sqrt{NT}}\right) + O_P\left(\|\bar{\lambda}\|^3\right) + \text{ terms independent of } \lambda,
\end{aligned}$$

where $D_{1it} = \tilde{X}_{it}'A_{1it}^{-1}\left(A_{2it\lambda}A_{1it}^{-1}C_{1it} - C_{2it\lambda}\right)$, $D_{2it} = \tilde{X}_{it}'A_{1it}^{-1}C_{1it}$ and $D_{3it} = \tilde{X}_{it}'A_{1it}^{-1}B_{2it\lambda}$.

$$\begin{aligned}
CV_2(\lambda) &= \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}\tilde{X}_{it}'(\bar{\beta}_{it} - \hat{\beta}_{it}) \\
&= \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}\tilde{X}_{it}'A_{1it}^{-1}A_{2it,\lambda}A_{1it}^{-1}\left(B_{2it,\lambda} + C_{1it} + C_{2it,\lambda}\right) \\
&\quad - \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}\tilde{X}_{it}'A_{1it}^{-1}\left(B_{2it,\lambda} + C_{1it} + C_{2it,\lambda}\right) + O_P\left(\frac{\|\bar{\lambda}\|^2}{NT}\right) + O_P\left(\frac{\|\bar{\lambda}\|^3}{\sqrt{NT}}\right).
\end{aligned}$$

Then it is easy to know that the leading term of $CV_2(\lambda)$ is

$$-\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}u_{it}X_{it}'A_{1it}^{-1}B_{2it,\lambda}=O_P\left(\frac{\|\bar{\lambda}\|}{\sqrt{NT}}\right).$$

For $CV_1(\lambda)$, the leading terms are

$$\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}D_{2it}D_{3it}=O_P\left(\frac{\|\bar{\lambda}\|}{\sqrt{NT}}\right)\quad\text{and}\quad\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}D_{3it}^2=O_P\left(\|\bar{\lambda}\|^2\right).$$

Note that the two leading terms $\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}D_{2it}D_{3it}$ and $-\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{it}\tilde{X}_{it}'A_{1it}^{-1}B_{2it,\lambda}$ cannot cancel each other as in proving Theorem 2.1 in the presence of irrelevant covariates. Thus, the first result of this theorem follows.

2). We now investigate the asymptotic behaviour of $\hat{\lambda}_s$ for $s=r_1+1,\ldots,r$. Based on the first result of this theorem, we know that

$$\begin{aligned}CV_1(\lambda)&=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(D_{1it}-D_{2it}-D_{3it}\right)^2+O_P\left(\frac{\|\bar{\lambda}\|^2}{\sqrt{NT}}\right)+O_P\left(\|\bar{\lambda}\|^3\right)\\&=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(D_{1it}-D_{2it}-D_{3it}\right)^2+o_P\left(\frac{1}{NT}\right).\end{aligned}$$

For simplicity, let $\Psi(\bar{Z}_{it})=p(\bar{Z}_{it})(\Sigma_X(\bar{Z}_{it})-\mu_X(\bar{Z}_{it})\mu_X(\bar{Z}_{it})')$. We first consider $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}D_{3it}^2$.

$$\begin{aligned}&\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}D_{3it}^2=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{X}_{it}'A_{1it}^{-1}B_{2it\lambda}\right)^2\\&=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{X}_{it}'\Psi^{-1}(\bar{Z}_{it})E[\tilde{L}_{jsit,\tilde{\lambda}}|\tilde{Z}_{it}]^{-1}B_{2it\lambda}\right)^2+o_P(\|\bar{\lambda}\|^2)\\&=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big(\tilde{X}_{it}'\Psi^{-1}(\bar{Z}_{it})E[\tilde{L}_{jsit,\tilde{\lambda}}|\tilde{Z}_{it}]^{-1}\\&\quad\cdot\sum_{n=1}^{r_1}\lambda_nE[X_{js}X_{js}'(\beta(\bar{Z}_{js})-\beta(\bar{Z}_{it}))\bar{1}_{n,jsit}|\bar{Z}_{it}]\cdot E[\tilde{L}_{jsit,\tilde{\lambda}}|\tilde{Z}_{it}]\Big)^2+o_P(\|\bar{\lambda}\|^2)\\&=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{X}_{it}'\Psi^{-1}(\bar{Z}_{it})\cdot\sum_{n=1}^{r_1}\lambda_nE[X_{js}X_{js}'(\beta(\bar{Z}_{js})-\beta(\bar{Z}_{it}))\bar{1}_{n,jsit}|\bar{Z}_{it}]\right)^2+o_P(\|\bar{\lambda}\|^2)\\&=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{X}_{it}'\Psi^{-1}(\bar{Z}_{it})\cdot\sum_{n=1}^{r_1}\lambda_nE[X_{js}X_{js}'(\beta(\bar{Z}_{js})-\beta(\bar{Z}_{it}))\bar{1}_{n,jsit}|\bar{Z}_{it}]\right)^2+o_P\left(\frac{1}{NT}\right)\end{aligned}$$

where the second equality follows from (2) of Lemma B.2, Assumption B and $B_{2it\lambda}=O_P(\|\bar{\lambda}\|^2)$; the third equality follows from a similar procedure as used for proving (2) of Lemma B.2 and Assumption B; the fifth equality follows from the first result of this theorem. Note that $E[\tilde{L}_{jsit,\tilde{\lambda}}|\tilde{Z}_{it}]$ gets canceled above. Therefore, the leading term on RHS of the above equation is unrelated with $\tilde{\lambda}$ and the remaining terms have an order of magnitude of $o_P\left(\frac{1}{NT}\right)$. Also we know that $D_{1it}^2=o_P\left(\frac{1}{NT}\right)$, $D_{1it}D_{2it}=o_P\left(\frac{1}{NT}\right)$ and $D_{1it}D_{3it}=o_P\left(\frac{1}{NT}\right)$ due to the first result of this theorem. Then we can further write

$$CV_1(\lambda)=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(2D_{2it}D_{3it}+D_{2it}^2\right)+o_P\left(\frac{1}{NT}\right)+\text{ terms unrelated to }\tilde{\lambda}.$$

Note that both of $2D_{2it}D_{3it}$ and D_{2it}^2 have an order of magnitude of $O_P\left(\frac{1}{NT}\right)$.

We now further investigate the leading terms of $CV_1(\lambda)$.

$$\begin{aligned}
&\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{2it}^2 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{X}_{it}' A_{1it}^{-1} C_{1it} C_{1it}' A_{1it}^{-1} \tilde{X}_{it} \\
&= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{X}_{it}' \Psi^{-1}(\bar{Z}_{it}) C_{1it} C_{1it}' \Psi^{-1}(\bar{Z}_{it}) \tilde{X}_{it} E[\tilde{L}_{jsit,\tilde{\lambda}} | \tilde{Z}_{it}]^{-2} + o_P\left(\frac{1}{NT}\right) \\
&= \frac{1}{N^3T^3}\sum_{i,t}\sum_{j,s}\sum_{k,r} \tilde{X}_{it}' \Psi^{-1}(\bar{Z}_{it}) \tilde{X}_{js} \tilde{u}_{js} \bar{1}_{jsit} \tilde{L}_{jsit,\tilde{\lambda}} \tilde{X}_{kr}' \tilde{u}_{kr} \bar{1}_{krit} \tilde{L}_{krit,\tilde{\lambda}} \Psi^{-1}(\bar{Z}_{it}) \tilde{X}_{it} E[\tilde{L}_{jsit,\tilde{\lambda}} | \tilde{Z}_{it}]^{-2} \\
&\quad + o_P\left(\frac{1}{NT}\right) \\
&= o_P\left(\frac{1}{NT}\right) + \frac{1}{N^3T^3}\sum_{i,t}\sum_{j,s} \tilde{X}_{it}' \Psi^{-1}(\bar{Z}_{it}) \tilde{X}_{js} \tilde{X}_{js}' \tilde{u}_{js}^2 \bar{1}_{jsit} \tilde{L}_{jsit,\tilde{\lambda}}^2 \Psi^{-1}(\bar{Z}_{it}) \tilde{X}_{it} E[\tilde{L}_{jsit,\tilde{\lambda}} | \tilde{Z}_{it}]^{-2} \\
&\quad + \frac{1}{N^3T^3}\sum_{i,t}\sum_{j,s}\sum_{k,r\neq j,s} \tilde{X}_{it}' \Psi^{-1}(\bar{Z}_{it}) \tilde{X}_{js} \tilde{u}_{js} \bar{1}_{jsit} \tilde{L}_{jsit,\tilde{\lambda}} \tilde{X}_{kr}' \tilde{u}_{kr} \bar{1}_{krit} \tilde{L}_{krit,\tilde{\lambda}} \Psi^{-1}(\bar{Z}_{it}) \tilde{X}_{it} E[\tilde{L}_{jsit,\tilde{\lambda}} | \tilde{Z}_{it}]^{-2} \\
&\equiv H_{1,NT} + H_{2,NT} + o_P\left(\frac{1}{NT}\right),
\end{aligned}$$

where the second equality follows from (2) of Lemma B.2, Assumption B and $C_{1it} = O_P\left(\frac{1}{\sqrt{NT}}\right)$.

Applying a similar procedure as used for deriving $CV_1(\lambda)$ in Lemma 2.1.1, we can obtain

$$H_{1,NT} = \frac{1}{NT} C \cdot E\left[E[\tilde{L}_{jsit,\tilde{\lambda}}^2 | \tilde{Z}_{it}] \cdot E[\tilde{L}_{jsit,\tilde{\lambda}} | \tilde{Z}_{it}]^{-2}\right] + o_P\left(\frac{1}{NT}\right), \tag{B.30}$$

where by the construction of $H_{1,NT}$ it is easy to know that C is a positive constant. Note that $E[\tilde{L}_{jsit,\tilde{\lambda}}^2 | \tilde{Z}_{it}] \geq E[\tilde{L}_{jsit,\tilde{\lambda}} | \tilde{Z}_{it}]^2$, where the equality holds if and only if $\lambda_s = 1$ for all $s = r_1 + 1, \ldots, r$. Hence, $H_{1,NT}$ is minimized at the upper bound values for $\lambda_s = 1$ for all $s = r_1 + 1, \ldots, r$.

For the term $\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{2it}D_{3it} = O_P\left(\frac{1}{NT}\right)$, denote $H_{3,NT} = \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} D_{2it}D_{3it}$.

For the term CV_2, by the first result of this theorem we further write

$$\begin{aligned}
CV_2(\lambda) &= \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{u}_{it} \tilde{X}_{it}' A_{1it}^{-1} A_{2it,\lambda} A_{1it}^{-1} \left(B_{2it,\lambda} + C_{1it} + C_{2it,\lambda}\right) \\
&\quad - \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{u}_{it} \tilde{X}_{it}' A_{1it}^{-1} \left(B_{2it,\lambda} + C_{1it} + C_{2it,\lambda}\right) + O_P\left(\frac{\|\bar{\lambda}\|^2}{NT}\right) + O_P\left(\frac{\|\bar{\lambda}\|^3}{\sqrt{NT}}\right) \\
&= -\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{u}_{it} \tilde{X}_{it}' A_{1it}^{-1} B_{2it,\lambda} - \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{u}_{it} \tilde{X}_{it}' A_{1it}^{-1} C_{1it} + o_P\left(\frac{1}{NT}\right) \\
&= H_{4,NT} + H_{5,NT} + o_P\left(\frac{1}{NT}\right).
\end{aligned}$$

Therefore,

$$CV(\lambda) = H_{1,NT} + H_{2,NT} + H_{3,NT} + H_{4,NT} + H_{5,NT} + o_P\left(\frac{1}{NT}\right), \tag{B.31}$$

where $H_{1,NT}$ to $H_{5,NT}$ all contain $\tilde{\lambda}$. Moreover, based on the first result of this theorem, it is easy to know that $H_{1,NT}$ to $H_{5,NT}$ all have an order of magnitude of $O_P\left(\frac{1}{NT}\right)$ and $H_{1,NT}$ is minimized at

$\lambda_s = 1$ for all $s = r_1 + 1, \ldots, r$. By a similar argument as in Li et al. (2013, p. 578), the second result of this theorem holds. ■

Proof of Theorem 2.2.2:

The kernel functions for the relevant and irrelevant covariates are given as follows.

$$\bar{L}_{js\bar{\lambda}} = \prod_{s=1}^{r_1} \hat{\lambda}_s^{1(Z_{it,s} \neq z_s)} \quad \text{and} \quad \tilde{L}_{js\tilde{\lambda}} = \prod_{s=r_1+1}^{r} \lambda_s^{1(Z_{it,s} \neq z_s)},$$

where $\hat{\lambda}_s$ for $s = 1, \ldots, r_1$ is the estimate of λ_s by minimizing the CV criterion function; and λ_s for $s = r_1 + 1, \ldots, r$ is any arbitrary constant belonging to $[0, 1]$.

Denote that $\hat{\bar{\lambda}} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_{r_1})'$, $\bar{1}_{\bar{Z}_{it},\bar{z}} = 1(\bar{Z}_{it} = \bar{z})$ and $\bar{1}_{n,\bar{Z}_{js},\bar{z}} = 1(Z_{it,n} \neq z_n) \prod_{m=1,m\neq n}^{r_1} 1(Z_{it,m} = z_m)$ for $n = 1, \ldots, r_1$. Thus, write

$$\begin{aligned}
\hat{\beta}(z) - \beta(\bar{z}) = A_{0,NT}^{-1} \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} \tilde{X}_{js} \tilde{X}'_{js} (\beta(\bar{Z}_{js}) - \beta(\bar{z})) \bar{L}_{js\bar{\lambda}} \tilde{L}_{it\tilde{\lambda}} \\
+ A_{0,NT}^{-1} \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} \tilde{X}_{js} \tilde{u}_{js} \bar{L}_{js\bar{\lambda}} \tilde{L}_{it\tilde{\lambda}} \\
+ A_{0,NT}^{-1} \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} \tilde{X}_{js} \gamma_{js} \bar{L}_{js\bar{\lambda}} \tilde{L}_{it\tilde{\lambda}},
\end{aligned}$$

where $A_{0,NT} = \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} \tilde{X}_{js} \tilde{X}'_{js} \bar{L}_{js\bar{\lambda}} \tilde{L}_{it\tilde{\lambda}}$.

By the proof of Theorem 2.2.1, $A_{0,NT}^{-1} = O_P(1)$, $\frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} \tilde{X}_{js} \tilde{u}_{js} \bar{L}_{js\bar{\lambda}} \tilde{L}_{it\tilde{\lambda}} = O_P\left(\frac{1}{\sqrt{NT}}\right)$ and $\frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} \tilde{X}_{js} \gamma_{js} \bar{L}_{js\bar{\lambda}} \tilde{L}_{it\tilde{\lambda}} = O_P\left(\|\hat{\bar{\lambda}}\|^p\right)$. Thus, we need only to focus on the second term on the RHS of the above equation:

$$\begin{aligned}
& \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} \tilde{X}_{js} \tilde{X}'_{js} (\beta(\bar{Z}_{js}) - \beta(\bar{z})) \left(\bar{1}_{\bar{Z}_{js},\bar{z}} + \sum_{n=1}^{r_1} \hat{\lambda}_n \bar{1}_{n,\bar{Z}_{js},\bar{z}} + O(\|\hat{\bar{\lambda}}\|^2) \right) \tilde{L}_{it\tilde{\lambda}} \\
= & \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} \tilde{X}_{js} \tilde{X}'_{js} (\beta(\bar{Z}_{js}) - \beta(\bar{z})) \bar{1}_{\bar{Z}_{js},\bar{z}} \tilde{L}_{it\tilde{\lambda}} \\
& + \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} \tilde{X}_{js} \tilde{X}'_{js} (\beta(\bar{Z}_{js}) - \beta(\bar{z})) \sum_{n=1}^{r_1} \hat{\lambda}_n \bar{1}_{n,\bar{Z}_{js},\bar{z}} \tilde{L}_{it\tilde{\lambda}} \\
& + O(\|\hat{\bar{\lambda}}\|^2) \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} \tilde{X}_{js} \tilde{X}'_{js} (\beta(\bar{Z}_{js}) - \beta(\bar{z})) \tilde{L}_{it\tilde{\lambda}} \\
= & 0 + \frac{1}{NT} \sum_{j=1}^{N} \sum_{s=1}^{T} X_{js} X'_{js} (\beta(\bar{Z}_{js}) - \beta(\bar{z})) \left(\sum_{n=1}^{r_1} \hat{\lambda}_n \bar{1}_{n,\bar{Z}_{js},\bar{z}} \right) \tilde{L}_{it\tilde{\lambda}} + O_P\left(\frac{1}{NT}\right) = O_P\left(\frac{1}{\sqrt{NT}}\right),
\end{aligned}$$

where the second equality follows from $(\beta(\bar{Z}_{js}) - \beta(\bar{z}))\bar{1}_{\bar{Z}_{js},\bar{z}} = 0$ and Theorem 2.2.1. The proof is then complete. ■

Proof of Theorem 2.3.1:

1). Let $\alpha_{NT} = \frac{1}{\sqrt{NT}}$ and U be an $(m \times q)$ matrix. We want to show that for any given $\epsilon > 0$, there exists a large constant C such that

$$\liminf_N \Pr\left\{ \inf_{\|U\|=C} Q_\tau\left(B_0 + \alpha_{NT} U\right) > Q_\tau(B_0) \right\} = 1 - \epsilon. \tag{B.32}$$

This implies with a probability of at least $1-\epsilon$ that there exists a local minimum in the ball $\{B_0+\alpha_{NT}U : \|U\| \leq C\}$. Hence, there exists a local minimizer such that $\|\hat{B} - B_0\| = O_P(\alpha_{NT})$. The above argument is in the same spirit of the proofs for Theorem 1 of Fan and Li (2001) and Lemma A.1 of Wang and Xia (2009).

For notational simplicity, let U_j be the transpose of the j^{th} row of the matrix U with $j = 1, \ldots, m$ and V_s be the s^{th} column of the matrix U with $s = 1, \ldots, p$; and denote

$$e_j = \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{X}_{it} \left(\tilde{X}_{it}' \beta(\bar{Z}_{it}) - \tilde{X}_{it}' \beta(\bar{z}^j) + \gamma_{it} + \tilde{u}_{it} \right) L(Z_{it}, z^j, \hat{\lambda}),$$

where $\gamma_{it} = \frac{1}{T_{it}} \sum_{s=1}^{T} X_{is}' \left(\beta(\bar{Z}_{is}) - \beta(\bar{Z}_{it}) \right) L_{is,it}^p$. By the proofs of Theorems 2.1.2 and 2.2.2, it is easy to know that $e_j = O_P(1)$ uniformly in j due to the fact that $\mathcal{D}$ is compact.

Then we write

$$\begin{aligned}
& Q_\tau(B_0 + \alpha_{NT} U) - Q_\tau(B_0) \\
=& \sum_{j=1}^{m} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\tilde{X}_{it}' \beta(\bar{Z}_{it}) + \gamma_{it} + \tilde{u}_{it} - \tilde{X}_{it}' \beta(\bar{z}^j) - \alpha_{NT} \tilde{X}_{it}' U_j \right)^2 L(Z_{it}, z^j, \hat{\lambda}) \\
& + \sum_{s=1}^{q^*} \tau_s \|b_{0s} + \alpha_{NT} V_s\| + \sum_{s=q^*+1}^{q} \tau_s \|\alpha_{NT} V_s\| \\
& - \sum_{j=1}^{m} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\tilde{X}_{it}' \beta(\bar{Z}_{it}) + \gamma_{it} + \tilde{u}_{it} - \tilde{X}_{it}' \beta(\bar{z}^j) \right)^2 L(Z_{it}, z^j, \hat{\lambda}) - \sum_{s=1}^{q^*} \tau_s \|b_{0s}\| \\
=& \sum_{j=1}^{m} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\alpha_{NT} \tilde{X}_{it}' U_j \right)^2 L(Z_{it}, z^j, \hat{\lambda}) + \sum_{s=q^*+1}^{q} \tau_s \|\alpha_{NT} V_s\| + \sum_{s=1}^{q^*} \tau_s \left(\|b_{0s} + \alpha_{NT} V_s\| - \|b_{0s}\| \right) \\
& - 2 \sum_{j=1}^{m} \sum_{i=1}^{N} \sum_{t=1}^{T} \alpha_{NT} U_j' \tilde{X}_{it} \left(\tilde{X}_{it}' \beta(\bar{Z}_{it}) - \tilde{X}_{it}' \beta(\bar{z}^j) + \gamma_{it} + \tilde{u}_{it} \right) L(Z_{it}, z^j, \hat{\lambda}) \\
\geq& \sum_{j=1}^{m} \sum_{i=1}^{N} \sum_{t=1}^{T} \alpha_{NT}^2 U_j' \tilde{X}_{it} \tilde{X}_{it}' U_j L(Z_{it}, z^j, \hat{\lambda}) + \sum_{s=1}^{q^*} \tau_s \left(\|b_{0s} + \alpha_N V_s\| - \|b_{0s}\| \right) \\
& - 2 \sum_{j=1}^{m} \sum_{i=1}^{N} \sum_{t=1}^{T} \alpha_{NT} U_j' \tilde{X}_{it} \left(\tilde{X}_{it}' \beta(\bar{Z}_{it}) - \tilde{X}_{it}' \beta(\bar{z}^j) + \gamma_{it} + \tilde{u}_{it} \right) L(Z_{it}, z^j, \hat{\lambda}) \\
\geq& \frac{\rho_1}{2} \sum_{j=1}^{m} \|U_j\|^2 - 2 \sum_{j=1}^{m} U_j' e_j + \sum_{s=1}^{q^*} \tau_s \left(\|b_{0s} + \alpha_{NT} V_s\| - \|b_{0s}\| \right) \\
\geq& \frac{\rho_1}{2} \sum_{j=1}^{m} \|U_j\|^2 - 2 \sum_{j=1}^{m} U_j' e_j - O(1) \sum_{s=1}^{q^*} \tau_s \frac{1}{\sqrt{NT}} \|V_s\|,
\end{aligned}$$

where the second inequality follows from (2) of Lemma B.2 and Assumption C; and the third inequality

follows from the Mean Value Theorem. Note that $\|U\| = C$, so we can further write

$$\begin{aligned}
&Q_\tau(B_0 + \alpha_{NT}U) - Q_\tau(B_0) \\
&\geq \frac{\rho_1}{2}\sum_{j=1}^{m}\|U_j\|^2 - 2\sum_{j=1}^{m}U_j'e_j - O(1)\sum_{s=1}^{q^*}\tau_s\frac{1}{\sqrt{NT}}\|V_s\| \\
&\geq \frac{\rho_1}{2}\sum_{j=1}^{m}\|U_j\|^2 - 2\left(\sum_{j=1}^{m}\|U_j\|^2\sum_{j=1}^{m}\|e_j\|^2\right)^{1/2} - O(1)\sum_{s=1}^{q^*}\tau_s\frac{1}{\sqrt{NT}}\|V_s\| \\
&\geq \frac{\rho_1}{2}C^2 - 2C\left(\sum_{j=1}^{m}\|e_j\|^2\right)^{1/2} - O(1)\frac{1}{\sqrt{NT}}\|\tau^*\|\left(\sum_{s=1}^{q^*}\|V_s\|^2\right)^{1/2} \\
&= \frac{\rho_1}{2}C^2 - 2C\left(\sum_{j=1}^{m}\|e_j\|^2\right)^{1/2} - O(1)C, \qquad \text{(B.33)}
\end{aligned}$$

where $\frac{1}{\sqrt{NT}}\|\tau^*\| = O(1)$ by the condition given in this theorem and $\|e_j\| = O_P(1)$ uniformly in j. Note that $\frac{\rho_1}{2}C^2$ is a quadratic function in C while the remaining terms on RHS of (B.33) are linear in C. Since C can be sufficiently large, it is easy to know that RHS of (B.33) is positive with an arbitrary probability close to 1. The proof for (B.32) is now complete. ■

2). For simplicity, we show that $\Pr(\|\hat{b}_{\tau,q}\| = 0) \to 1$ only. The proofs for $\hat{b}_{\tau,j}$ with $j = q^*+1, \ldots, q-1$ are the same. If $\|\hat{b}_{\tau,q}\| \neq 0$, $\hat{B}_\tau$ must satisfy the following equation

$$0 = \frac{\partial}{\partial b_q}Q_\tau(B) = A_1 + A_2, \qquad \text{(B.34)}$$

where

$$A_1 = -\sum_{i=1}^{N}\sum_{t=1}^{T}2\tilde{X}_{it,q}\left((\tilde{Y}_{it} - \tilde{X}_{it}'\hat{\beta}_{\tau,1})L(Z_{it}, z^1, \hat{\lambda}), \ldots, (\tilde{Y}_{it} - \tilde{X}_{it}'\hat{\beta}_{\tau,m})L(Z_{it}, z^m, \hat{\lambda})\right)'$$

and $A_2 = \frac{\tau_q}{\|\hat{b}_{\tau,q}\|}\hat{b}_{\tau,q}$. For $s = 1, \ldots, m$, we can further write each element of A_1 as follows:

$$\begin{aligned}
\frac{1}{\sqrt{NT}}A_{1,s} &= -\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}2\tilde{X}_{it,q}\left(\tilde{X}_{it}'(\beta(\bar{Z}_{it}) - \hat{\beta}_{\tau,s}) + \gamma_{it} + \tilde{u}_{it}\right) \\
&= -\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}2\tilde{X}_{it,q}\tilde{X}_{it}'(\beta(\bar{Z}_{it}) - \hat{\beta}_{\tau,s})L(Z_{it}, z^s, \hat{\lambda}) \\
&\quad -\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}2\tilde{X}_{it,q}(\gamma_{it} + \tilde{u}_{it})L(Z_{it}, z^s, \hat{\lambda}) \\
&= -\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}2\tilde{X}_{it,q}\tilde{X}_{it}'(\beta(\bar{Z}_{it}) - \beta(\bar{z}^s))L(Z_{it}, z^s, \hat{\lambda}) \\
&\quad -\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}2\tilde{X}_{it,q}\tilde{X}_{it}'(\beta(\bar{z}^s) - \hat{\beta}_{\tau,s})L(Z_{it}, z^s, \hat{\lambda}) + O_P(1) = O_P(1),
\end{aligned}$$

where the third equality follows from the proof of the first result of this theorem; and the fourth equality follows from Theorem 2.1.1 (or 2.2.1) and the first result of this theorem.

On the other hand, $\left\|\frac{1}{\sqrt{NT}}A_2\right\| \geq \frac{1}{\sqrt{NT}}\min_{s\in\{q^*+1,\ldots,q\}}\tau_s \geq \omega_2$ by the condition given in the theorem, where ω_2 is sufficiently large. Therefore, $\Pr(\|A_1\| < \|A_2\|) \to 1$, which implies that, with a probability tending to 1, (B.34) does not hold. The above analysis implies that $\hat{b}_{\tau,q}$ must be located at a place where the objective function (2.7) is not differentiable with respect to b_q. Since equation (2.7) of the main file is not differentiable with respect to b_q only at the origin, we immediately obtain that $\Pr(\|\hat{b}_{\tau,q}\| = 0) \to 1$. In a similar fashion, we can show that $\Pr(\hat{b}_{\tau,j} = 0) \to 1$ with $j = q^*+1,\ldots,q-1$. The proof is then complete. ■

Proof of Theorem 2.3.2:

By Theorem 2.3.1, we know that $\|\hat{b}_{\tau,s}\| = 0$ for $s = q^*+1,\ldots,q$ with a probability tending to one. After some simple algebra, we can obtain the first derivative of $Q_\tau(B)$ with respect to β_j for $j = 1,\ldots,m$. Then it is easy to know that $\hat{\beta}_{\tau,jU}$ must be the solution of the following equation

$$\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{itU}\left(\tilde{Y}_{it} - \tilde{X}'_{itU}\hat{\beta}_{\tau,jU}\right)L(Z_{it}, z^j, \hat{\lambda}) + \frac{1}{NT}D\hat{\beta}_{\tau,jU} = 0,$$

where $\tilde{X}_{itU} = (\tilde{X}_{it,1},\ldots,\tilde{X}_{it,q^*})'$ and $D = \mathrm{diag}\Big(\tau_1\|\hat{b}_{\tau,1}\|^{-1},\ldots,\tau_{q^*}\|\hat{b}_{\tau,q^*}\|^{-1}\Big)$. It implies that $\hat{\beta}_{\tau,jU}$ must have the form

$$\hat{\beta}_{\tau,jU} = \left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{itU}\tilde{X}'_{itU}L(Z_{it}, z^j, \hat{\lambda}) + \frac{1}{2NT}D\right)^{-1}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{itU}\tilde{Y}_{it}L(Z_{it}, z^j, \hat{\lambda}).$$

In contrast, the oracle estimator has the following form

$$\left\|\hat{\beta}_{\tau,jU} - \hat{\beta}_{ora}(\bar{z}^j)\right\| \leq \left\|\Sigma_{NT}(z^j)\right\|\left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{itU}\tilde{Y}_{it}L(Z_{it}, z^j, \hat{\lambda})\right\|, \tag{B.35}$$

where

$$\Sigma_{NT}(z^j) = \left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{itU}\tilde{X}'_{itU}L(Z_{it}, z^j, \hat{\lambda}) + \frac{1}{2NT}D\right)^{-1} - \left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{itU}\tilde{X}'_{itU}L(Z_{it}, z^j, \hat{\lambda})\right)^{-1}.$$

Since $\Sigma_{NT}(z^j)$ has finite dimensions, it is easy to know that the rate of $\left\|\Sigma_{NT}(z^j)\right\|$ converging to 0 is the same as

$$\left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{itU}\tilde{X}'_{itU}L(Z_{it}, z^j, \hat{\lambda}) + \frac{1}{2NT}D - \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{itU}\tilde{X}'_{itU}L(Z_{it}, z^j, \hat{\lambda})\right\|$$
$$= \left\|\frac{1}{2NT}D\right\| = O_P\left(\frac{\|\tau^*\|}{NT}\right).$$

Moreover, as with the proof of Theorem 2.1.1 (or 2.2.1), $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{X}_{itU}\tilde{Y}_{it}L(Z_{it}, z^j, \hat{\lambda}) = O_P(1)$. Therefore, for $j = 1,\ldots,m$, $\left\|\hat{\beta}_{\tau,jU} - \hat{\beta}_{ora}(\bar{z}^j)\right\| = O_P\left(\frac{\|\tau^*\|}{NT}\right)$. The proof is now complete. ■

Proof of Theorem 2.3.3:

1). For an arbitrary model S, we say it is under-fitted if it misses at least one variable with a nonzero coefficient (i.e. $S \subset \mathcal{A}^c$ but $\mathcal{A}^c \neq S$); it is over fitted if S covers all relevant variables but also includes at least one redundant regressor (i.e. $\mathcal{A}^c \subset S$ but $\mathcal{A}^c \neq S$). Then, depending on whether the model S is under fitted, correctly fitted, or over fitted, we create three mutually exclusive sets $A^- = \{\tilde{\tau} \in \mathbb{R} : S \subset \mathcal{A}^c, S \neq \mathcal{A}^c\}$, $A^0 = \{\tilde{\tau} \in \mathbb{R} : S = \mathcal{A}^c\}$ and $A^+ = \{\tilde{\tau} \in \mathbb{R} : S \supset \mathcal{A}^c, S \neq \mathcal{A}^c\}$. Suppose that $\tilde{\beta}_j$ for $j = 1, \ldots, m$ are unregularized estimates and there is a sequence $\{\hat{\tau}_{NT}\}$ that ensures (2.15) of the main file satisfies the conditions required by Theorem 2.3.1 (e.g. those used in Monte Carlo study).

Case 1: In this case, we consider under-fitted models, where $S \subset \mathcal{A}^c$ but $\mathcal{A}^c \neq S$. Without losing generality, we assume that only one variable is missing, so we assume that the first $q^* - 1$ elements of $\hat{\beta}_{\tilde{\tau},j}$ are obtained from the under-fitted model and the remaining $q - q^* + 1$ elements of $\hat{\beta}_{\tilde{\tau},j}$ are 0.

We then write

$$
\begin{aligned}
RSS_{\tilde{\tau}} &= \frac{1}{NT}\sum_{j=1}^{m}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{Y}_{it} - \tilde{X}_{it}'\hat{\beta}_{\tilde{\tau},j}\right)^2 L(Z_{it}, z^j, \hat{\lambda}) \\
&= \frac{1}{NT}\sum_{j=1}^{m}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{Y}_{it} - \tilde{X}_{it}'\tilde{\beta}_j + \tilde{X}_{it}'\tilde{\beta}_j - \tilde{X}_{it}'\hat{\beta}_{\tilde{\tau},j}\right)^2 L(Z_{it}, z^j, \hat{\lambda}) \\
&= \frac{1}{NT}\sum_{j=1}^{m}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{Y}_{it} - \tilde{X}_{it}'\tilde{\beta}_j\right)^2 L(Z_{it}, z^j, \hat{\lambda}) \\
&\quad + \frac{1}{NT}\sum_{j=1}^{m}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{X}_{it}'\tilde{\beta}_j - \tilde{X}_{it}'\hat{\beta}_{\tilde{\tau},j}\right)^2 L(Z_{it}, z^j, \hat{\lambda}) \\
&\quad + \frac{2}{N}\sum_{j=1}^{m}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{\beta}_j - \hat{\beta}_{\tilde{\tau},j}\right)'\tilde{X}_{it}\left(\tilde{Y}_{it} - \tilde{X}_{it}'\tilde{\beta}_j\right) L(Z_{it}, z^j, \hat{\lambda}) \\
&= \frac{1}{NT}\sum_{j=1}^{m}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{Y}_{it} - \tilde{X}_{it}'\tilde{\beta}_j\right)^2 L(Z_{it}, z^j, \hat{\lambda}) \\
&\quad + \frac{1}{NT}\sum_{j=1}^{m}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{X}_{it}'\tilde{\beta}_j - \tilde{X}_{it}'\hat{\beta}_{\tilde{\tau},j}\right)^2 L(Z_{it}, z^j, \hat{\lambda}) \\
&\equiv RSS^* + R_{2\tilde{\tau}},
\end{aligned}
$$

where the fourth equality is due to the construction of the unregularized estimators.

We now consider $R_{2\tilde{\tau}}$ and write

$$
\begin{aligned}
R_{2\tilde{\tau}} &= \frac{1}{NT}\sum_{j=1}^{m}\sum_{i=1}^{N}\sum_{s=1}^{T}\left(\tilde{\beta}_j - \hat{\beta}_{\tilde{\tau},j}\right)'\tilde{X}_{it}\tilde{X}_{it}' L(Z_{it}, z^j, \hat{\lambda})\left(\tilde{\beta}_j - \hat{\beta}_{\tilde{\tau},j}\right) \\
&= \sum_{j=1}^{m}\left(\tilde{\beta}_j - \hat{\beta}_{\tilde{\tau},j}\right)'\Sigma_1(z^j)\left(\tilde{\beta}_j - \hat{\beta}_{\tilde{\tau},j}\right) + o_P(1) \\
&\geq \sum_{j=1}^{m}\rho_{min}(\Sigma_1(z^j))\left\|\tilde{\beta}_j - \hat{\beta}_{\tilde{\tau},j}\right\|^2 + o_P(1)
\end{aligned}
$$

$$= O(1)\sum_{j=1}^{m}\left\|\tilde{\beta}_j - \hat{\beta}_{\tilde{\tau},j}\right\|^2 + o_P(1) \geq O(1)\sum_{j=1}^{m}\tilde{\beta}_{j,q^*}^2 + o_P(1),$$

where $\Sigma_1(z^j) = \Sigma_{XX}(\bar{z})E[\tilde{L}(\tilde{Z}_{it}, \tilde{z}, \hat{\tilde{\lambda}})|\hat{\tilde{\lambda}}]$; $\rho_{min}(\Sigma_1(z^j))$ denotes the minimum eigenvalue of $\Sigma_1(z^j)$; $\tilde{\beta}_{j,q^*}$ denotes the q^{*th} element of $\tilde{\beta}_j$; the second equality follows from (2) of Lemma B.2 of the Appendix and Theorem 2.1.1 (or 2.2.1); and the first inequality follows from Assumption C.2.

Similarly, we can obtain that $RSS_{\hat{\tau}_{NT}} \equiv RSS^* + R_{2\hat{\tau}_{NT}}$, where

$$\begin{aligned}
R_{2\hat{\tau}_{NT}} &= \frac{1}{NT}\sum_{j=1}^{m}\sum_{i=1}^{N}\sum_{s=1}^{T}\left(\tilde{\beta}_j - \hat{\beta}_{\hat{\tau}_{NT},j}\right)'\tilde{X}_{it}\tilde{X}_{it}'L(Z_{it}, z^j, \hat{\lambda})\left(\tilde{\beta}_j - \hat{\beta}_{\hat{\tau}_{NT},j}\right) \\
&= \sum_{j=1}^{m}\left(\tilde{\beta}_j - \hat{\beta}_{\hat{\tau}_{NT},j}\right)'\Sigma_1(z^j)\left(\tilde{\beta}_j - \hat{\beta}_{\hat{\tau}_{NT},j}\right) + o_P(1) \\
&\leq \sum_{j=1}^{m}\rho_{max}(\Sigma_1(z^j))\left\|\tilde{\beta}_j - \hat{\beta}_{\hat{\tau}_{NT},j}\right\|^2 + o_P(1) \\
&\leq O(1)\sum_{j=1}^{m}\left\|\tilde{\beta}_j - \hat{\beta}_{\hat{\tau}_{NT},j}\right\|^2 + o_P(1) \\
&\leq O(1)\sum_{j=1}^{m}\left\|\tilde{\beta}_j - \beta(\bar{z}^j)\right\|^2 + O(1)\sum_{j=1}^{m}\left\|\beta(\bar{z}^j) - \hat{\beta}_{\hat{\tau}_{NT},j}\right\|^2 = o_P(1),
\end{aligned}$$

where $\rho_{max}(\Sigma_1(z^j))$ denotes the maximum eigenvalue of $\Sigma_1(z^j)$; the second equality follows from (2) of Lemma B.2 of the Appendix and Theorem 2.1.1 (or 2.2.1); the second inequality follows from Assumption C.2; and the last equality follows from Theorem 2.3.1 and the fact that both $\tilde{\beta}_j$ and $\hat{\beta}_{\hat{\tau}_{NT},j}$ are regularized estimators.

Note that by (1) of Lemma B.2 we can obtain that $RSS^* \to_P \sum_{j=1}^{m}\Pr(\bar{z}^j)\sigma_u^2$. Based on the analysis on $R_{2\tilde{\tau}}$ and $R_{2\hat{\tau}_{NT}}$, we then can further conclude that

$$\Pr\left(\inf_{\tilde{\tau}\in A^-} BIC_{\tilde{\tau}} > BIC_{\hat{\tau}_{NT}}\right) \to 1.$$

Case 2: In this case, we consider over-fitted models, where $S \supset \mathcal{A}^c$ but $\mathcal{A}^c \neq S$. Consider $\forall\tilde{\tau} \in A^+$ and recall that $\hat{B}_{\tilde{\tau}}$ determines $S_{\tilde{\tau}}$. Under such a model $S_{\tilde{\tau}}$, we can define another unpenalized estimator $\check{B}_{\tilde{\tau}}$ as

$$\check{B}_{\tilde{\tau}} = \underset{\beta_1,\ldots,\beta_m}{\operatorname{argmin}}\frac{1}{NT}\sum_{j=1}^{m}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{Y}_{it} - \tilde{X}_{it}'\beta_j\right)^2 L(Z_{it}, z^j, \hat{\lambda}),$$

where, for $j = 1, \ldots, m$, $\|\beta_{j,s}\| = 0$ with $\forall s \notin S_{\tilde{\tau}}$ and $\beta_{j,s}$ denotes the s^{th} element of β_j. In other words, $\check{B}_{\tilde{\tau}} = (\check{\beta}_1, \ldots, \check{\beta}_m)'$ is the unregularized estimator under the model determined by $\hat{B}_{\tilde{\tau}}$. By definition, we obtain immediately that $RRS_{\tilde{\tau}} \geq RRS_{S_{\tilde{\tau}}}$, where

$$RRS_{S_{\tilde{\tau}}} = \frac{1}{NT}\sum_{j=1}^{m}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{Y}_{it} - \tilde{X}_{it}'\check{\beta}_j\right)^2 L(Z_{it}, z^j, \hat{\lambda}).$$

It follows that

$$
\begin{aligned}
&\ln RRS_{\tilde{\tau}} - \ln RRS^* \geq \ln RRS_{S_{\tilde{\tau}}} - \ln RRS^* \\
&= \ln\left\{\frac{RRS^*}{RRS^*} + \frac{1}{NT\cdot RRS^*}\sum_{j=1}^{m}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{\beta}_j - \check{\beta}_j\right)' \tilde{X}_{it}\tilde{X}_{it}' L(Z_{it}, z^j, \hat{\lambda})\left(\tilde{\beta}_j - \check{\beta}_j\right)\right\} \\
&\geq -\frac{O(1)}{NT\cdot RRS^*}\sum_{j=1}^{m}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{\beta}_j - \check{\beta}_j\right)' \tilde{X}_{it}\tilde{X}_{it}' L(Z_{it}, z^j, \hat{\lambda})\left(\tilde{\beta}_j - \check{\beta}_j\right) \\
&\geq -\frac{O_P(1)}{RRS^*}\sum_{j=1}^{m}\rho_{max}(\Sigma_1(z^j))\left\|\tilde{\beta}_j - \check{\beta}_j\right\|^2 \\
&\geq -\frac{O_P(1)}{RRS^*}\sum_{j=1}^{m}\rho_{max}(\Sigma_1(z^j))\left\|\tilde{\beta}_j - \beta(\bar{z}^j)\right\|^2 - \frac{O_P(1)}{RRS^*}\sum_{j=1}^{m}\rho_{max}(\Sigma_1(z^j))\left\|\beta(\bar{z}^j) - \check{\beta}_j\right\|^2 \\
&\geq -\left|O_P\left(\frac{1}{NT}\right)\right|,
\end{aligned}
$$

where $\tilde{\beta}_j$ for $j = 1, \ldots, m$ are unregularized estimators as those used in **Case 1**; the second inequality follows from (2) of Lemma B.2 and Theorem 2.1.1 (or 2.2.1); and the fourth inequality follows from Theorem 2.3.1.

Similarly, we can obtain that $\ln RRS_{\hat{\tau}_{NT}} - \ln RRS^* = O_P\left(\frac{1}{NT}\right)$. Thus, we obtain

$$
\ln RRS_{\tilde{\tau}} - \ln RRS_{\hat{\tau}_{NT}} \geq -\left|O_P\left(\frac{1}{NT}\right)\right|.
$$

We then write

$$
\inf_{\tilde{\tau}\in A^+} BIC_{\tilde{\tau}} - BIC_{\hat{\tau}_{NT}} = \ln RRS_{\tilde{\tau}} - \ln RRS_{\hat{\tau}_{NT}} + (df_{\tilde{\tau}} - df_{\hat{\tau}_{NT}})\frac{\ln(NT)}{NT}.
$$

By Theorem 2.3.1, we know that $\Pr(df_{\hat{\tau}_{NT}} \to q^*) = 1$. Since $\tilde{\tau} \in A^+$, we must have that $\Pr(df_{\tilde{\tau}} \geq q^* + 1) \to 1$. Then it is clear that

$$
\Pr\left(\inf_{\tilde{\tau}\in A^+} BIC_{\tilde{\tau}} > BIC_{\hat{\tau}_{NT}}\right) \to 1.
$$

Combining Cases 1 and 2, we obtain that $\Pr\left(\inf_{\tilde{\tau}\in A^-\cup A^+} BIC_{\tilde{\tau}} > BIC_{\hat{\tau}_{NT}}\right) \to 1$, which in turn implies $\Pr\left(S_{\hat{\tilde{\tau}}} \to \mathcal{A}^c\right) = 1$. The proof is complete.

2)-3). The second and third results of this theorem follow by noting that setting $\tilde{\tau}$ to a large constant satisfies all the conditions required by Theorem 2.3.2 and the first result of this theorem. Thus, we have

$$
\hat{\beta}_{\hat{\tilde{\tau}},jU} - \beta_U(\bar{z}^j) = \hat{\beta}_{ora}(\bar{z}^j) - \beta_U(\bar{z}^j) + O_P\left(\frac{1}{NT}\right).
$$

Then the results follow from Theorem 2.1.2 and Theorem 2.2.2 immediately. ■

References

Aitchison, J. and Aitken, C. (1976), 'Multivariate binary discrimination by the kernel method', *Biometrika* **63**(3), 413–420.

Bosq, D. (1996), *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*, Springer, New York.

Chen, J., Gao, J. and Li, D. (2012), 'A new diagnostic test for cross–section uncorrelatedness in nonparametric panel data models', *Econometric Theory* **28**(5), 1–20.

Fan, J. and Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', *Journal of the American Statistical Association* **96**(456), 1348–1360.

Fan, J. and Yao, Q. (2003), *Nonlinear Time Series: Nonparametric and Parametric Methods*, Springer-Verlag.

Gao, J. (2007), *Nonlinear Time Series: Semiparametric and Nonparametric Methods*, Chapman and Hall.

Hunter, D. R. and Li, R. (2005), 'Variable selection using mm algorithms', *The Annals of Statistics* **189**, 1617–1642.

Li, Q., Ouyang, D. and Racine, J. S. (2013), 'Categorical semiparametrics varying–coefficient models', *Journal of Applied Econometrics* **28**(4), 551–579.

Newey, W. K. and McFadden, D. (1994), 'Large sample estimation and hypothesis testing', *Handbook of Econometrics* pp. 2113–2245.

Su, L. and Jin, S. (2012), 'Sieve estimation of panel data models with cross section dependence', *Journal of Econometrics* **169**(1), 34–47.

Wang, H. and Xia, Y. (2009), 'Shrinkage estimation of the varying coefficient', *Journal of the American Statistical Association* **104**(486), 747–757.