Cardiff Economics Working Papers

Working Paper No. E2022/1

Varying Coefficient Model with Correlated Error Components and Application to Disparities Between Mental Health Service by Councils in England

Pipat Wongs-arta, Namhyun Kim, Yingcun Xia, Francesco Moscone

January 2022

ISSN 1749-6010

Cardiff Business School
Cardiff University
Colum Drive
Cardiff CF10 3EU
United Kingdom
t: +44 (0)29 2087 4000
f: +44 (0)29 2087 4419
business.cardiff.ac.uk

Varying Coefficient Model with Correlated Error Components and Application to Disparities Between Mental Health Service by Councils in England

Pipat Wongs-art[a,*], Namhyun Kim[b], Yingcun Xia[c], Francesco Moscone[d]

[a] *Cardiff Business School, Cardiff University, United Kingdom*
[b] *Exeter Business School, University of Exeter, United Kingdom*
[c] *Department of Statistics&Applied Probability, National University of Singapore, Singapore*
[d] *Brunel Business School, Brunel University London, United Kingdom*

Abstract

In this paper, we discuss estimation procedure and various inferential methods for varying coefficient panel data models that include spatially correlated error components. Our estimation procedure is an extension of the quasi-maximum likelihood method for spatial panel data regression to the conditional local kernel-weighted likelihood. We allow both relevant and irrelevant regressors in our model and propose a variable selection procedure that we show to perform well for models that involve spatial error dependence. We also extend our procedure so that it allows empirical modelling and testing of the so-called semi-varying coefficient specification. To ensure the statistical validity of our methods, we derive a set of asymptotic properties based on a collection of primitive assumptions that appear regularly in the nonparametric literature. Finally, we use the proposed model and methods to analyse the municipal disparities in mental health service spending by local authorities in England in order to illustrate practicability and empirical relevance.

Keywords: Spatial models, Error components, Local maximum likelihood, Varying coefficient, Variable selection, Mental health services and expenditures

JEL: C14; C51; C52; G12; G17

*Corresponding authors

Email addresses: `wongsa-artp@cardiff.ac.uk` (Pipat Wongs-art), `n.kim@exeter.ac.uk` (Namhyun Kim)

1. Introduction

Panel data consist of repeated observations over time on the same set of cross-sectional units, which can be individuals, firms, schools, cities, local authorities or any collection of units one can follow over time. In the recent years, in the attempt to answer complex empirical questions many researchers recognise the need to exploit the rich information available in panel data sets. Accordingly, panel data and methods of econometric analysis appropriate to such data have become increasingly important in the discipline. Recently, we have witnessed fast methodological developments in various areas of panel data analysis. This paper focuses on two important areas, namely (i) spatial error dependence (SED), and (ii) varying coefficient panel data models.

Spatial models for panel data are important tools in economics, regional science and geography in analysing a wide range of empirical issues. By far the most widely used spatial models are variants of that originated by Cliff (1973) (see also Cliff and Ord (1981)). Spatial panel data models with spatial autoregressive (SAR) disturbances are considered in Baltagi et al. (2003), Kapoor et al. (2007), Liu and Yang (2015), and Su and Yang (2015) among others. Gao et al. (2020) consider an alternative assumption on the dependence that is shown to be closely related to the SAR. Moreover, varying coefficient models are a useful extension of classical linear models and have been the main focus of many methodological studies in the literature (see e.g. Fan and Zhang (1999), Cai et al. (2000) and Xia et al. (2004)). Varying coefficient panel data models have also attracted considerable attention in the past two decades. For example, Sun et al. (2009) propose a panel data varying coefficient model by imposing a widely-used identification restriction such that the sum of the fixed effects is zero, whereas Rodriguez-Poo and Soberon (2014) propose to use the first difference to remove the fixed effects. Furthermore, Feng et al. (2017) consider varying coefficient panel data models where all covariates are categorical.

So far, methodological developments in these two areas have progressed as two separate directions. In this paper, we establish estimation procedure and various novel inference methods for varying-coefficient panel data models that include spatially correlated error components. We begin by constructing alternative varying-coefficient specifications in which both relevant and irrelevant regressors are included. In addition, our spatial model allows the individual effects themselves to be spatially correlated. This differs from previous studies in the literature (e.g. Baltagi et al. (2012)), who consider only spatial dependence in the error term. Then, we establish the model's estimation procedure that can be viewed as an extension of the quasi-maximum likelihood method for spatial panel data regression (e.g. Lee and Yu (2010), and Liu and Yang (2015)) to the conditional local kernel-weighted likelihood (see e.g. Fan et al. (1998), Cai et al. (2000), and Fan and Zhang (2008)). In this regard, we derive a set of asymptotic results based on a collection of primitive assumptions that also appear in other existing studies (e.g. Robinson (2011)).

Asymptotically, our analysis is tailored for the case where the time dimension is fixed and the cross section dimension tends to infinity. In the other words, T is fixed and $N \to \infty$ and thus is geared towards samples where N is large relative to T as is frequently the case. Moreover, we establish a novel procedure for selecting necessary regressors. Wang and Xia (2009) introduced the so-called Kernel Least Absolute Shrinkage and Selection Operator (KLASSO) technique. We show that this technique is ineffective when applied to the panel data where there exists the Cliff-Ord-type models of spatial error dependence and suggest an alternative procedure. We also extend our procedure to handle selection in a more complex specification known in the literature as the semi-varying coefficient model. Finally, we conduct extensive simulation exercises in order to examine the finite sample performance and robustness of our proposed (estimation and inference) procedures. Importantly, we show that addressing spatial error dependence and using random effects enables important efficiency gain that leads to more effective statistical inference.

The analytical tools developed in this paper can be used for a broad range of applications. To illustrate their empirical relevance and applicability, we apply the newly established model and methods to analyse municipal disparities in mental health service (MHS) spending by councils in England. Our study explains the MHS spending in relation to a set of risk factors (e.g. percentages of male population and of population under 14 year of age) and supply factors of mental health needs (e.g. medians house price and weekly wage). Moreover, we study the interaction between these explanatory variables and some important local authority-specific attributes, namely (i) political preferences and ideology, and (ii) level of total public health expenditure by local authorities. The idea behind the former stems from the hypothesis that some councils may give more weight in terms of resources to some risk factors (e.g. standardised mortality ratio and percentages of male population) and such decision is influenced by political preferences or beliefs within the local authorities, for example whether left- or right-wing political party is in power. On the other hand, studying the latter, in which the MHS is a part of, can help to highlight local authorities' views about each of the explanatory variables. To understand this idea more clearly, let us first recall the Engel's law in economics which suggests that the poorer a family is, the larger the budget share it spends on nourishment. Correspondingly to the Engel's law, a relatively stronger impact of the share of population under 14 on MHS when the total public health expenditure is low, for example, suggests that the variable is considered by the local authorities to be an essential determinant. On the contrary, a relatively smaller impact of the percentage of male population on MHS when the total public health expenditure is low suggests, for instance, that the variable is considered to be important though not essential. A more detailed discussion of these points is presented in Section 4.

The rest of the paper is organised as follows. In Section 2.1, we propose a varying-coefficient panel data model that includes spatially correlated error components. In Sections 2.2, we illustrate the model's estimation procedure and establish its relevant asymptotic properties, whereas in Section 2.3 we introduce the alternative variable selection procedure. In Section 3, we conduct extensive simulation exercises in order to examine the finite sample performance and robustness of our proposed procedures, while we present the application of our model and methods to analyse municipal disparities in mental health service spending by councils in England in Section 4. Finally, Section 5 presents conclusions and further discussion. Technical proofs are provided in the Appendix.

2. Varying coefficient model with spatially correlated error components

We begin by introducing the model specification and basic assumptions.

2.1. Model specification

Let $y_{it} \in \mathbb{R}^1$ be a response of interest, $X_{it} = (X_{it,1}, \ldots, X_{it,D}) \in \mathbb{R}^D$ and $Z_{it} \in [0, 1]$, which are referred to as the D-dimensional "regressors" and "covariate", respectively, in order to differentiate and avoid confusion. For $i = 1, \ldots, N$ and $t = 1, \ldots, T$, we assume that we observe y_{it} generated by

$$y_{it} = X_{it}\beta_0(Z_{it}) + u_{it}, \tag{2.1}$$

where $\beta_0(z) = \{\beta_{1,0}(z), \ldots, \beta_{D,0}(z)\}^\top \in \mathbb{R}^D$ is a vector of smooth nonparametric functions in z and $u_{it} \in \mathbb{R}^1$ denotes an error term of which $E(u_{it}|X_{it}, Z_{it}) = 0$ almost surely. To specify the spatial error dependence we define

$$u_N = (u_{11}, u_{21}, \ldots, u_{N1}, u_{12}, \ldots, u_{N2}, u_{13}, \ldots, u_{NT})^\top,$$

where we have grouped the data by time periods rather than spatial units as commonly done in panel data literature, y_N as a $NT \times 1$ vector of y_{it} with a similar structure, and

$$X_N = (X_{11}^\top, X_{21}^\top, \ldots, X_{N1}^\top, X_{12}^\top, \ldots, X_{N2}^\top, X_{13}^\top, \ldots, X_{NT}^\top)^\top.$$

Accordingly, the model in (2.1) can be expressed in matrix notation as

$$y_N = (B_0 \circ X_N)e_D + u_N, \tag{2.2}$$

where $B_0 = \{\beta_0(Z_{11}), \beta_0(Z_{21}), \ldots, \beta_0(Z_{N1}), \beta_0(Z_{12}), \ldots, \beta_0(Z_{NT})\}^\top$, e_D is $D \times 1$ vector of 1s and "$\circ$" denotes the Hadamard product. We assume that u_N follows the SAR process

$$u_N = (I_T \otimes \rho_0 W_N)u_N + \varepsilon_N, \tag{2.3}$$

where $\otimes$ signifies the Kronecker product, W_N is an $N \times N$ spatial weights matrix which is nonstochastic, ρ_0 is a scalar auto-regressive parameter and ε_N is an $NT \times 1$ vector of innovations. Moreover, ε_N follows a classical one-way error component model (see e.g. Baltagi et al. (2008))

$$\varepsilon_N = (e_T \otimes I_N)\mu_N + v_N,$$

where μ_N denotes a vector of the unit specific error component, e_T is a $T \times 1$ vector of 1s and v_N is an $NT \times 1$ vector of independent and identically distributed (i.i.d) idiosyncratic errors. For the sake of clarity, hereafter we refer to u_N and ε_N as vectors of "disturbance" and "innovation", respectively.

With regard to the model in (2.1), we impose:

Assumption A1. *W_N is row-normalized in the sense that elements in a given row sum up to one and non-stochastic spatial weights matrix with zero diagonal elements.*

Assumption A2. *Let $S_N(\rho) = I_N - \rho W_N$ for an arbitrary ρ. $S_N(\rho)$ is invertible for all $\rho \in P$, where the parameter space P is compact and $\rho_0 \in (-1, 1)$ is in the interior of P.*

Assumption A3. *W_N and $S_N^{-1}(\rho)$ are uniformly bounded in both row and column sums in absolute value.*

Assumption A4. *Let T be a fixed positive integer. In addition, $\{v_{it}\}$, $i = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$, are i.i.d. across i and t with zero mean, variance $\sigma^2_{v,0}$, $E(|v_{it}|^{4+\varsigma}) < \infty$ for some $\varsigma = 2\varrho$, where $0 < \varrho \leq 2$.*

Assumption A5. *The unit specific error components $\{\mu_i\}$, $i = 1, 2, \ldots, N$ are i.i.d. across i with zero mean, variance $\sigma^2_{\mu,0}$ and $E(|\mu_i|^{4+\varsigma}) < \infty$ for some $\varsigma = 2\varrho$, where $0 < \varrho \leq 2$.*

Assumption A6. *The processes $\{v_{it}\}$ and $\{\mu_i\}$ are independent of each other.*

Assumption A1 implies that no unit is a neighbour to itself. Although the elements of W_N are assumed independent of t, the number of neighbours, which a given unit has, may depend on the number of cross-sectional units, N. Assumption A2 ensures that the model is closed in the sense that we can write

$$u_N = [I_T \otimes (I_N - \rho_0 W_N)^{-1}]\varepsilon_N, \tag{2.4}$$

which clearly suggests that our SAR random effect model allows the individual effects themselves to be spatially correlated. This differs from previous studies in the literature (see e.g. Baltagi et al. (2012)) who focus only on spatial dependence on the error term. Assumption A3 restricts the extent of association between the cross sectional units. In practice these are satisfied given that each of the units is associated only with a limited

number of neighbours, or in the other words, if W_N is sparse. Alternatively, they may be satisfied when W_N is not sparse if its elements decline with a distance measure that increases sufficiently rapidly as the sample size increases.

The remaining assumptions are standard and lead to the variance-covariance matrix $E[u_N u_N^\top]$ of the form

$$\Omega^0_{u,N} = [I_T \otimes (I_N - \rho_0 W))^{-1}]\Omega^0_{\varepsilon,N}[I_T \otimes (I_N - \rho_0 W^\top)^{-1}], \tag{2.5}$$

where $\Omega^0_{\varepsilon,N} \equiv E[\varepsilon_N \varepsilon_N^\top] = \sigma^2_{v,0} Q_{0,N} + \sigma^2_{1,0} Q_{1,N}$ and $\sigma^2_{1,0} = \sigma^2_{v,0} + T\sigma^2_{\mu,0}$. As such, $Q_{0,N} = (I_T - (J_T/T)) \otimes I_N$ and $Q_{1,N} = (J_T/T) \otimes I_N$ are symmetric, idempotent and orthogonal to each other, where $J_T = e_T e_T^\top$ is a $T \times T$ matrix of ones, and are standard transformation matrices frequently used in the error component literature.

Alternatively, we can write $\Omega^0_{u,N} = \sigma^2_{v,0}\mathbb{Q}^0_N$, where

$$\mathbb{Q}^0_N = \left[I_T \otimes (I_N - \rho_0 W_N)^{-1}\right] \{Q_{0,N} + (1 + \phi_0 T) Q_{1,N}\} \left[I_T \otimes (I_N - \rho_0 W_N^\top)^{-1}\right] \tag{2.6}$$

and $\phi_0 = \sigma^2_{\mu,0}/\sigma^2_{v,0}$, which suggests that $\mathbb{Q}^0_N = (1/\sigma^2_{v,0})E[u_N u_N^\top]$. In this regard,

$$\left(\mathbb{Q}^0_N\right)^{-1} = \bar{\mathbb{Q}}^{0\top}_N \bar{\mathbb{Q}}^0_N, \tag{2.7}$$

where $\bar{\mathbb{Q}}^0_N = \left\{Q_{0,N} + (1 + T\phi_0)^{-1/2} Q_{1,N}\right\} [I_T \otimes (I_N - \rho_0 W_N)]$ by using the orthogonality of $Q_{0,N}$ and $Q_{1,N}$. We now use these results to establish the model estimation procedure.

2.2. Estimation procedure

To establish the estimation procedure, we need to first introduce a transformation of the original model. Let $\ddot{X}_{0N} = \bar{\mathbb{Q}}^0_N X_N$ and $\ddot{u}_{0N} = \bar{\mathbb{Q}}^0_N y_N$, where $\bar{\mathbb{Q}}^0_N$ is defined in (2.7). Then, the transformed model is written as

$$\ddot{y}_{0N} = (B_0 \circ \ddot{X}_{0N}) e_D + \ddot{u}_{0N}, \tag{2.8}$$

where $\ddot{y}_{0N} = \bar{\mathbb{Q}}^0_N y_N - \{\bar{\mathbb{Q}}^0_N (B_0 \circ X_N) e_D - (B_0 \circ \bar{\mathbb{Q}}^0_N X_N)\}$. Such a model can be viewed as a combination of the Cochrane-Orcutt/RE-GLS transformations in econometrics.

Correspondingly, let $\bar{\mathbb{Q}}_N = \left\{Q_{0,N} + (1 + T\phi)^{-1/2} Q_{1,N}\right\} [I_T \otimes (I_N - \rho W_N)]$ represent an arbitrary expression of $\bar{\mathbb{Q}}^0_N$, and $\ddot{y}_N$ and $\ddot{X}_N$ denote that of $\ddot{y}_{0N}$ and $\ddot{X}_{0N}$, respectively. Also, let $\ddot{X}_{js}$ be the js-th row of $\ddot{X}_N$, $\ddot{u}_{js} = \ddot{y}_{js} - \ddot{X}_{js}\beta$ be the js-th element of $\ddot{u}_N = \ddot{y}_N - (B \circ \ddot{X}_N) e_D$, and

$$\mathbb{Q}_N = \left[I_T \otimes (I_N - \rho W_N)^{-1}\right] \{Q_{0,N} + (1 + \phi T) Q_{1,N}\} \left[I_T \otimes (I_N - \rho W_N^\top)^{-1}\right]$$

be an arbitrary expression of $\mathbb{Q}^0_N$.

For a given bandwidth parameter h and a kernel function $K(\cdot)$ in $K_h(\cdot) = K(\cdot/h)/h$, we can then construct the conditional *local kernel-weighted log-likelihood* as follows

$$\begin{aligned}\ell(\beta, \sigma_v^2, \phi, \rho) &= -\frac{1}{2}\log\{2\pi\sigma_v^2\}\sum_{j=1}^{N}\sum_{s=1}^{T}K_h(Z_{js}-z) + \log\{|\mathbb{Q}_N|\}\sum_{j=1}^{N}\sum_{s=1}^{T}K_h(Z_{js}-z) \\ &\quad - \frac{1}{2\sigma_v^2}\sum_{j=1}^{N}\sum_{s=1}^{T}\{\ddot{y}_{js} - \ddot{X}_{js}\beta\}^2 K_h(Z_{js}-z). \end{aligned} \tag{2.9}$$

Given $\delta = (\phi, \rho)^\top \in \Delta$, where Δ denotes a compact parameter space, the local likelihood function in (2.9) is maximised at

$$\tilde{\beta}(z) = \left[\sum_{j=1}^{N}\sum_{s=1}^{T}\ddot{X}_{js}^\top\ddot{X}_{js}K_h(Z_{js}-z)\right]^{-1}\sum_{j=1}^{N}\sum_{s=1}^{T}\ddot{X}_{js}^\top\ddot{y}_{js}K_h(Z_{js}-z), \tag{2.10}$$

which can also be expressed as $\tilde{\beta}(z) = \left\{\ddot{X}_N^\top K_N \ddot{X}_N\right\}^{-1}\ddot{X}_N^\top K_N \ddot{y}_N$ in matrix notation, where $K_N = \mathrm{diag}\{K_h(Z_{11}-z), \ldots, K_h(Z_{N1}-z), K_h(Z_{12}-z), \ldots, K_h(Z_{NT}-z)\}$, and

$$\tilde{\sigma}_v^2(z) = \left[\sum_{j=1}^{N}\sum_{s=1}^{T}K_h(Z_{js}-z)\{\ddot{y}_{js} - \ddot{X}_{js}\tilde{\beta}(z)\}^2\right]\left[\sum_{j=1}^{N}\sum_{s=1}^{T}K_h(Z_{js}-z)\right]^{-1}.$$

These suggest formulating the concentrated log-likelihood $\tilde{\ell}_z^c(\delta) \equiv \max_{\beta,\sigma_v^2}\ \ell(\beta, \sigma_v^2, \phi, \rho)$ as follows

$$\tilde{\ell}_z^c(\delta) = -\frac{1}{2}\left[\log(2\pi\tilde{\sigma}_v^2) + \log|\mathbb{Q}_N| + 1\right]\sum_{j=1}^{N}\sum_{s=1}^{T}K_h(Z_{js}-z). \tag{2.11}$$

Suppose that $\tilde{\ell}_z^c(\delta)$ is maximised by $\hat{\delta} = (\hat{\phi}, \hat{\rho})^\top$, i.e. $\hat{\delta}$ is the quasi-maximum likelihood estimator of $\delta_0 = (\phi_0, \rho_0)^\top$. To establish its asymptotic properties requires a number of additional conditions. We introduce first some conditions of standard nature on the kernel function and the bandwidth.

Assumption B1. *$K(z)$ is a symmetric density function with a compact support. In addition, $K(z)$ has a first derivative $K'(z)$, by which $\int v(K'(z))^2 dv$ is bounded.*

Assumption B2. *The bandwidth parameter is any monotonic sequence $h = h(N) \propto (N)^{-1/5}$ implying that $\{Nh\}^{-1} = N^{-4/5}$.*

Now we introduce some conditions on the regressors and the covariate.

Assumption C1. *Z_{it} is independent and identically distributed (i.i.d.) over i and t. In addition, the density function $f(z)$ of Z_{it} is continuous, positively bounded away from 0 on [0,1], and has bounded second derivative.*

Assumption C2. *For $\forall z \in \mathcal{D}$, $i = 1, \ldots, N$ and $t = 1, \ldots, T$, $E[X_{it}|Z_{it} = z] = \mu_X(z)$ and $E[X_{it}^\top X_{it}|Z_{it} = z] = \Sigma_X(z)$ is non-singular and has bounded second order derivatives on $[0, 1]$. Also, $E[\|X_N^\top X_N\|_F^2|Z_{it} = z]$ is bounded, where $\|\cdot\|_F$ denotes the Frobenius norm. X_{it} is independent of Z_{js} for $(i, t) \neq (j, s)$.*

Finally, we impose some standard conditions on the functional coefficient.

Assumption D1. *Second order derivatives of $\beta_{0,d}(z)$, $d = 1, \ldots, D$, are continuous. Also, $E(\|\beta_{0,d}(z)\|^4)$ is bounded.*

Assumptions B1 and B2 are primitive and used regularly in nonparametric studies (see e.g. Okui and Takahide (2018)), whereas Assumption C1 ensures that the observed index values are sufficiently dense on the support. This implies that maximal distance between two consecutive index variables is only of the order $O_P\left(\frac{\log NT}{NT}\right)$. In addition, Assumption C2 replaces assumptions in spatial panel regression models. For example, it is required in those studies that elements of X_N are uniformly bounded constants for all N and $\lim_{NT\to\infty}(NT)^{-1}X_N^\top Q_N^{-1}X_N$ exists and is nonsingular for all $\delta \in \Delta$, (e.g. Assumption 6 of Lee (2004)). For an arbitrary index value $z \in [0, 1]$, let z^* be its nearest neighbor among the observed index values, i.e. $z^* = \arg\min_{\tilde{z}\in\{Z_{it}:1\leq i\leq N, 1\leq t\leq T\}} |z - \tilde{z}|$. Assumption D1 imposes a smoothness condition on the functional coefficient, which implies that $\|\beta_0(z) - \beta_0(z^*)\| = O_P\left(\frac{\log NT}{NT}\right)$ (see e.g. Xia et al. (2004)). This is an order substantially smaller than the optimal nonparametric convergence rate, which is $(NT)^{-2/5}$.

Furthermore, Lemma 2.1 below is useful for deriving consistency and identifiability of the spatial estimation, which are stated in Theorem 2.1.

Lemma 2.1. *Let Assumption A to D hold. Also, let $E\{\ell_z(\beta, \sigma_v^2, \phi, \rho)|z\} \equiv \bar{\ell}_z(\beta, \sigma_v^2, \phi, \rho)$ and $\bar{\ell}_z^c(\delta) \equiv \max_{\beta,\sigma_v^2} \bar{\ell}_z(\beta, \sigma_v^2, \phi, \rho)$. Then (a)*

$$\begin{aligned}
\bar{\ell}(\beta, \sigma_v^2, \phi, \rho) \;=\; & -\frac{1}{2}\log\{2\pi\sigma_v^2\}\sum_{j=1}^N\sum_{s=1}^T K_h(Z_{js} - z) + \log\{|\mathbb{Q}_N|\}\sum_{j=1}^N\sum_{s=1}^T K_h(Z_{js} - z) \\
& - \frac{1}{2\sigma_v^2}\{(\beta_0(z) - \beta(z))^\top E[\ddot{X}_N^\top K_N \ddot{X}_N|z](\beta_0(z) - \beta(z))\} \\
& - \frac{1}{2\sigma_v^2}\{\sigma_{0v}^2 tr[\mathbb{Q}_{0N}\bar{\mathbb{Q}}_N^\top K_N \bar{\mathbb{Q}}_N]\}, \qquad (2.12)
\end{aligned}$$

where $E[\ddot{X}_N^\top K_N \ddot{X}_N|z] = \sum_{j=1}^N\sum_{s=1}^T E[\ddot{X}_{js}^\top \ddot{X}_{js}|z]K_h(Z_{js} - z)$, and (b)

$$\bar{\ell}_z^c(\delta) = -\frac{1}{2}\left[\log(2\pi\bar{\sigma}_v^2) + \log|\mathbb{Q}_N| + 1\right]\sum_{j=1}^N\sum_{s=1}^T K_h(Z_{js} - z) \qquad (2.13)$$

where $\bar{\sigma}_v^2 = (1/NT)\sigma_{v_0}^2 TR[\mathbb{Q}_{0N}\bar{\mathbb{Q}}_N^\top K_N \bar{\mathbb{Q}}_N]\left[\sum_{j=1}^N\sum_{s=1}^T K_h(Z_{js} - z)\right]^{-1}$.

Theorem 2.1. *Let Assumption A to D hold. Also, let*

$$\limsup_{N\to\infty}\left\{\max_{\delta\in\bar{D}_\epsilon(\delta_0)\cap\Delta}\bar{\ell}_z^c(\delta)\right\}\neq\limsup_{N\to\infty}\bar{\ell}_z^c(\delta_0)$$

for any δ, where $\bar{D}_\epsilon(\delta_0)$ is the complement of ϵ-neighbourhood of δ_0. Then, δ_0 is uniquely identifiable and $\hat{\delta}=\delta_0+O_P\left((NT)^{-1/2}\right)$ as $N\to\infty$.

Moreover, the estimators of $\beta_0(z)$ and $\sigma_{v,0}^2$, which are based explicitly on $\hat{\delta}$, can be formulated as follows

$$\begin{aligned}\hat{\beta}(z) &= \left[\sum_{j=1}^N\sum_{s=1}^T\hat{X}_{js}^\top\hat{X}_{js}K_h(Z_{js}-z)\right]^{-1}\sum_{j=1}^N\sum_{s=1}^T\hat{X}_{js}^\top\hat{y}_{js}K_h(Z_{js}-z)\\ &= \left\{\hat{X}_N^\top K_N\hat{X}_N\right\}^{-1}\hat{X}_N^\top K_N\hat{y}_N,\end{aligned}\tag{2.14}$$

where $\hat{X}_N=\hat{\hat{\mathbb{Q}}}_N X_N$ in which $\hat{\hat{\mathbb{Q}}}=\{Q_{0,N}+(1+T\hat{\phi})^{-1/2}Q_{1,N}\}[I_T\otimes(I_N-\hat{\rho}W_N)]$, and

$$\hat{\sigma}_v^2(z)=\left[\sum_{j=1}^N\sum_{s=1}^T K_h(Z_{js}-z)\{\hat{y}_{js}-\hat{X}_{js}\hat{\beta}(z)\}^2\right]\left[\sum_{j=1}^N\sum_{s=1}^T K_h(Z_{js}-z)\right]^{-1}.\tag{2.15}$$

Since $\sigma_{v,0}^2$ does not depend on the location z, it can be estimated based on

$$\hat{\sigma}_v^2=\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T\hat{\sigma}_v^2(Z_{it}).$$

Hereafter, we refer to $\hat{\beta}(z)$ and $\hat{\sigma}_v^2$ as *"unpenalised estimators"* and present their asymptotic properties in Theorems 2.2 and 2.3. To state these properties, let $\mathscr{D}(z)=f_z(z)E(\ddot{X}_{0js}^\top\ddot{X}_{0js}|z)$, $\mathscr{B}=E[\ddot{X}_{0js}^\top\ddot{X}_{0js}|z](\beta_0'(z)f'(X_{js},Z_{js})/f(X_{js}|Z_{js}=z)+\frac{1}{2}\beta_0''(z)f_z(z))]$, and $V(z)=\sigma_{v,0}^2E(\ddot{X}_{0js}^\top\ddot{X}_{0js}|z)f(z)\mathscr{K}^2$, where $\mathscr{K}^2=\int K^2(u)du$ and $\mathscr{K}_2=\int u^2K(u)du$.

Theorem 2.2. *Let Assumption A to D hold. Then (a) $\hat{\sigma}_v^2=\sigma_{v,0}^2+O_P\left((NT)^{-1/2}\right)$, and (b) $\hat{\beta}(z)=\beta_0(z)+O_P\left((NT)^{-2/5}\right)$ as $N\to\infty$.*

Theorem 2.3. *Let Assumption A to D hold and $\inf_{||z||\le c_N}f_z(z)>0$, where $c_N=h^{-\delta}$ with $\delta>0$ being arbitrarily small. Then, as $N\to\infty$,*

$$\sqrt{NTh}\left(\hat{\beta}(z)-\beta_0(z)-Bias\right)\to_D N(0,\Sigma),$$

where $Bias=\mathscr{D}^{-1}(z)\mathscr{K}_2h^2\mathscr{B}$ and $\Sigma=\mathscr{D}^{-1}(z)V(z)\mathscr{D}^{-1}(z)$.

2.3. SAREC-KLASSO method

So far, we have assumed that all the regressors are necessary. In this section, we relax this assumption, and introduce a procedure for selecting relevant regressors. Particularly, we extend the KLASSO technique to the panel data context of (2.1), where there exists the Cliff-Ord-type models of spatial error dependence. We refer to this procedure as spatial autoregressive error component KLASSO or SAREC-KLASSO. To this end, we assume without loss of generality that there exists an integer D_0 such that $\infty > E\{\beta_{d,0}^2(Z_{it})\} > 0$ for any $d \leq D_0$, while $E\{\beta_{d,0}^2(Z_{it})\} = 0$ for any $D_0 < d$. Accordingly, define $X_{ita} = \{X_{it,1}, \ldots, X_{it,D_0}\} \in \mathbb{R}^{D_0}$ and $X_{itb} = \{X_{it,D_0+1}, \ldots, X_{it,D}\} \in \mathbb{R}^{D-D_0}$. In other words, there are D_0 regressors, which are truly relevant, but the rest are not.

Let $B = \{\beta(Z_{11}), \ldots, \beta(Z_{N1}), \beta(Z_{12}), \ldots, \beta(Z_{NT})\}^\top \equiv \{b_1, \ldots, b_{D_0},\ b_{D_0+1}, \ldots, b_D\}$, which is an $NT \times D$ matrix, and

$$\tilde{Q}_\lambda(B) = \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{N}\sum_{s=1}^{T}\{\ddot{y}_{js} - \ddot{X}_{js}\beta(Z_{it})\}^2 K_h(Z_{it} - Z_{js}) + \sum_{d=1}^{D}\lambda_d\|b_d\|, \tag{2.16}$$

where $\lambda = (\lambda_1, \ldots, \lambda_D)^\top \in \mathbb{R}^D$ are the tuning parameters, $b_d \in \mathbb{R}^{TN\times 1}$ is the dth column of B and $\|\cdot\|$ stands for the usual Euclidean norm. Under the conditions of the model, the last $(D - D_0)$ columns of the B matrix should be 0, which suggests that the task of variable selection is equivalent to identifying sparse columns in the B matrix. By following the group LASSO idea of Yuan and Lin (2006), we note firstly the penalized estimation

$$\begin{aligned}\tilde{B}_\lambda &= \{\tilde{\beta}_\lambda(Z_{11}), \ldots, \tilde{\beta}_\lambda(Z_{N1}), \tilde{\beta}_\lambda(Z_{12}), \ldots, \tilde{\beta}_\lambda(Z_{NT})\}^\top \\ &= \operatorname{argmin}_{B\in\mathbb{R}^{TN\times D}}\ \tilde{Q}_\lambda(B) \ \equiv\ (\tilde{b}_{\lambda,1}, \ldots, \tilde{b}_{\lambda,D_0}, \tilde{b}_{\lambda,D_0+1}, \ldots, \tilde{b}_{\lambda,D}).\end{aligned} \tag{2.17}$$

The above estimator can be viewed as the penalized counterpart of that in (2.10). In other words, the (i,t)-row of $\tilde{B}_\lambda$ is defined as the transpose of

$$\tilde{\beta}_\lambda(Z_{it}) = \left[\sum_{j=1}^{N}\sum_{s=1}^{T}\ddot{X}_{js}^\top\ddot{X}_{js}K_h(Z_{it} - Z_{js}) + \tilde{\mathfrak{D}}\right]^{-1}\sum_{j=1}^{N}\sum_{s=1}^{T}\ddot{X}_{js}{}^\top\ddot{y}_{js}K_h(Z_{it} - Z_{js}), \tag{2.18}$$

where $\tilde{\mathfrak{D}} = \operatorname{diag}(\lambda_1/\|\tilde{b}_1\|, \ldots, \lambda_D/\|\tilde{b}_D\|)$.

Similarly to the unpenalised procedure, $\tilde{\beta}_\lambda(z)$ is useful for formulating conditional local kernel-weighted log-likelihood

$$\begin{aligned}\tilde{\ell}_{\lambda,z}(\theta) &= -\frac{1}{2}\log\{2\pi\sigma_v^2\}\sum_{j=1}^{N}\sum_{s=1}^{T}K_h(Z_{js} - z) + \log\{|\mathbb{Q}_N|\}\sum_{j=1}^{N}\sum_{s=1}^{T}K_h(Z_{js} - z) \\ &\quad -\frac{1}{2\sigma_v^2}\sum_{j=1}^{N}\sum_{s=1}^{T}\{\ddot{y}_{js} - \ddot{X}_{js}\tilde{\beta}_\lambda(z)\}^2 K_h(Z_{js} - z),\end{aligned} \tag{2.19}$$

where $\theta = (\sigma_v^2, \phi, \rho)^\top \in \Theta$ for which Θ is a compact parameter space. Furthermore, the estimator

$$\tilde{\sigma}_{\lambda,v}^2 = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \tilde{a}_\lambda(Z_{it}),$$

where

$$\tilde{a}_\lambda(Z_{it}) = \left[\sum_{j=1}^{N} \sum_{s=1}^{T} K_h(Z_{js} - Z_{it}) \{ \ddot{y}_{js} - \ddot{X}_{js} \tilde{\beta}_\lambda(Z_{it}) \}^2 \right] \left[\sum_{j=1}^{N} \sum_{s=1}^{T} K_h(Z_{js} - z) \right]^{-1}$$

enables the formulation of the concentrated log-likelihood

$$\begin{aligned} \tilde{\ell}_\lambda^c(\delta) &= -\frac{1}{2} \left[\log\{2\pi\} + 1 + \log(\tilde{\sigma}_{\lambda,v}^2) \right] \sum_{j=1}^{N} \sum_{s=1}^{T} K_h(Z_{js} - z) \\ &\quad + \log |\mathbb{Q}_N| \sum_{j=1}^{N} \sum_{s=1}^{T} K_h(Z_{js} - z). \end{aligned} \tag{2.20}$$

Now, let $\hat{\delta}_\lambda$ denotes quasi-maximum likelihood estimates of δ_0. Then, the penalized estimate of $B_0 = \{\beta_0(Z_{11}), \ldots, \beta_0(Z_{N1}), \beta_0(Z_{12}), \ldots, \beta_0(Z_{NT})\}^\top$ is

$$\begin{aligned} \hat{B}_\lambda &= \{\hat{\beta}_\lambda(Z_{11}), \ldots, \hat{\beta}_\lambda(Z_{N1}), \hat{\beta}_\lambda(Z_{12}), \ldots, \hat{\beta}_\lambda(Z_{NT})\}^\top \\ &= \operatorname{argmin}_{B \in \mathbb{R}^{TN \times D}} \hat{Q}_\lambda(B) \equiv \left(\hat{b}_{\lambda,1}, \ldots, \hat{b}_{\lambda,D_0}, \hat{b}_{\lambda,D_0+1}, \ldots, \hat{b}_{\lambda,D} \right) \end{aligned} \tag{2.21}$$

in which

$$\hat{Q}_\lambda(B) = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{s=1}^{T} \left\{ \hat{y}_{js} - \hat{X}_{js} \beta(Z_{it}) \right\}^2 K_h(Z_{it} - Z_{js}) + \sum_{d=1}^{D} \lambda_d \|b_d\|, \tag{2.22}$$

where $\lambda = (\lambda_1, \ldots, \lambda_D)^\top \in \mathbb{R}^D$ are the tuning parameters, $b_d \in \mathbb{R}^{TN \times 1}$ is the dth column of B and $\|\cdot\|$ stands for the usual Euclidean norm. In other words, the (i,t)-row of $\hat{B}_\lambda$ is defined as the transpose of

$$\hat{\beta}_\lambda(Z_{it}) = \left[\sum_{j=1}^{N} \sum_{s=1}^{T} \hat{X}_{js}^\top \hat{X}_{js} K_h(Z_{it} - Z_{js}) + \hat{\mathcal{D}} \right]^{-1} \sum_{j=1}^{N} \sum_{s=1}^{T} \hat{X}_{js}^\top \hat{y}_{js} K_h(Z_{it} - Z_{js}), \tag{2.23}$$

where $\hat{\mathcal{D}} = \operatorname{diag}(\lambda_1/\|\hat{b}_1\|, \ldots, \lambda_D/\|\hat{b}_D\|)$.

To discuss the asymptotic properties of the penalized estimators, we need to impose some conditions on the amount of shrinkages being applied to the relevant and irrelevant coefficients as follows.

Assumption E1. *For* $a_N = \max\{\lambda_d : 1 \le d \le D_0\}$ *and* $b_N = \min\{\lambda_d : D_0 < d \le D\}$, *assume that* $(N)^{11/10} a_N \to 0$ *and* $(N)^{11/10} b_N \to \infty$.

We now present theoretical properties of the above penalized estimators. To this end, let $\hat{\beta}_\lambda(Z_{it}) = \{\hat{\beta}_{\lambda,a}(Z_{it}), \hat{\beta}_{\lambda,b}(Z_{it})\}^\top$, where $\hat{\beta}_{\lambda,a}(Z_{it}) = \{\hat{\beta}_{\lambda,1}(Z_{it}), \ldots, \hat{\beta}_{\lambda,D_0}(Z_{it})\}^\top$ and $\hat{\beta}_{\lambda,b}(Z_{it}) = \{\hat{\beta}_{\lambda,D_0+1}(Z_{11}), \ldots, \hat{\beta}_{\lambda,D}(Z_{11})\}^\top$.

Theorem 2.4. *Let Assumptions A to E hold. Then*

$$P\left(\sup_{z\in[0,1]} \|\hat{\beta}_{\lambda,b}(z)\| = 0\right) \to 1,$$

where $\hat{\beta}_{\lambda,b}(z) = (\hat{\beta}_{\lambda,(D_0+1)}(z), \ldots, \hat{\beta}_{\lambda,D}(z))^\top$.

Theorem 2.5. *Let Assumptions A to E hold. Then*

$$\sup_{z\in[0,1]} \|\hat{\beta}_{\lambda,a}(z) - \hat{\beta}_a(z)\| = o_P\{(NT)^{-2/5}\},$$

where

$$\hat{\beta}_a(z) = \left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \hat{X}_{ita}^\top \hat{X}_{ita} K_h(Z_{it} - z)\right]^{-1} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \hat{X}_{ita}^\top \hat{y}_{it} K_h(Z_{it} - z). \quad (2.24)$$

Theorem 2.4 suggests that the true model can be consistently selected as long as the tuning parameters satisfy the conditions listed in Assumption E1. Moreover, since it is associated with D_0, $\hat{\beta}_a(z)$ can be viewed as the oracle estimators. Theorem 2.5 suggests that the asymptotically optimal nonprametric convergence rate can be achieved as long as the tuning parameters satisfy the conditions listed in Assumption E1.

In spite of the results in Theorems 2.4 and 2.5, practical selection of up to D shrinkage parameters, i.e. $\lambda_1, \ldots, \lambda_D$, is not straightforward. In order to overcome such a difficulty, we follow an idea often used in the literature (see e.g. Zou (2006), Wang and Leng (2007) and Zou and Li (2007)) that is to specify

$$\lambda_d = \frac{\lambda_0}{(NT)^{-1/2}\|\hat{b}_d\|}, \quad (2.25)$$

where $\hat{b}_d$ is the d-th column of the unpenalised estimate $\hat{B}$ and $\lambda_0 > 0$. In this regard, it is important to note that

$$(NT)^{-1/2}\|\hat{b}_d\| = \left\{(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} \hat{\beta}_k^2(Z_{it})\right\}^{1/2} \to_P \left\{E[\beta^2(Z_{it})]\right\}^{1/2}, \ 1 \le d \le D_0 \quad (2.26)$$

and

$$\left\{(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} \hat{\beta}_k^2(Z_{it})\right\}^{1/2} = O_P\{(NT)^{-2/5}\}, \ (D_0+1) \le d \le D, \quad (2.27)$$

which are direct results of Lemma C.5 in the appendix. While (2.26) suggests that λ_d converges to a positive constant for $1 \leq d \leq D_0$, (2.27) implies λ_d converges to infinity for $(D_0+1) \leq d \leq D$. Therefore, in order to maintain $(N)^{11/10}a_N \to 0$ and $(N)^{11/10}b_N \to \infty$, it must be the case that $\lambda_0(NT)^{11/10} \to 0$ and $\lambda_0(NT)^{3/2} \to \infty$.

The specification in (2.25) helps to reduce the original D-dimensional problem about $\lambda \in \mathbb{R}^D$ to a univariate problem about selecting $\lambda_0 > 0$. In practice, such a selection is done by minimising the following BIC-type criterion

$$\mathrm{BIC}_\lambda = \log\{RSS_\lambda\} + df \times \frac{\log\{(NT)h\}}{(NT)h}, \tag{2.28}$$

where $0 \leq df \leq D$ is the number of nonzero coefficients identified by $\hat{B}_\lambda$ and

$$RSS_\lambda = (NT)^{-2} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{j=1}^{N} \sum_{s=1}^{T} \left\{ \hat{y}_{js} - \hat{X}_{js}\hat{\beta}_\lambda(Z_{it}) \right\}^2 K_h(Z_{it} - Z_{js}). \tag{2.29}$$

Let $\hat{B}_{\hat{\lambda}}$ denote a penalised estimator in (2.21), which corresponds to $\hat{\lambda} = \mathrm{argmin}_\lambda \, \mathrm{BIC}_\lambda$, and $\mathcal{S}_{\hat{\lambda}}$ represent the model identified by $\hat{B}_{\hat{\lambda}}$. Theorem 2.6 below states that the turning parameter $\hat{\lambda}$ selected by the BIC criterion is able to consistently identify the true model.

Theorem 2.6. *Let Assumptions A to E hold. Then*

$$P(\mathcal{S}_{\hat{\lambda}} = \mathcal{S}_T) \to 1 \tag{2.30}$$

as $N \to \infty$, where $\mathcal{S}_T = \{1, \ldots, D_0\}$ denotes the true model.

Remark 2.1. *A final point to clarify regarding the variable-selection procedure is the use of $\hat{y}_{it}$ and $\hat{X}_{it}$ in the calculation of RSS_λ in (2.29). In this regard, the conceptual discussion in Section suggests that we can rely on the following steps: (i) compute spatial estimates of $\delta = (\phi, \rho)^\top$ based on maximizing the concentrated log-likelihood under the unpenalized estimation in (2.10), (ii) compute $\hat{y}_{it}$ and $\hat{X}_{it}$, then (iii) apply the SAREC-KLASSO method as discussed in the previous and current sections.*

2.4. Identifying constant coefficients in semi-varying coefficient models

Another useful procedure is to identify constant coefficients amongst those associated with the relevant regressors selected in the previous section. This enables modelling of the so-called semi-varying coefficient specification. In this section, we suggest an approach for identifying constant coefficients, which can be viewed as an alternative to the shrinkage method introduced in Hu and Xia (2012). Our approach consists of two important steps. In the first step, we select the relevant variables using the shrinkage method introduced in Section 2.3. Theorems 2.4 to 2.6 ensures that all relevant variables that are associated with nonzero (functional or constant) coefficients are consistently identified as long as

the tuning parameters satisfy the conditions listed in Assumption E1. Let $\hat{D}$ denote the number of relevant regressors identified by $\hat{B}_{\hat{\lambda}}$. The second step involves hypothesis testing for coefficient constancy in the varying-coefficient model. More specifically, we test the hypotheses

$$H_0 : \beta_d(z) = C_d \quad \text{versus} \quad H_1 : \beta_d(z) \neq C_d, \quad 1 \leq d \leq \hat{D}, \tag{2.31}$$

where C_d is a constant.

Corollary 2.1. *Let the conditions of Theorem 2.2(a) hold. Then*

$$\sup_{z\in[0,1]} \|\hat{\beta}(z) - \tilde{\beta}(z)\|^2 = o_P\{(NT)^{-2/5}\}. \tag{2.32}$$

Corollary 2.1 suggests that the difference between $\hat{\beta}(z)$ and $\tilde{\beta}(z)$ (defined in (2.10)) is negligible uniformly over the entire index support. This is critical since it ensures that above hypothesis test can be implemented based on asymptotic property established in Cai et al. (2000), Fan and Zhang (2000b). The test statistic is written as

$$T_j = (-2\log h)^{1/2} \left[\sup_{z\in[0,1]} \left| \{\widehat{var}(\hat{\beta}_j|\mathfrak{D})\}^{-1/2} (\hat{\beta}_j(z) - \hat{C}_j - \widehat{bias}(\hat{b}_j(z)|\mathfrak{D})) \right| - d_N \right], \tag{2.33}$$

in which the components of the test can be defined as follows:

$$\widehat{var}(\hat{\beta}_j(z)|\mathfrak{D}) = e_{j,p}^\top \left\{ \hat{X}_N^\top K_N \hat{X}_N \right\}^{-1} \hat{X}_N^\top K_N^2 \hat{X}_N \left\{ \hat{X}_N^\top K_N \hat{X}_N \right\}^{-1} e_{j,p} \hat{\sigma}_v^2,$$
$$d_N = (-2\log h)^{1/2} + \tfrac{1}{(-2\log h)^{1/2}} \log\left\{ \tfrac{1}{4\nu_0 \pi} \textstyle\int \{K'(t)\}^2 dt \right\},$$
$$\widehat{bias}(\hat{\beta}_j(z)|\mathfrak{D}) \approx e_{j,p}^\top \left\{ \hat{X}_N^\top K_N \hat{X}_N \right\}^{-1} \hat{X}_N^\top K_N \hat{a}_N,$$
$$\hat{C}_j = \tfrac{1}{NT} \textstyle\sum_{i=1}^T \sum_{t=1}^T \hat{\beta}_j(Z_{it}),$$

where $a_{it} = \left\{ \hat{\beta}^{(1)}(z)(Z_{it} - z) + 2^{-1}\hat{\beta}^{(2)}(z)(Z_{it} - z)^2 \right\} \hat{X}_{it}$ and $K'(t) = \partial K(t)/\partial t$. Finally, the null hypothesis is rejected when the test statistic exceeds the asymptotic critical value $c_\alpha = -\log(-0.5\log\alpha)$.

2.5. Local quadratic approximation of the penalty function

In the spirit of Hunter and Li (2005) (see also Fan and Li (2001)), the computation in practice is based on an iterative algorithm in which the loss function in (2.22) is locally approximated by

$$\sum_{i=1}^N \sum_{t=1}^T \sum_{j=1}^N \sum_{s=1}^T \left\{ \hat{y}_{js} - \hat{X}_{js}^\top \beta(Z_{it}) \right\}^2 K_h(Z_{it} - Z_{js}) + \sum_{d=1}^D \lambda_d \frac{\|b_d\|^2}{\|\hat{b}_{\lambda,d}^{(m)}\|} \tag{2.34}$$
$$= \sum_{i=1}^N \sum_{t=1}^T \left\{ \sum_{j=1}^N \sum_{s=1}^T \left\{ \hat{y}_{js} - \hat{X}_{js}^\top \beta(Z_{it}) \right\}^2 K_h(Z_{it} - Z_{js}) + \sum_{d=1}^D \lambda_d \frac{\beta_d^2(Z_{it})}{\|\hat{b}_{\lambda,d}^{(m)}\|} \right\},$$

where $\hat{B}_\lambda^{(m)} = \left\{\hat{\beta}_\lambda^{(m)}(Z_{11}), \hat{\beta}_\lambda^{(m)}(Z_{21}), \ldots, \hat{\beta}_\lambda^{(m)}(Z_{NT})\right\}^\top = \left(\hat{b}_{\lambda,1}^{(m)}, \hat{b}_{\lambda,2}^{(m)}, \ldots, \hat{b}_{\lambda,D}^{(m)}\right)$ denotes the estimates obtained in the mth iteration. The minimiser of which is $\hat{B}_\lambda^{(m+1)}$ such that the (i,t)-row is defined as the transpose of

$$\begin{aligned}\hat{\beta}_\lambda^{(m+1)}(Z_{it}) &= \left\{\sum_{j=1}^{N}\sum_{s=1}^{T}\hat{X}_{js}^\top\hat{X}_{js}K_h(Z_{it}-Z_{js}) + \hat{\mathcal{D}}^{(m)}\right\}^{-1} \\ &\quad\times\left\{\sum_{j=1}^{N}\sum_{s=1}^{T}\hat{X}_{js}^\top\hat{y}_{js}K_h(Z_{it}-Z_{js})\right\} \equiv \hat{\beta}_{it}^{(m+1)}, \qquad (2.35)\end{aligned}$$

where $\hat{\mathcal{D}}^{(m)} = \mathrm{diag}(\lambda_1/\|\hat{b}_{\lambda,1}^{(m)}\|, \ldots, \lambda_D/\|\hat{b}_{\lambda,D}^{(m)}\|)$.

We next study the dynamics of $\hat{\beta}_\lambda^{(m+1)}(z)$ as $m \to \infty$. The results are presented as corollaries of Theorems 2.4 and 2.5 since their mathematical proof is closely related.

Corollary 2.2. *Let Assumptions A to E hold. Then*

$$P\left(\sup_{z\in[0,1]}\|\hat{\beta}_{\lambda,b}^{(m+1)}(z)\| = 0\right) \to 1,$$

where $\hat{\beta}_{\lambda,b}^{(m+1)}(z) = (\hat{\beta}_{\lambda,(D_0+1)}^{(m+1)}(z), \ldots, \hat{\beta}_{\lambda,D}^{(m+1)}(z))^\top$.

Corollary 2.3. *Let Assumptions A to E hold. Then, we have*

$$\sup_{z\in[0,1]}\|\hat{\beta}_{\lambda,a}^{(m+1)}(z) - \hat{\beta}_a(z)\| = o_P\{(NT)^{-2/5}\},$$

where $\hat{\beta}_{\lambda,a}^{(m+1)}(z) = (\hat{\beta}_{\lambda,1}^{(m+1)}(z), \ldots, \hat{\beta}_{\lambda,D_0}^{(m+1)}(z))^\top$.

3. Simulation studies

In this section, we present a set of simulation exercises that examine the finite-sample performance of the procedures introduced in the previous sections. These are (3.1) spatial estimation, which involves estimation of $\delta_0 = (\phi_0, \rho_0)^\top$, using the concentrated likelihood; (3.2) nonparametric estimation of coefficient functions $\beta_0(z) = \{\beta_{0,1}(z), \ldots, \beta_{0,D}(z)\}^\top$ based on the oracle, unpenalised and the penalised estimators; (3.3) variable selection, i.e. relevant versus irrelevant variables, based on the SAREC-KLASSO and KLASSO methods; and (3.4) hypothesis testing of coefficient constancy, i.e $H_0 : \beta_{0,d}(z) = C_d$ versus $H_1 : \beta_{0,d}(z) \neq C_d$, where $1 \le d \le \hat{D}$ and C_d is a constant.

In order to achieve these, we assume that y_{it} is generated based on two types of data generating process, namely:

$$\begin{aligned}\text{Model I} \quad y_{it} &= 2\sin(2\pi Z_{it})X_{it,1} + 2\cos(2\pi Z_{it})X_{it,2} + u_{it} \\ \text{Model II} \quad y_{it} &= 2\sin(2\pi Z_{it})X_{it,1} + 2\cos(2\pi Z_{it})X_{it,2} + 0.5X_{it,3} + 0.7X_{it,4} + u_{it}\end{aligned}$$

Table 1: Spatial estimation: Model I

$P=2$	$N=100$			$N=200$			$N=300$		
	$\hat{\rho}_{or}$	$\hat{\rho}$	$\hat{\rho}_\lambda$	$\hat{\rho}_{or}$	$\hat{\rho}$	$\hat{\rho}_\lambda$	$\hat{\rho}_{or}$	$\hat{\rho}$	$\hat{\rho}_\lambda$
MAE	0.063	0.068	0.075	0.044	0.045	0.046	0.037	0.039	0.038
$RMSE$	0.078	0.086	0.090	0.055	0.056	0.059	0.047	0.048	0.048
	$\hat{\phi}_{or}$	$\hat{\phi}$	$\hat{\phi}_\lambda$	$\hat{\phi}_{or}$	$\hat{\phi}$	$\hat{\phi}_\lambda$	$\hat{\phi}_{or}$	$\hat{\phi}$	$\hat{\phi}_\lambda$
MAE	0.221	0.253	0.245	0.172	0.203	0.189	0.113	0.159	0.141
$RMSE$	0.269	0.303	0.291	0.207	0.241	0.228	0.146	0.195	0.178
$P=5$	$N=100$			$N=200$			$N=300$		
	$\hat{\rho}_{or}$	$\hat{\rho}$	$\hat{\rho}_\lambda$	$\hat{\rho}_{or}$	$\hat{\rho}$	$\hat{\rho}_\lambda$	$\hat{\rho}_{or}$	$\hat{\rho}$	$\hat{\rho}_\lambda$
MAE	0.090	0.081	0.099	0.064	0.066	0.073	0.054	0.055	0.059
$RMSE$	0.109	0.103	0.128	0.081	0.083	0.092	0.068	0.072	0.077
	$\hat{\phi}_{or}$	$\hat{\phi}$	$\hat{\phi}_\lambda$	$\hat{\phi}_{or}$	$\hat{\phi}$	$\hat{\phi}_\lambda$	$\hat{\phi}_{or}$	$\hat{\phi}$	$\hat{\phi}_\lambda$
MAE	0.214	0.252	0.241	0.173	0.203	0.190	0.115	0.162	0.143
$RMSE$	0.265	0.300	0.288	0.208	0.241	0.230	0.148	0.197	0.181
$P=8$	$N=100$			$N=200$			$N=300$		
	$\hat{\rho}_{or}$	$\hat{\rho}$	$\hat{\rho}_\lambda$	$\hat{\rho}_{or}$	$\hat{\rho}$	$\hat{\rho}_\lambda$	$\hat{\rho}_{or}$	$\hat{\rho}$	$\hat{\rho}_\lambda$
MAE	0.096	0.092	0.123	0.078	0.078	0.096	0.063	0.065	0.074
$RMSE$	0.120	0.116	0.151	0.095	0.099	0.119	0.082	0.084	0.098
	$\hat{\phi}_{or}$	$\hat{\phi}$	$\hat{\phi}_\lambda$	$\hat{\phi}_{or}$	$\hat{\phi}$	$\hat{\phi}_\lambda$	$\hat{\phi}_{or}$	$\hat{\phi}$	$\hat{\phi}_\lambda$
MAE	0.213	0.254	0.238	0.173	0.204	0.190	0.115	0.162	0.144
$RMSE$	0.265	0.303	0.287	0.209	0.243	0.231	0.148	0.198	0.182

Note: Estimates computed based on maximizing the concentrated log-likelihood (i) under the unpenalised estimation, $\hat{\rho}$, (ii) under the penalized estimation, $\hat{\rho}_\lambda$, and (iii) under the oracle estimation, $\hat{\rho}_{or}$.

The difference between Models I and II lies in the fact that the former includes zero constant-coefficients, whereas the latter includes two, i.e. 0.5 and 0.7. Hence, Model II is an example of the semi-varying coefficient models. We set $X_{it,1} = 1$, and generate $(X_{it,2}, \ldots, X_{it,7})^\top$ from multivariate normal distribution by setting $cov(X_{it,j_1}, X_{it,j_2}) = 0.5^{|j_1-j_2|}$ for any $2 \leq j_1, j_2 \leq 7$. We also generate Z_{it} from uniform distribution $U[0,1]$. Furthermore, the disturbance follows the SAR and EC processes explained in Section 2.1 with $\rho_0 = 0.3$ and $\sigma^2_{\mu,0} = \sigma^2_{v,0} = 1$. Regarding the required spatial weight matrix, we follow Kelejian and Prucha (1999) and employ matrices that differ in their degree of sparseness. Particularly, we construct what known in the literature as the "P-ahead-and-P-behind" spatial association. For example, $P = 1$ leads to the "1-ahead-and-1-behind" matrix, whose ith row has nonzero elements in positions $i+1$ and $i-1$, so that the ith element is directly related only to two other elements, namely the ones in front and behind it. We construct three spatial weight matrices based on $P = 2$, $P = 5$ or $P = 8$, which lead to 4, 10 and 16 nonzero elements in a given row, respectively.

Moreover, in the practical computation, we follow the results of Magnus and Muris (2010) and compute inverse and determinant of the matrix $\mathbb{Q}_N$ based on

$$(\mathbb{Q}_N)^{-1} = (1/T)J_T \otimes C_1^{-1} + \{I_T - (1/T)J_T\} \otimes C_2^{-1} \quad \text{and} \quad |\mathbb{Q}_N| = |C_1||C_2|^{T-1},$$

where $C_1 = (1+\phi T)C_2$ and $C_2 = \{(I_N - \rho W_N)^\top (I_N - \rho W_N)\}^{-1}$. This can help to alleviate a serious computational burden caused by repeated evaluations of this $TN \times TN$ matrix during the optimisation process. Other necessary computational parameters are selected as follows. For each simulation repetition, we select the optimal bandwidth based on the method of leaving-one-out cross validation within the context of the unpenalised estimation since the asymptotic theory for such selection is already well developed in the literature (see e.g. Lee and Yu (2010)). The bandwidth selected in this step is also used in the penalized estimation. In addition, the optimal shrinkage parameter is selected based on the BIC criterion defined in (2.6), whereas the total number of iteration of the iterative algorithm in Section 2.5 is set at 15. In the simulation exercises that follow, a total of 200 repetitions are conducted for each of the model setups. Tables 1 to 3 summarise the simulation-results obtained. Below, we discuss a number of important findings.

3.1. Autoregressive parameter and variance ratio

In Tables 1 and 2, $\hat{\rho}_{or}$, $\hat{\rho}_{un}$ and $\hat{\rho}_{\lambda}$ denote estimates of the spatial parameter ρ_0 that are computed based on maximizing the concentrated log-likelihood under the oracle, unpenalised and penalised estimation, respectively. Similarly, $\hat{\phi}_{or}$, $\hat{\phi}_{un}$ and $\hat{\phi}_{\lambda}$, are those of the variance ratio $\phi_0 = \sigma^2_{\mu,0}/\sigma^2_{v,0}$. For comparison, we consider two measures of accuracy, namely the mean absolute error (MAE) and root mean squared error (RMSE). While the RMSE closely resembles a standard definition that is often seen in the literature, it is based instead on quantiles, which exist with certainty, rather than moments (see also Kapoor et al. (2007)). In particular, we compute

$$RMSE = \left\{ bias^2 + \left(\frac{IQ}{1.35} \right)^2 \right\}^{1/2}, \tag{3.1}$$

where *bias* refers to the difference between the median of the estimates and ρ_0, IQ is the inter-quantile range $c_1 - c_2$ in which c_1 and c_2 are the 0.75 and 0.25 quantiles, respectively.

The results in the tables show that $\hat{\rho}_{or}$ and $\hat{\rho}_{un}$ perform almost equally well when N is small. Although MAE and RMSE for $\hat{\rho}_{\lambda}$ converge to zero as N increases, the estimator does not perform as well as the oracle and the unpenalised-based counterpart at small N. However, all the three estimators of the spatial parameter perform almost equally well at larger N. Regarding those of the variance ratio, it is clear that $\hat{\phi}_{or}$ performs the best. Unlike that of the spatial parameter, here $\hat{\phi}_{\lambda}$ performs much better than its unpenalised-based counterpart. These results are not surprising given the fact that the oracle and

Table 2: Spatial estimation: Model II

$P=2$	$N=100$			$N=200$			$N=300$		
	$\hat{\rho}_{or}$	$\hat{\rho}$	$\hat{\rho}_\lambda$	$\hat{\rho}_{or}$	$\hat{\rho}$	$\hat{\rho}_\lambda$	$\hat{\rho}_{or}$	$\hat{\rho}$	$\hat{\rho}_\lambda$
MAE	0.068	0.068	0.079	0.044	0.046	0.048	0.038	0.039	0.039
$RMSE$	0.083	0.086	0.096	0.056	0.056	0.061	0.047	0.048	0.050
	$\hat{\phi}_{or}$	$\hat{\phi}$	$\hat{\phi}_\lambda$	$\hat{\phi}_{or}$	$\hat{\phi}$	$\hat{\phi}_\lambda$	$\hat{\phi}_{or}$	$\hat{\phi}$	$\hat{\phi}_\lambda$
MAE	0.236	0.254	0.257	0.170	0.203	0.191	0.123	0.159	0.155
$RMSE$	0.283	0.303	0.304	0.206	0.241	0.229	0.157	0.195	0.190
$P=5$	$N=100$			$N=200$			$N=300$		
	$\hat{\rho}_{or}$	$\hat{\rho}$	$\hat{\rho}_\lambda$	$\hat{\rho}_{or}$	$\hat{\rho}$	$\hat{\rho}_\lambda$	$\hat{\rho}_{or}$	$\hat{\rho}$	$\hat{\rho}_\lambda$
MAE	0.095	0.081	0.102	0.066	0.066	0.075	0.054	0.055	0.061
$RMSE$	0.119	0.103	0.129	0.084	0.083	0.098	0.070	0.072	0.080
	$\hat{\phi}_{or}$	$\hat{\phi}$	$\hat{\phi}_\lambda$	$\hat{\phi}_{or}$	$\hat{\phi}$	$\hat{\phi}_\lambda$	$\hat{\phi}_{or}$	$\hat{\phi}$	$\hat{\phi}_\lambda$
MAE	0.232	0.252	0.252	0.171	0.203	0.194	0.125	0.162	0.157
$RMSE$	0.279	0.300	0.301	0.207	0.241	0.232	0.159	0.198	0.193
$P=8$	$N=100$			$N=200$			$N=300$		
	$\hat{\rho}_{or}$	$\hat{\rho}$	$\hat{\rho}_\lambda$	$\hat{\rho}_{or}$	$\hat{\rho}$	$\hat{\rho}_\lambda$	$\hat{\rho}_{or}$	$\hat{\rho}$	$\hat{\rho}_\lambda$
MAE	0.095	0.092	0.121	0.080	0.078	0.096	0.063	0.065	0.075
$RMSE$	0.119	0.116	0.148	0.098	0.099	0.121	0.082	0.084	0.099
	$\hat{\phi}_{or}$	$\hat{\phi}$	$\hat{\phi}_\lambda$	$\hat{\phi}_{or}$	$\hat{\phi}$	$\hat{\phi}_\lambda$	$\hat{\phi}_{or}$	$\hat{\phi}$	$\hat{\phi}_\lambda$
MAE	0.232	0.254	0.253	0.170	0.204	0.194	0.125	0.162	0.158
$RMSE$	0.279	0.303	0.301	0.207	0.243	0.232	0.159	0.198	0.193

Note: $\hat{\rho}$, $\hat{\rho}_\lambda$, and $\hat{\rho}_{or}$ are defined as in Table 1.

the penalised estimation are able to provide the much more accurate estimates of the coefficient functions (we will discuss this further below). At $N = 300$, $\hat{\phi}_\lambda$ performs almost as well as the oracle-based counterpart. Moreover, an increase in P, which leads to a higher number of nonzero elements in a give row of the weighting matrix, renders less accurate estimation of both the spatial parameter and the variance ratio. However, that of the former seems to be affected more significantly. Finally, similar results are obtained for both of the model examples.

3.2. Nonparametric estimation of the coefficient functions

In order to investigate the relative accuracy of the penalized estimators compared to that of the unpenalised and oracle based counterparts, we compute the following relative estimation error (REE)

$$REE = 100 \times \frac{\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{t=1}^{T}|\hat{\beta}_{\lambda,k}(Z_{it}) - \beta_{0,k}(Z_{it})|}{\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{t=1}^{T}|\vartheta_k(Z_{it}) - \beta_{0,k}(Z_{it})|}, \tag{3.2}$$

where $\vartheta_k(Z_{it})$ may be either the unpenalised $\hat{\beta}_k(Z_{it})$ or oracle estimator $\hat{\beta}_{or,k}(Z_{it})$.

Table 3: Nonparametric estimation of the coefficient functions

Model I	$N = 100$		$N = 200$		$N = 300$	
	REE_{or}	REE_{un}	REE_{or}	REE_{un}	REE_{or}	REE_{un}
$P = 2$	1.278	0.391	1.079	0.357	1.032	0.350
$P = 5$	1.289	0.393	1.075	0.354	1.032	0.350
$P = 8$	1.291	0.394	1.074	0.354	1.030	0.349
Model II	N = 100		N = 200		N = 300	
	REE_{or}	REE_{un}	REE_{or}	REE_{un}	REE_{or}	REE_{un}
$P = 2$	1.048	0.608	1.011	0.604	0.998	0.599
$P = 5$	1.057	0.611	1.014	0.605	0.998	0.601
$P = 8$	1.057	0.609	1.013	0.604	0.997	0.600

Table 3 presents the related simulation results. In the table, REE_{or} and REE_{un} represent the REE measures when $\vartheta_k(Z_{it})$ is $\hat{\beta}_{or,k}$ and $\hat{\beta}_k(Z_{it})$, respectively. In all cases, it is clear that REE_{or} converges to one, while REE_{un} converges away from one as N increases. This implies the penalised based estimator performs at least as well as the oracle estimator as $N \to \infty$, but definitely better than the unpenalised counterpart. Moreover, the penalised based estimator perform well asymptotically for the models that involve zero coefficients. However, it performs even better asymptotically for the model that involves a mixture of functional and constant coefficients. In fact, the penalised estimator is already performing as well as the oracle counterpart at N as low as 300. Finally, these results are quite robust across P.

3.3. Variable selection

We now discuss finite sample performance of the SAREC-KLASSO procedure for selecting between relevant and irrelevant regressors. Table 4 summarises the simulation results. Prior to considering these results, it is useful to note that the vector of relevant regressors is $X_{ita}^{\top} = \{X_{it,1}, X_{it,2}\}^{\top}$ for Model I, whereas it is $X_{ita}^{\top} = \{X_{it,1}, \ldots, X_{it,4}\}^{\top}$ for Model II, so that the numbers of relevant regressors are $K_0 = 2$ and $K_0 = 4$, respectively. Table 4 presents percentages of the simulation repetitions where the SAREC-KLASSO procedure is not only able to obtain the correct number of relevant regressors, but also able to accurately select the regressors in questions.

These results show that the performance of our procedure is not affected by the fact that Model II contains constant coefficients. Such a finding paves way for identifying constant coefficients in semivarying coefficient models using the procedure introduced in Section 2.4. A higher number of nonzero coefficients leads to better finite sample performance at smaller N. Nonetheless, the results for the two models converge when N increases to 300. Moreover, the finite sample performance of our selection procedure seems to be

Table 4: Variable selection

Model	$P = 2$	$N = 100$	$N = 200$	$N = 300$
I	KLASSO	0.020	0.167	0.533
	SAREC-KLASSO	0.353	0.800	0.960
II	KLASSO	0.007	0.053	0.340
	SAREC-KLASSO	0.627	0.880	0.967
Model	$P = 5$	$N = 100$	$N = 200$	$N = 300$
I	KLASSO	0.027	0.200	0.560
	SAREC-KLASSO	0.360	0.820	0.960
II	KLASSO	0.007	0.067	0.353
	SAREC-KLASSO	0.613	0.860	0.960
Model	$P = 8$	$N = 100$	$N = 200$	$N = 300$
I	KLASSO	0.027	0.213	0.580
	SAREC-KLASSO	0.367	0.867	0.967
II	KLASSO	0.007	0.087	0.387
	SAREC-KLASSO	0.593	0.833	0.967

worsen as P increases, but only marginally. This likely reflects the performance of the spatial estimation, which was discussed in the previous section. Finally, it is important to note that the KLASSO procedure is totally incapable of operating under models associated with spatially correlated error components.

3.4. Hypothesis testing for coefficient constancy

In the current section, we examine finite sample performance of the Fan and Zhang's (2000) hypothesis testing procedure of coefficient constancy for models associated with spatially correlated error components. We compare two scenarios, namely Fan and Zhang's (2000) procedure *with* and *without* spatial error dependence being addressed and the random effect being utilised in order to obtain efficiency gain. Furthermore, in order to allow an investigation into the ability of the test to reject an untrue null hypothesis we assume that observation y_{it} is generated based on:

$$\text{Model III} \quad y_{it} = 2\sin(2\pi Z_{it})X_{it,1} + 0.5\cos(2\pi Z_{it})X_{it,2} + 0.5Z_{it}(1 - Z_{it})X_{it,3} + u_{it}$$

Otherwise, the remaining details are as previously specified.

Table 5 summarises the simulation results. The table shows percentages of correct rejections and non-rejections (out of 150 replications) obtained by applying the Fan and Zhang's (2000) testing procedure *with* and *without* spatial error dependence being addressed and the random effect being utilised in order to obtain efficiency gain. Before discussing our findings, it is important to note that, in Model III, $\beta_{0,1}(z)$ demonstrates a much strong nonlinearity compared to $\beta_{0,1}(z)$ and $\beta_{0,1}(z)$. The results obtained seem to

Table 5: Hypothesis test of coefficient constancy

$P = 2$	$N = 100$		$N = 200$		$N = 300$	
Null Hypothesis	*with*	*without*	*with*	*without*	*with*	*without*
$H_0 : \beta_1(z) = C_1$	100	100	100	100	100	100
$H_0 : \beta_2(z) = C_2$	65	60	84	75	94	84
$H_0 : \beta_3(z) = C_3$	67	69	81	77	85	77
$H_0 : \beta_4(z) = 0$	74	69	85	83	89	84
$H_0 : \beta_5(z) = 0$	76	64	84	77	83	81
$H_0 : \beta_6(z) = 0$	72	70	83	74	84	77
$H_0 : \beta_7(z) = 0$	81	66	82	77	88	83
$P = 5$	$N = 100$		$N = 200$		$N = 300$	
Null Hypothesis	*with*	*without*	*with*	*without*	*with*	*without*
$H_0 : \beta_1(z) = C_1$	100	100	100	100	100	100
$H_0 : \beta_1(z) = C_2$	67	63	81	79	94	85
$H_0 : \beta_1(z) = C_3$	69	65	77	76	80	76
$H_0 : \beta_1(z) = 0$	74	67	85	79	90	86
$H_0 : \beta_1(z) = 0$	73	66	85	77	83	79
$H_0 : \beta_1(z) = 0$	73	72	82	74	83	79
$H_0 : \beta_1(z) = 0$	81	67	83	78	88	84
$P = 8$	$N = 100$		$N = 200$		$N = 300$	
$H_0 : \beta_1(z) = C_1$	100	100	100	100	100	100
$H_0 : \beta_1(z) = C_2$	65	63	81	81	94	85
$H_0 : \beta_1(z) = C_3$	66	65	77	75	80	75
$H_0 : \beta_1(z) = 0$	75	70	83	79	90	85
$H_0 : \beta_1(z) = 0$	73	66	85	77	82	77
$H_0 : \beta_1(z) = 0$	74	72	83	71	83	77
$H_0 : \beta_1(z) = 0$	81	68	83	75	88	85

Note: The table shows percentages of correct rejections and non-rejections obtained by applying the Fan and Zhang's (2000) testing procedure *with* and *without* spatial error dependence being addressed and the random effect being utilised in order to obtain efficiency gain.

reflect this fact. The null hypothesis of a constant coefficient is easily rejected for $\beta_{0,1}(z)$ such that the percentages of rejections reach 100% even for $N = 100$. For $\beta_{0,1}(z)$ and $\beta_{0,1}(z)$, having addressed spatial error dependence and utilised random effect in order to obtain efficiency gain clearly makes a significant impact on the power of the test. A similar benefit is also evidence for $\beta_{0,4}(z)$ to $\beta_{0,7}(z)$. In this regard, the correct null hypothesis is rejected much less frequently. Finally, the results are robust across P.

4. Public mental health expenditure in England

In the UK, Department for Communities and Local Government's (DCLG) revenue account budget records Mental Health Support (MHS), which covers services where the

Table 6: Our data and its sources

Symbols	Descriptions
vote	Percentage of voters with right-wing ideology Source: Percentage of voters that have voted for the Conservative and UK Independence Parties in local government elections available at www.electionscentre.co.uk
tph	Population-standardised total public health by local authority Source: Reported in the DCLG's Revenue Outturn, Social Care and Public Health data available at www.ons.gov.uk
mhs	Per capita measure of standardised MHS for persons age between 18 and 64 Source: Reported in the DCLG's Revenue Outturn, Social Care and Public Health data available at www.ons.gov.uk
nuc	Claimants of unemployment-related benefits on Benefits Agency Administrative System Source: Regional labour market Claimant Count by unitary and local authority available at www.ons.gov.uk
pmp	Percentage of male population by local authority Source: Estimates of the population for the UK available at www.ons.gov.uk
*pu*14	Percentage of population under 14 year of age Source: Estimates of the population for the UK available at www.ons.gov.uk
smr	Age-standardised mortality rates for 2016 to 2019 standardised to the 2013 European Standard Population expressed per 100,000 population Source: Deaths registered by area of usual residence available at https://data.gov.uk
noj	Number of jobs is measured by the Labour Force Survey as the sum of employee jobs; self-employment jobs, and government-supported trainees Source: Regional labour market available at https://data.gov.uk
plp	Percentage of households headed by lone parent by local authority Source: Estimated number of households by household types, local authorities in England available at www.ons.gov.uk
mhp	Median house price paid by local authority Source: Median house prices for administrative geographies available at www.ons.gov.uk
mww	Median weekly wage-Gross (£) for all employee jobs by local authority in England Source: Earnings and hours worked, place of residence by local authority available at www.ons.gov.uk
psq	Population density defined as population per square kilometre Source: Estimates of the population for the UK available at www.ons.gov.uk

primary support reason for their care is related to mental health support. These include nursing, supported accommodation, direct payments, homecare, supported living, other long term care, and other short term support, which are recorded under "Social Care". Intriguingly, the DCLG revenue account reveals evidence that the budget allocated to MHS varies substantially across the English (upper tier) local authorities. For example, in 2016/17 MHS spending by these local authorities ranged between 0.11% (Wandsworth) and just below 53% (Harrow) of their total public health budgets, respectively. While the figures were similar in 2017/18, they were between 0.43% (Halton) and just above 61%

(North Somerset) in 2018/19. Such disparities grew significantly in 2019/20.

In this section, we employ the newly established model and methods to analyse the municipal disparities in the MHS spending in England. Being able to explain such phenomenons is an important step toward having a better understanding of impacts and implications of the UK 2013 public health reforms. Particularly, we would like to understand whether their intended objectives, i.e. to improve the nation's health and well-being and to reduce inequalities at both national and local levels, are achieved. A number of previous studies have applied a traditional reduced form demand and supply framework in which local authorities are treated as statistical units and the municipal disparity in their spending is explained in relation to a set of risk factors. In our view, this is an example of empirical questions where a varying-coefficient panel data model incorporating spatially correlated error components, can render the investigation much more fruitful. Within the context of the MHS, the varying-coefficient process enables non-linear interactions between risk factors of mental health need (e.g. percentage of people aged under 14) and authority specific attributes that represent local preferences and policies. Moreover, the error component structure on the disturbance incorporates unobservable spatial interaction among the local authorities as well as individual heterogeneity.

The study in this section focuses on 151 councils in England, which have social services responsibility, out of 333 local authorities. However, two local authorities, namely City of London and Isles of Scilly, are excluded from our analysis due their unusual socio-economic and demographic characteristics. On the time dimension, we focus on the observation period between 2016/17 and 2019/20, which reflects our interest on the impact of the 2013 government's public health reform and the reduction in its spending on the public health grant during the period. These lead therefore to $N = 149$ and $T = 4$. Below we begin by first establishing the empirical model.

4.1. Empirical model

Since the objective of our study is to analyse the disparities of mental health spendings across local authorities in England, our dependent variable is the MHS by local authority standardized by the total population in each local authority. In the study that follows, we denote per capita measure of the standardised MHS by mhs, and assume that the data generating process behind the mhs is

$$mhs = X\beta_0(Z) + u, \tag{4.1}$$

where $\beta_0(Z)$ is a vector of smooth functions and u follows a spatially correlated error component process, which was thoroughly defined in Section 2.1. We will now discuss the individual components that comprise model (4.1) in more detail.

4.1.1. Regressors and covariate

Let us begin with $X = (X_1, \ldots, X_D)$ and Z. Regarding the former, the first proposition is to include an intercept term in the model by setting $X_1 = 1$, which implies that

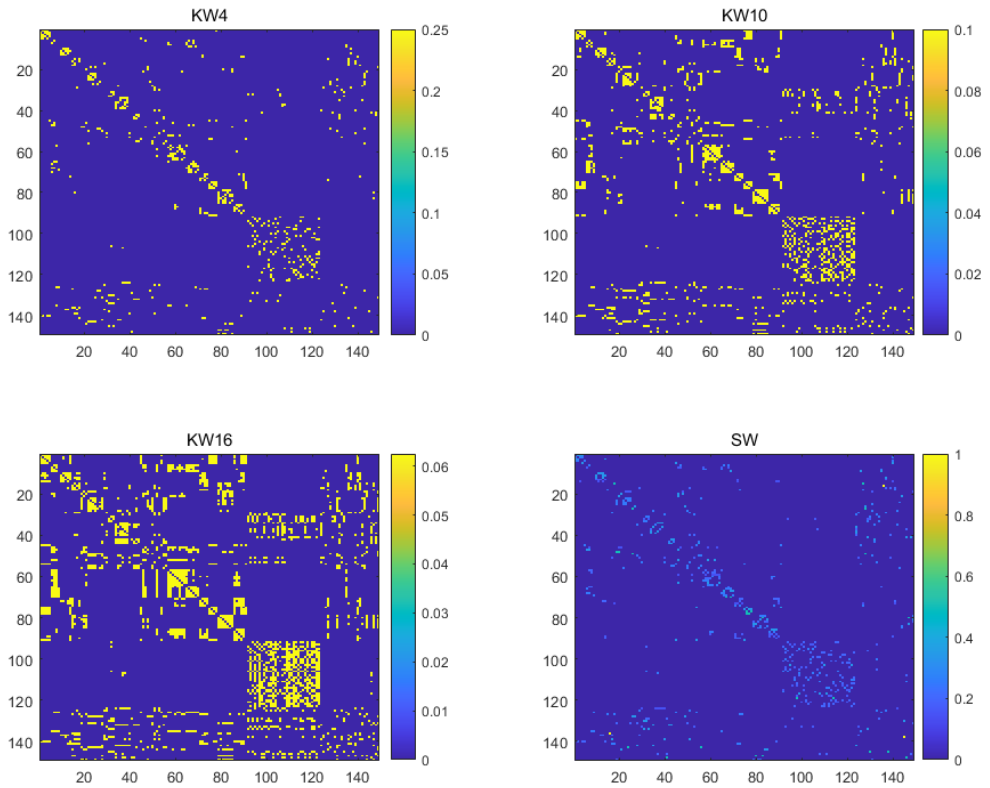
$$mhs = \beta_{0,1}(Z) + X^* \beta_0^*(Z) + u, \tag{4.2}$$

where $X^* = (X_2, \ldots, X_D)$ and $\beta_0^*(Z) = (\beta_{0,2}(Z), \ldots, \beta_{0,D}(Z))^\top$. The remaining regressors are derived from two sources. Firstly, we selected a set of explanatory variables suggested by the literature as area-level characteristics potentially linked to mental health needs (see e.g. McCrone and Jacobson (2004), Aziz et al. (2003), and Moscone et al. (2007)). Our study explains the municipal disparity in mental health spending based on a set of risk factors, namely: (i) Population density, (ii) Percentage of male population, (iii) Percentage of population under 14 year of age, (iv) Standardized mortality ratio, (v) Number of jobs, (vi) Percentage of households headed by lone parent, and (vii) Number of unemployment claimants. Finally, we include (ix) Median house price, and (x) Median weekly wage in order to control for the supply-side factors. Table 6 presents descriptions and sources of the data used in detail.

Moreover, it is important to note that the English local authorities have considerable autonomy under the reformed system to (a) allocate resources from central government among various local services (e.g. education, housing, leisure, community resources and social services), and (b) prioritise particular areas and client groups in line with local interpretations of need. Therefore, their actual spending will likely reflect local policies and preference rather than the standard spending assessment by the central government. Hence, it is implausible to assume, for example, that the percentage of people aged under 14 (%POPu14) would have the same effect on *mhs* across all the local authorities. In the light of this argument, we will focus our study on two strategies. Firstly, we define the covariate Z in order to take into account a political influence. The basis of this idea is from a hypothesis that there might be councils that decide (based on their political beliefs, for example) to give more weight in terms of resources to the elderly while others to the youths. In the practical analysis, the covariate is represented by the percentage of voters with right-wing ideology, *vote* hereafter. More specifically, it is the percentage of voters that have voted for the Conservative and UK Independence Parties in the local government elections. Secondly, the covariate Z is defined as total public health spending by local authority standardized by the total population in each authority, *tph* hereafter. The purpose of this is to take into consideration a type of Engel's law, which might be in operation within the context of MHS. To understand this idea more clearly, let us recall the Engel's law in economics which suggests that the poorer a family is, the larger the budget share it spends on nourishment. Within our panel data model, the varying coefficient specification helps to highlight local authorities' views about each of

the exogenous variables. For example, effect of the percentage of population under 14 on MHS is higher when TPH is low suggests that the variable is considered an essential determinant. On the other hand, effect of the percentage of male population on MHS is lower when TPH is low suggests that the variable is considered to be important though not essential.

Figure 1: Dependence implied by weight matrices under consideration



4.1.2. Spatial Error Dependence (SED) versus Spatial Lag Dependence (SLD)

There often exists an association between MHS spendings made by two or more local authorities. In the literature, such an association is often modelled based on either the SLD or SED. The SLD model is useful for modelling endogenous effects, which explain variations in individual behaviour by the prevalence of the behaviour in a group, contextual effects, which explain individual behaviour by the variation of background characteristics of the group, and correlated effects, which assess whether individuals facing a similar environment or sharing similar individual characteristics will behave the same way. However,

Table 7: Descriptive statistics

	Mean	StD	Min	Max
tph	65.793	24.372	29.739	172.647
mhs	13.265	7.163	0.100	53.710
nuc	5,169.98	4,228.38	105.00	48,145.00
pmp	0.495	0.009	0.473	0.553
*pu*14	0.172	0.020	0.135	0.247
smr	967.366	131.461	583.100	1345.800
noj	202,926	175,994	19,000	2,130,000
plp	0.107	0.028	0.042	0.216
mhp	274,133	173,263	105,000	1,425,000
mww	469.094	77.396	332.100	784.400
psq	2823.152	3367.706	63.000	16425.320

its usefulness is diminished by its inability to disentangle or to identify these effects, i.e. the so-called reflection problem posed by Manski (1993). Within the context of our model, there are enough reasons to believe that a more relevant type of dependence is the SED. In (4.2), since Z represents the authorities specific socio-demographic/economic attributes, $\beta_{0,1}(Z)$ can help to indirectly model the contextual and correlated effects. Furthermore, the non-linear interactions, which are modelled via our varying coefficient specification, can help to capture these effects even more effectively. In addition, measurement errors that spill across grid boundaries, for example, can easily lead to the SED. Otherwise, there may exist unobservable latent variables that might be unaccounted for in the model. For instance, closure of a large psychiatric hospital, which serves patients from various municipalities, clearly has an impact on social care sector across a wide territory. Such a closure of hospitals, which has been one of the most prominent features of mental health policy in the UK for some years, will substantially increase the need for social care services across a wide area and ultimately influencing expenditure. Another example would be the provision of high-secure and medium-secure units for people with forensic needs, often organised at multi-regional level in line with nationally agreed population catchment areas. Their funding may be a NHS hospital, NHS trust, or other independent provider's responsibility, but there will be again a social care shared (spatial) resource effect of not providing these services. Other sources of unobserved spatial concentration could be suggested as the high psychiatric hospital admission in two or more neighbouring authorities, which may be caused by noise pollution from airports. Noise has been the major environmental issue in the field of aviation, primarily impacting residential communities close to airports by affecting community annoyance, sleep deprivation, and mental health issues.

4.1.3. Spatial weight matrices

Regarding the SAR in (2.3), how and to what extent the MHS by a local authority depends on that of the others are captured by elements in the matrix W_N. In spatial econometrics in general, a weight matrix is constructed based either on geographical or socio-economic/demographic distances of individuals. Within the context of our model, we argue that geographical-based weight matrices are more effective due to a number of reasons. Firstly, they are exogenous to the model and also time-non-varying. These conditions are required, but cannot be guaranteed when adopting weights based on socio-economic/demographic distance metrics. Moreover, our varying-coefficient specification enables modelling non-linear interactions between the risk factors of mental health need and authorities specific socio-economic/demographic attributes. Hence, it is reasonable to assume that socio-economic/demographic interactions of MHS spending are fully captured within the model. In the current section, we consider a similar set of weight matrices to that used in Section 3, so that we can analyse if and how estimation results change with weight matrices that differ in their degree of sparseness. In particular, we consider weights matrices, which are constructed based on (i) the k-nearest neighbours criteria, where k is 4, 10, or 16, and (ii) sphere of influence. Figure 1 depicts spatial dependence implied by these weight matrices, which are referred to hereafter as K4, K10, K16 and SW, respectively.

4.2. Empirical analysis

We begin with basic data exploration before discussing estimation results in detail.

4.2.1. Basic data exploration

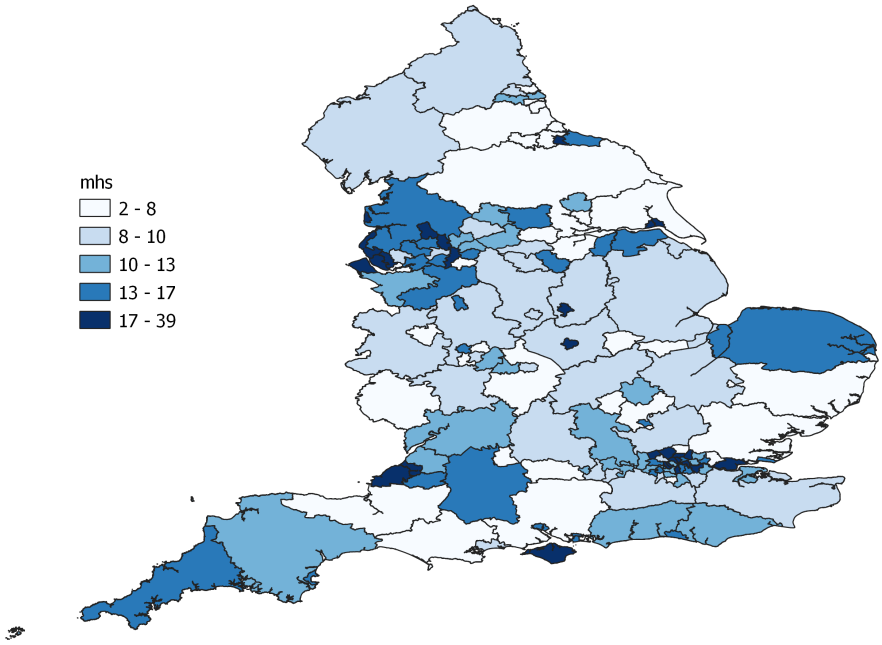
Table 7 presents descriptive statistics, which describe basic features of the data used in our study. Figure 2 presents the average mhs (i.e. per capita measure of standardised MHS for persons age between 18 and 64) for all the local authorities over 2016/17 to 2019/20. It is evident that mhs tends to distribute in clusters, with the highest concentrations in metropolitan areas such as Greater London, Greater Manchester and Birmingham.

4.2.2. Estimation results

The steps taken in our analysis coincide with the methodological development in Section 2. We first estimate the spatial parameters using the likelihood methods discussed in Section 2.2, then perform variable selection using the SAREC-KLASSO method. Once irrelevant regressors are identified, we employ the test procedure discussed in Section 2.4 (as the third step) to check whether associated functional coefficients are constant functions. The estimation results are summarised in Table 8, and graphically presented in Figures 3 to 12. In these figures, the red lines are confidence bands drawn at the 90% confidence level

$$\left[\hat{\beta}_j(z) - \hat{d}_N, \hat{\beta}_j(z) + \Delta(z)\right], \tag{4.3}$$

Figure 2: Per capita measure of standardised MHS for persons age between 18 and 64 (mhs)



where $\hat{\beta}_j(z)$ is an unpenalised estimate obtained after excluding irrelevant regressors,

$$\Delta(z) = \left\{ d_N + \left[\log 2 - \log\left\{-\log(1-\alpha)\right\}\right] (-2\log h)^{-1/2} \right\} \times \hat{SD}\left\{\hat{\beta}_j(z)\right\},$$

$\alpha = 0.1$, d_N and $\hat{SD}^2\left\{\hat{\beta}_j(z)\right\}$ are both defined in (2.33), whereas the broken blue line is computed as follows

$$\hat{C}_j = \frac{1}{NT}\sum_{i=1}^{T}\sum_{t=1}^{T}\hat{\beta}_j(Z_{it}). \tag{4.4}$$

Below we begin by discussing some important findings on the modelling specifications.

- In both panels of the table, the estimates of the autoregressive parameter increase as higher number of nearest neighbours being taken into consideration. Moreover, the estimates listed in panel (a) are closely similar to those in (b).
- Also in both cases, the outcomes of the variable selection do not variate across different weight matrices used. The selected number of relevant variables are 5 and 3 when z is defined as the percentage of right-wing voters, $vote$, and total public health, tph, respectively. Similarly, the outcomes of the coefficient constancy test

Table 8: Estimation results

(a) z is defined as percentage of right-wing voters ($vote$)

W	ρ	ϕ	$\hat{K}$	$inct$	nuc	pmp	$pu14$	smr	noj	plp	mhp	mww	psq
K4	0.159	1.767	5	✖	·	×	·	×	·	·	·	×	×
				3.833	·	2.561	·	2.277	·	·	·	0.151	0.723
K10	0.208	1.786	5	✖	·	✖	·	×	·	·	·	×	×
				3.738	·	3.415	·	2.139	·	·	·	0.455	1.182
K16	0.266	1.812	5	✖	·	✖	·	×	·	·	·	×	×
				3.268	·	3.453	·	1.872	·	·	·	0.127	1.239
SI	0.208	1.747	5	✖	·	✖	·	×	·	·	·	×	×
				3.633	·	3.745	·	2.341	·	·	·	0.374	0.752
K0	·	·	·	✖	·	✖	✖	✖	✖	✖	✖	✖	✖
				13.007	·	8.765	4.967	10.846	10.743	6.547	4.292	9.829	8.586

(b) z is defined as total public health (tph)

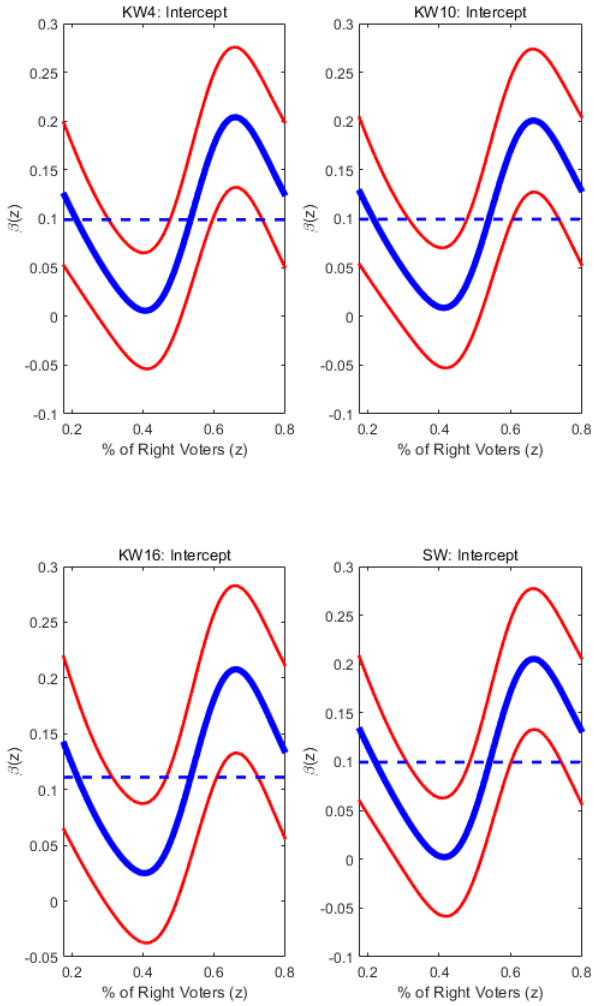
W	ρ	ϕ	$\hat{K}$	$inct$	nuc	pmp	$pu14$	smr	noj	plp	mhp	mww	psq
KW4	0.185	1.785	3	✖	·	·	·	·	·	·	·	✖	✖
				5.891	·	·	·	·	·	·	·	3.376	5.489
KW10	0.255	1.761	3	✖	·	·	·	·	·	·	·	✖	✖
				5.11	·	·	·	·	·	·	·	3.259	5.294
KW16	0.298	1.815	3	✖	·	·	·	·	·	·	·	✖	✖
				4.743	·	·	·	·	·	·	·	3.334	5.416
SW	0.212	1.769	3	✖	·	·	·	·	·	·	·	✖	✖
				5.709	·	·	·	·	·	·	·	3.511	5.021
K0	·	·	·	✖	·	✖	×	✖	✖	×	✖	✖	✖
				17.025	·	5.327	2.137	13.679	8.945	2.214	5.898	5.282	6.606

Note: "×" signifies variables (i) which are selected to be relevant and (ii) whose associated functional coefficients are statistically tested to be constant functions at 5% level. "✖" signifies variables (i) which are selected to be relevant and (ii) whose associated functional coefficients are statistically tested to be non-linear functions at 5% level.

remain largely unchanged across different weight matrices used. However, without taking into consideration the potential SAR and error component structure, the selection suggests that all (but one) variables in each of the panels are relevant. Such a finding is in significant contrast to that based on the SAREC-KLASSO method. In addition, the test statistics of the coefficient constancy test are much larger compared to those associated with $K4$, $K10$, $K16$, SI and $K0$.

- By applying the SAREC-KLASSO method, we have found that the intercept and two other regressors are relevant in explaining the disparities between mental health spending by councils in England, namely median weekly wage (mww) and population per square kilometre (psq). While the effects of these regressors depends non-linearly on total public health (tph), they are independent of the percentage of right-wing voters ($vote$). These findings are also supported by the confidence bands drawn in Figures 3 to 8. These confidence bands are drawn at the 10% significance level and suggest consentaneously that estimates of the functional coefficients for the $intercept$, mww and psq are statistically significant. Furthermore, the dependence of their effects on tph and independence from $vote$, which we highlighted earlier, are also confirmed.
- Interestingly, the percentage of male population (pmp) and standardized mortality

Figure 3: Estimates coefficient function of the intercept: Z represents *vote*

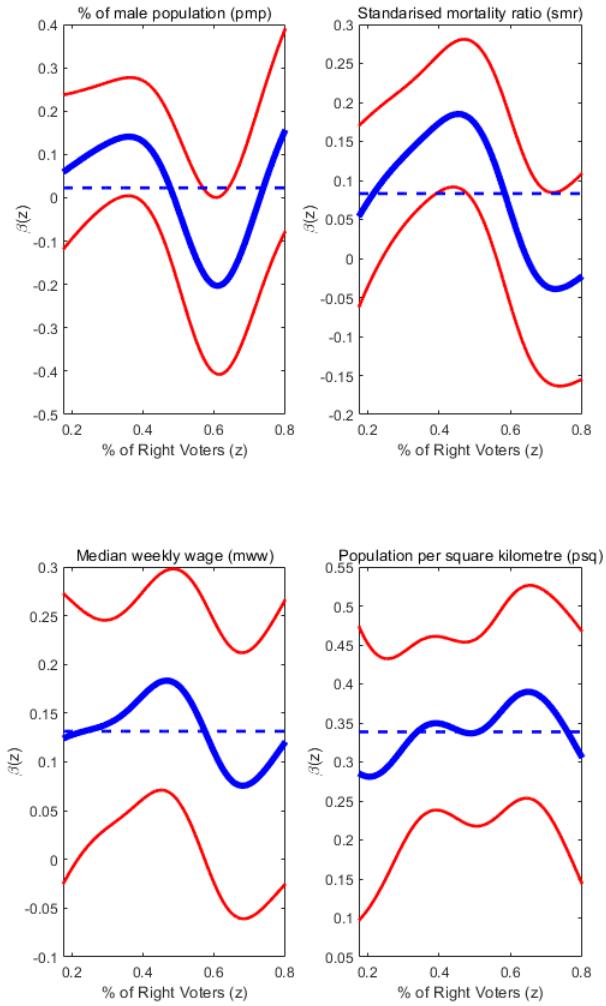


Note: The red solid curves are 90% confidence bands defined in (4.3). The blue broken line is $\hat{C}_j$ in (4.4).

ratio (*smr*) are also selected as relevant, but only when z is defined as *vote*. In this regard, the coefficient constancy test suggests that the effect of *pmp* on *mhs* is dependent of *vote*, whereas that of *smr* is independent. The second part of this finding is unusual because, if they exist, such constant effects of *smr* should also be found in the bottom panel of the table. A closer inspection of the test statistic suggests that the effect of *smr* is a borderline case in which the null hypothesis of coefficient constancy can be rejected by increasing the significant level slightly. Furthermore, the confidence bands drawn in the top-right panel Figures 4 to 8 confirm that the

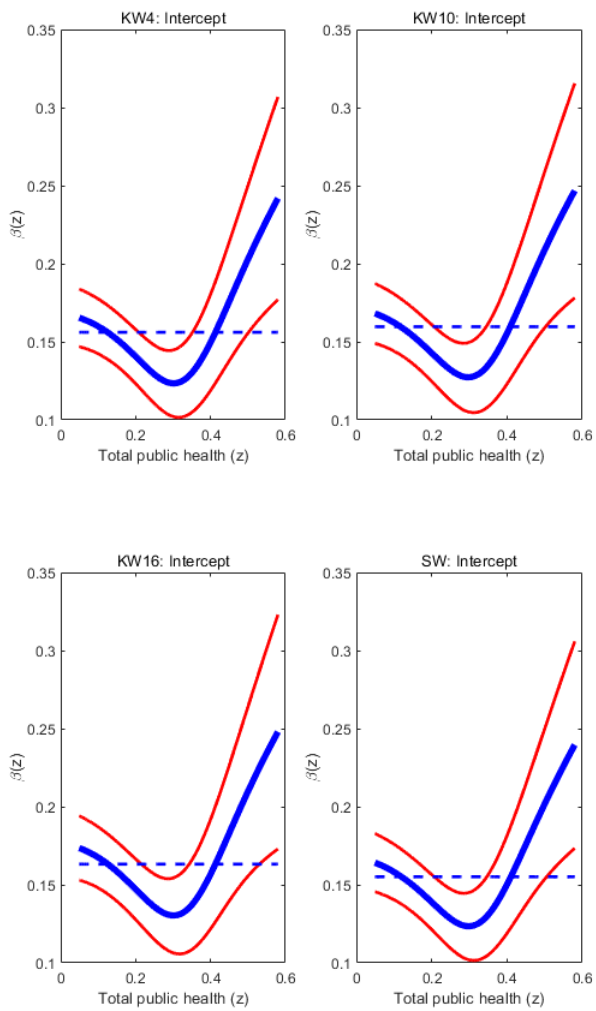
effects of *smr* on *mhs* indeed are dependent of *vote*.

Figure 4: Estimates coefficient functions based on KW4: Z represents *vote*



Note: The red solid curves are 90% confidence bands defined in (4.3). The blue broken line is $\hat{C}_j$ in (4.4).

Figure 5: Estimates coefficient function of the intercept: Z represents tph



Note: The red solid curves are 90% confidence bands defined in (4.3). The blue broken line is $\hat{C}_j$ in (4.4).

We now shift our attention to the empirical implications of the functional coefficients. Let us begin with Figures 3 and 5, which present the estimates of $\beta_{0,1}(Z)$ in (4.2) for each of the weight matrices used.

- From the figures, it is clear that these estimates (and their associated confidence bands) are consistent across the weight matrices used. As the results, our discussion will only concentrate on the top left panels of each figure, which are based on $KW4$, for Z defined by *vote* and by *tph*, respectively.
- The estimate of the functional coefficient suggests that per capita spending on mental health services is higher among the councils, which are dominated by central-right

politics. However, being dominated by central-left politics does not seem to have statistically significant effect.

- Furthermore, we find that *mhs* is treated as a luxury goods in the lower region of the *tph* support, whereas it is viewed as an inferior goods in the higher region.

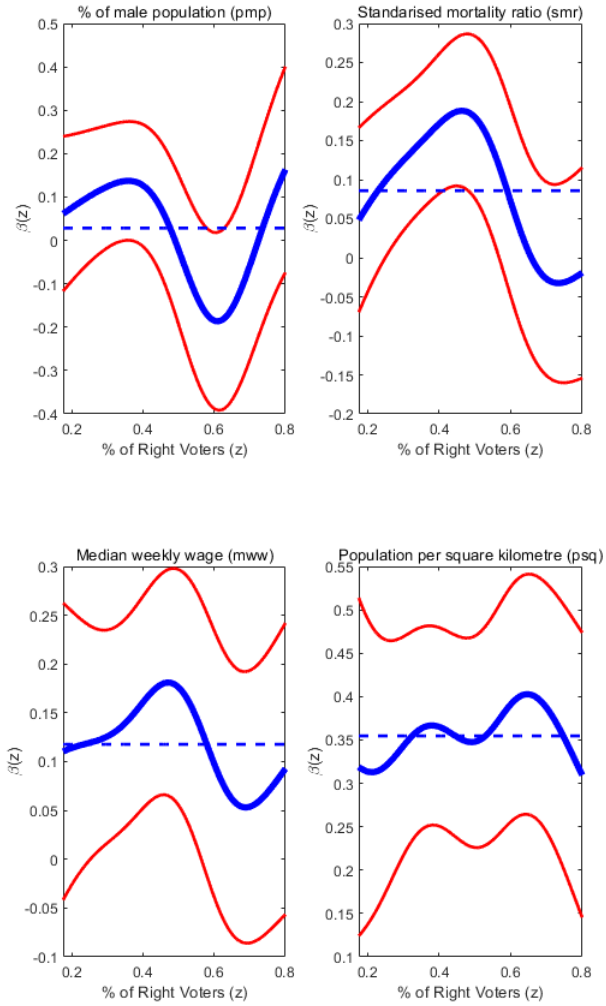
We now analyse estimates of the remaining functional coefficients, which are presented in Figures 3 to 12.

- From the figures, it is clear that these estimates (and their associated confidence bands) are quite consistent across the weight matrices used. As the results, our discussion will only concentrate on Figures 4 and 9, which are based on $KW4$, for Z defined by *vote* and by *tph*, respectively.
- The impact of population density on spending is positive and significant as we expected since we anticipate higher mental health expenditure in inner-city areas that are more densely populated.
- Furthermore, we find strong evidence that *mww* should have a demand-side interpretation instead since councils with higher median weekly earnings seem to spend more on mental health services. Such a result is consistent with that reported in Moscone et al. (2007) for their spatial error model. The estimate of the functional coefficient suggests that impact of *mww* on *mhs* increases between low to medium *tph* (i.e. *mww* plays a similar role to the luxury goods in the Engel curve literature), but decreases between medium to high *mhs* (i.e. *mww* plays a similar role to the inferior goods in the Engel curve literature).
- Moreover, independently to *vote* and *tph*, the percentage of male population does not seem to have a significant effect on the mental health service spending across councils. However, by conditioning its effect on the UK political spectrum, it seems that *pmp* has a positive (negative) impact on *mhs* in councils that are dominated by central-left (central-right) politics.
- Similarly, independently to *vote* and *tph*, the standardised mortality ratio does not seem to have a significant effect on the mental health service spending across councils. However, by conditioning its effect on the UK political spectrum, it seems that *smr* has a positive impact on *mhs* in councils that are dominated by central-left politics.

5. Conclusions

The research in this paper focuses on two of the most discussed areas of methodological development in panel data analysis, namely spatial error dependence and varying coefficient panel data models. We established a new varying-coefficient panel data model that includes spatially correlated error components and introduced the model's estimation procedure and various novel inference methods. Our estimation procedure is an extension of the quasi-maximum likelihood method for spatial panel data regression to the conditional

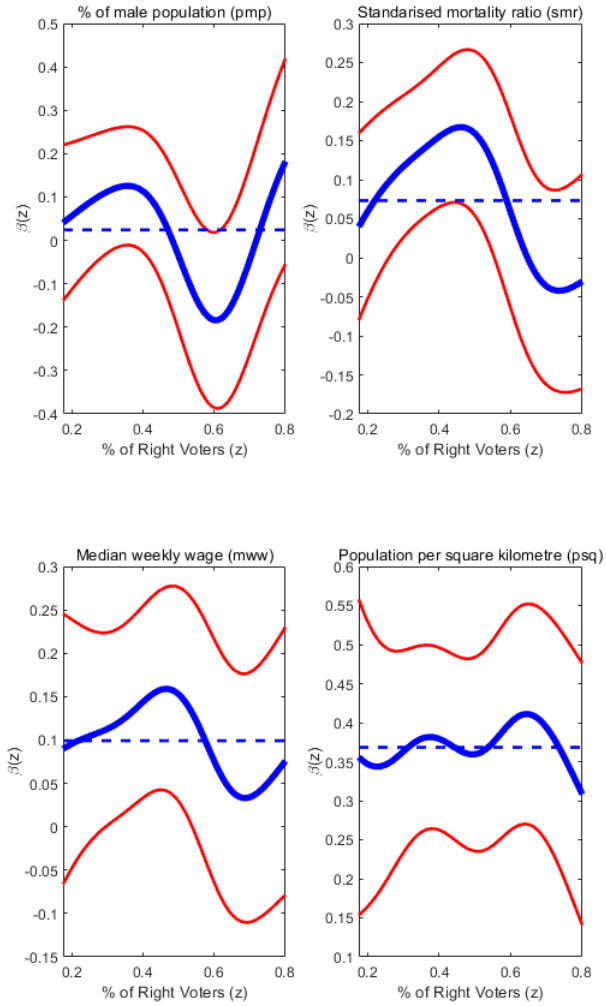
Figure 6: Estimates coefficient functions based on KW10: Z represents *vote*



Note: The red solid curves are 90% confidence bands defined in (4.3). The blue broken line is $\hat{C}_j$ in (4.4).

local kernel-weighted likelihood, whose asymptotic properties were established based on a set of primitive assumptions often seen in studies in the nonparametric literature. Moreover, we established a novel variable selection procedure by extending the Kernel Least Absolute Shrinkage and Selection Operator technique to panel data analysis where there exists the Cliff-Ord-type models of spatial error dependence. We also extended our procedure to handle selection in a more complex specification known in the literature as the semi-varying coefficient model. Furthermore, we conduct extensive simulation exercises in order to examine the finite sample performance and robustness of our proposed procedures

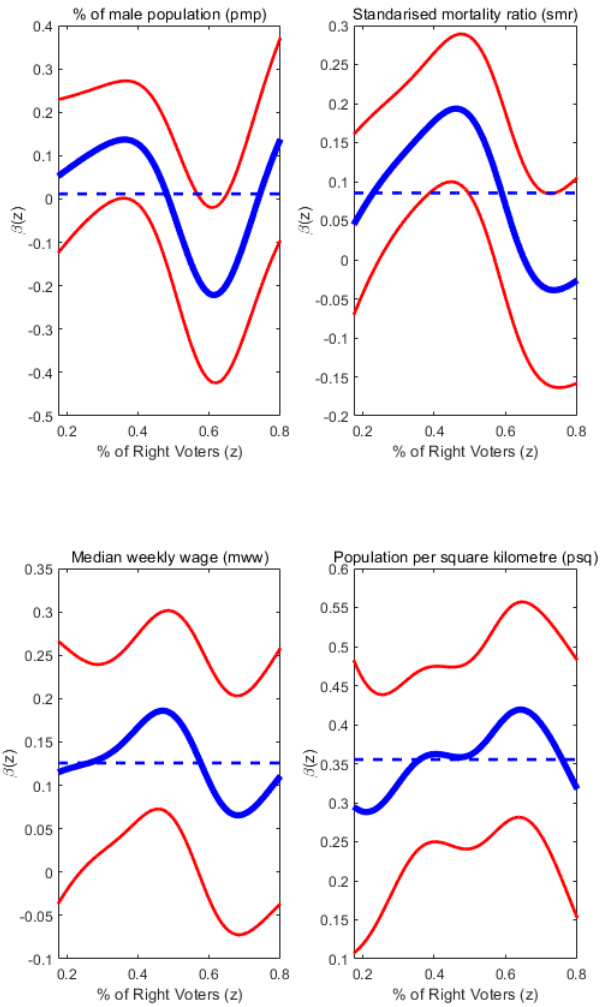
Figure 7: Estimates coefficient functions based on KW16: Z represents *vote*



Note: The red solid curves are 90% confidence bands defined in (4.3). The blue broken line is $\hat{C}_j$ in (4.4).

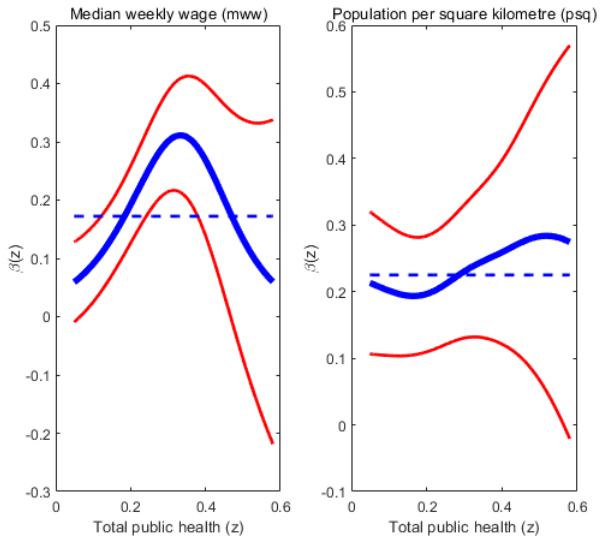
and illustrated their practical applicability by applying them to analyse municipal disparities in MHS spending by councils in England. Specifically, we studied the interaction between a set of demand and supply factors of mental health needs and local authorities specific political and economic attributes, namely the political preference/ideology and total public health spending.

Figure 8: Estimates coefficient functions based on SW: Z represents *vote*



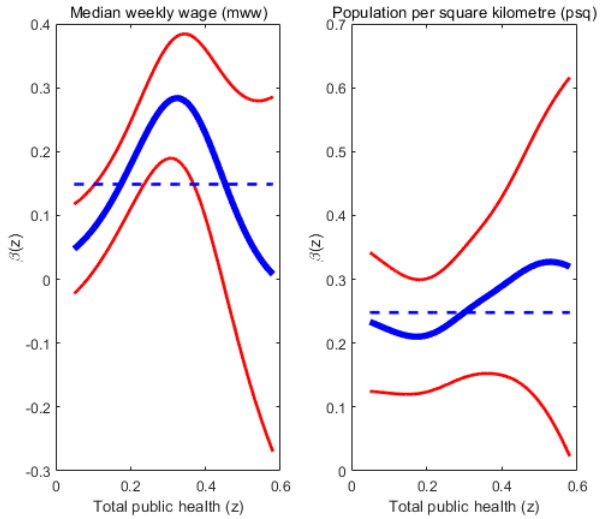
Note: The red solid curves are 90% confidence bands defined in (4.3). The blue broken line is $\hat{C}_j$ in (4.4).

Figure 9: Estimates coefficient functions based on KW4: Z represents *tph*



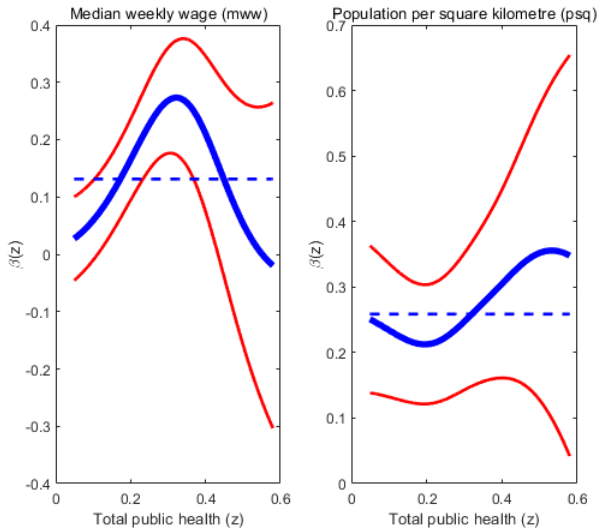
Note: The red solid curves are 90% confidence bands defined in (4.3). The blue broken line is $\hat{C}_j$ in (4.4).

Figure 10: Estimates coefficient functions based on KW10: Z represents *tph*



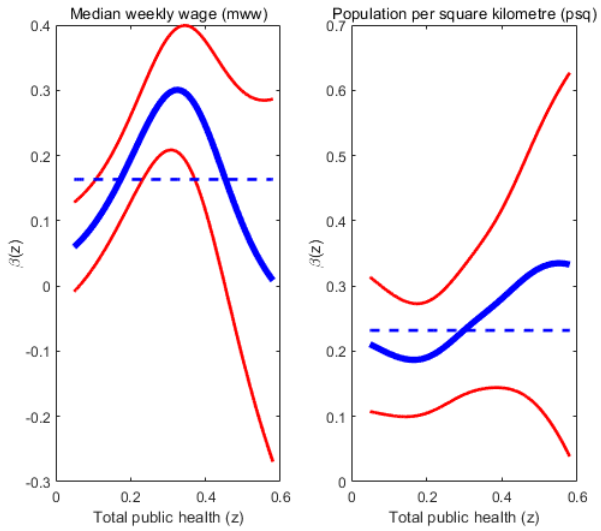
Note: The red solid curves are 90% confidence bands defined in (4.3). The blue broken line is $\hat{C}_j$ in (4.4).

Figure 11: Estimates coefficient functions based on KW16: Z represents *tph*



Note: The red solid curves are 90% confidence bands defined in (4.3). The blue broken line is $\hat{C}_j$ in (4.4).

Figure 12: Estimates coefficient functions based on SW: Z represents *tph*



Note: The red solid curves are 90% confidence bands defined in (4.3). The blue broken line is $\hat{C}_j$ in (4.4).

6. Appendices

This section consists of seven appendices. Appendix A provides a set of useful definitions, whereas Appendix B discusses the proof of Lemma 2.1 , Theorem 2.1 , and Theorem 2.2(a). Appendix C presents a set of lemmas that will be useful for the proof in Appendices D to G that follow. Appendices D presents the proof of Theorem 2.2(b) and Theorem 2.3, while Appendices E to G provide detailed proof for the results in Sections 2.3 to 2.5, respectively.

A. Definitions

For an arbitrary vector $P \in \mathbb{R}^{n\times 1}$, define its Euclidean norm as

$$\|P\| := \sqrt{p_1^2 + \cdots + p_n^2}.$$

For an arbitrary matrix $Q \in \mathbb{R}^{n\times m}$, define its Frobenius norm as $\|V\|_F := \left(\sum_{i=1}^n \sum_{j=1}^m |v_{ij}|^2\right)^{1/2}$. Also, define a sequence of combination pairs $(i,t) \equiv it = \{11, 21, \ldots, N1, 12, \ldots, N2, 13, \ldots, NT\}$. Let $m = (m_{it,d}) \in \mathbb{R}^{NT\times D}$ denote an arbitrary $NT \times D$ matrix with rows

$$m_{11}^\top, m_{21}^\top, \ldots, m_{N1}^\top, m_{12}^\top, \ldots, m_{N2}^\top, m_{13}^\top, \ldots, m_{NT}^\top,$$

i.e. m_{it} is $D \times 1$ and $m_{it}^\top$ is $1 \times D$, and columns $v_1, v_2, \ldots, v_D$, i.e. v_d is $NT \times 1$. Moreover, define

$$B_0 = (\beta_0^\top(Z_{11}), \ldots, \beta_0^\top(Z_{N1}), \beta_0^\top(Z_{12}), \ldots, \beta_0^\top(Z_{N2}), \ldots, \beta_0^\top(Z_{NT}))^\top \in \mathbb{R}^{NT\times D}$$

and

$$\tilde{B} = (\tilde{\beta}^\top(Z_{11}), \ldots, \tilde{\beta}^\top(Z_{N1}), \tilde{\beta}^\top(Z_{12}), \ldots, \tilde{\beta}^\top(Z_{N2}), \ldots, \tilde{\beta}^\top(Z_{NT}))^\top \in \mathbb{R}^{NT\times D}.$$

B. Proof of Lemma 2.1, Theorem 2.1 and Theorem 2.2(a)

B.1. Proof of Lemma 2.1:

The derivation of (2.12) is straightforward since the third term of (2.9) can be written as

$$\begin{aligned} -\frac{1}{2\sigma_v^2}\{u_N^\top \bar{\mathbb{Q}}_N^\top K_N \bar{\mathbb{Q}}_N u_N\} &= -\frac{1}{2\sigma_v^2}\{(\beta_0(z) - \beta(z))^\top \ddot{X}_N^\top K_N \ddot{X}_N (\beta_0(z) - \beta(z))\} \\ &\quad - \frac{1}{2\sigma_v^2}\{\ddot{u}_{0N}^\top K_N \ddot{u}_{0N}\}. \end{aligned} \tag{B.1}$$

In addition, the solutions for the optimization problem $\bar{\ell}_z^c(\delta) \equiv \max_{\beta,\sigma_v^2} \bar{\ell}_z(\beta, \sigma_v^2, \phi, \rho)$ are:

$$\bar{\beta}(z) = \{E[\ddot{X}_N^\top K_N \ddot{X}_N | z]\}^{-1} E[\ddot{X}_N^\top K_N \ddot{X}_N | z]\beta_0(z) = \beta_0(z)$$

and

$$\bar{\sigma}_v^2 = (1/NT)\sigma_{v_0}^2 \mathrm{TR}[\mathbb{Q}_{0N}\bar{\mathbb{Q}}_N^\top K_N \bar{\mathbb{Q}}_N] \left[\sum_{j=1}^N \sum_{s=1}^T K_h(Z_{js} - z)\right]^{-1}$$

Hence, substitution of these into (2.12) leads immediately to (2.13).

B.2. Proof of Theorem 2.1:

We first recall $\tilde{\ell}_z^c(\delta) \equiv \max_{\beta,\sigma_v^2} \ \ell(\beta, \sigma_v^2, \phi, \rho)$, which was shown in (2.11) to be

$$\tilde{\ell}_z^c(\delta) = -\frac{1}{2}\left[\log(2\pi\tilde{\sigma}_v^2) + \log|\mathbb{Q}_N| + 1\right]\sum_{j=1}^{N}\sum_{s=1}^{T} K_h(Z_{js} - z), \tag{B.2}$$

and $\bar{\ell}_z^c(\delta) \equiv \max\limits_{\beta,\sigma_v^2} \bar{\ell}_z(\beta, \sigma_v^2, \phi, \rho)$ for which

$$\bar{\ell}_z^c(\delta) = -\frac{1}{2}\left[\log(2\pi\bar{\sigma}_v^2) + \log|\mathbb{Q}_N| + 1\right]\sum_{j=1}^{N}\sum_{s=1}^{T} K_h(Z_{js} - z) \tag{B.3}$$

established in Lemma 2.1. Given the unique identification of δ_0, consistency of $\hat{\delta} = (\hat{\phi}, \hat{\rho})^\top$ as stated in the theorem follows from the convergence of $\frac{1}{NT}[\tilde{\ell}_z^c(\delta) - \bar{\ell}_z^c(\delta)]$ uniformly to zero on Δ. This can be established in two steps, namely (i) establishing point-wise convergence of $\frac{1}{NT}\tilde{\ell}_z^c(\delta)$ to $\frac{1}{NT}\bar{\ell}_z^c(\delta)$, and (ii) establishing uniform Lipschitz continuity of $\frac{1}{NT}[\tilde{\ell}_z^c(\delta) - \bar{\ell}_z^c(\delta)]$ over $\delta \in \Delta$.

To perform the first step, we need to note firstly that

$$\frac{1}{NT}[\tilde{\ell}_z^c(\delta) - \bar{\ell}_z^c(\delta)] = -\frac{1}{2}\log\left(\frac{\tilde{\sigma}_v^2}{\bar{\sigma}_v^2}\right).$$

Therefore, we only have to show that $\tilde{\sigma}_v^2 = \bar{\sigma}_v^2 + O_P((NT)^{-1/2})$. To this end, let us write $\tilde{u}_{js} = \ddot{y}_{js} - \ddot{X}_{js}\tilde{\beta}_{it} = \ddot{X}_{js}(\beta_0 - \tilde{\beta}_{it}) + \ddot{u}_{0js}$, where $\ddot{u}_{0js} = \ddot{y}_{js} - \ddot{X}_{js}\beta_0$, so that

$$\begin{aligned}
\tilde{u}_{js}^\top\tilde{u}_{js} &= \{\ddot{X}_{js}(\beta_0 - \tilde{\beta}_{it}) + \ddot{u}_{0js}\}^\top\{\ddot{X}_{js}(\beta_0 - \tilde{\beta}_{it}) + \ddot{u}_{0js}\} \\
&= \{\beta_0 - \tilde{\beta}_{it}\}^\top\ddot{X}_{js}^\top\ddot{X}_{js}\{\beta_0 - \tilde{\beta}_{it}\} + \{\beta_0 - \tilde{\beta}_{it}\}^\top\ddot{X}_{js}^\top\ddot{u}_{0js} + \ddot{u}_{0js}^\top\ddot{X}_{js}\{\beta_0 - \tilde{\beta}_{it}\} + \ddot{u}_{0js}^\top\ddot{u}_{0js}.
\end{aligned}$$

Making use of (2.10) enables writing $\tilde{\beta}_{it} = \beta_0 + \hat{u}_{0it}$, where

$$\tilde{u}_{0it} = \left[\sum_{j=1}^{N}\sum_{s=1}^{T}\ddot{X}_{js}^\top\ddot{X}_{js}K_h(Z_{js} - Z_{it})\right]^{-1}\sum_{j=1}^{N}\sum_{s=1}^{T}\ddot{X}_{js}^\top\ddot{u}_{0js}K_h(Z_{js} - Z_{it})$$

and hence

$$\tilde{u}_{js}^\top\tilde{u}_{js} = \tilde{u}_{0it}^\top\ddot{X}_{js}^\top\ddot{X}_{js}\tilde{u}_{0it} - \tilde{u}_{0it}^\top\ddot{X}_{js}^\top\ddot{u}_{0js} - \ddot{u}_{0js}^\top\ddot{X}_{js}\tilde{u}_{0it} + \ddot{u}_{0js}^\top\ddot{u}_{0js}. \tag{B.4}$$

Moreover,

$$\begin{aligned}
\tilde{\sigma}_v^2 &= \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\tilde{u}_{js}^\top\tilde{u}_{js}K_h(Z_{js} - Z_{it}) \\
&= \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\left\{\tilde{u}_{0it}^\top\ddot{X}_{js}^\top\ddot{X}_{js}\tilde{u}_{0it} - \tilde{u}_{0it}^\top\ddot{X}_{js}^\top\ddot{u}_{0js} - \ddot{u}_{0js}^\top\ddot{X}_{js}\tilde{u}_{0it} + \ddot{u}_{0js}^\top\ddot{u}_{0js}\right\}K_h(Z_{js} - Z_{it})
\end{aligned}$$

and

$$\begin{aligned}
\tilde{\sigma}_v^2 - \bar{\sigma}_v^2 &= \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\{\ddot{u}_{0js}^\top\ddot{u}_{0js} - E(\ddot{u}_{0js}^\top\ddot{u}_{0js}) \\
&+ [\mathbf{R}_{1,js} - E(\mathbf{R}_{1,js})] - [\mathbf{R}_{2,js} - E(\mathbf{R}_{2,js})] + E(\mathbf{R}_{1,js}) - 2E(\mathbf{R}_{2,js})\}K_h(Z_{js} - Z_{it})
\end{aligned}$$

where

$$\mathbf{R}_{1,js} = \tilde{u}_{0it}^\top \ddot{X}_{js}^\top \ddot{X}_{js} \tilde{u}_{0it} \text{ and } \mathbf{R}_{2,js} = \tilde{u}_{0it}^\top \ddot{X}_{js}^\top \ddot{u}_{0js}. \tag{B.5}$$

In this regard, we note firstly that

$$\frac{1}{NT} \ddot{u}_{0N}^\top \ddot{u}_{0N} = \bar{\sigma}_v^2 + O_P((NT)^{-1/2}) \tag{B.6}$$

under conditions required in Assumption A. The remaining terms in (B.5) can be dealt with:

$$E\left(\frac{1}{NT}\sum_{j=1}^N \sum_{s=1}^T \mathbf{R}_{1,js}\right) = \frac{1}{NT}\sum_{j=1}^N \sum_{s=1}^T E\{\tilde{u}_{0it}^\top \ddot{X}_{js}^\top \ddot{X}_{js} \tilde{u}_{0it}\} = O((NT)^{-2}h^{-1}) \tag{B.7}$$

and

$$E\left(\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \mathbf{R}_{2,js}\right) = \frac{1}{NT}\sum_{j=1}^N \sum_{s=1}^T E\left\{\tilde{u}_{0it}^\top \ddot{X}_{js}^\top \ddot{u}_{0it}\right\} = O((NTh^1)^{-1/2}). \tag{B.8}$$

To obtain these results requires noting firstly that we have, by using the spectral decomposition of a symmetric positive (or negative) definite matrix and the Cauchy-Schwartz inequality,

$$\begin{aligned}
\sum_{j=1}^N \sum_{s=1}^T E\left\{\tilde{u}_{0it}^\top \ddot{X}_{js}^\top \ddot{X}_{js} \tilde{u}_{0it}\right\} &\leq \left(\frac{\sqrt{p}}{\tilde{\gamma}_{it}^{\min}}\right)^2 \sum_{j=1}^N \sum_{s=1}^T E\left\{\check{u}_{0it}^\top \ddot{X}_{js}^\top \ddot{X}_{js} \check{u}_{0it}\right\} \\
&\leq \left(\frac{1}{\tilde{\gamma}_{js}^{\min}}\right)^2 \ddot{\lambda}_{js}^{\max} \{E||\check{u}_{0N}||^2\}^{1/2} \{E||\check{u}_{0N}||^2\}^{1/2} = O((NT)^{-2}h^{-1}),
\end{aligned} \tag{B.9}$$

where $\check{u}_{0it} = \frac{1}{NT}\sum_{j=1}^N \sum_{s=1}^T \ddot{X}_{js}^\top \ddot{u}_{0js} K_h(Z_{js} - Z_{it})$, $\ddot{\lambda}_{it}^{\max}$ is the maximum eigenvalue of the $D \times D$ symmetric positive definite matrix of $\ddot{X}_N^\top \ddot{X}_N$, and

$$E\left|\left|\sum_{i=1}^N \sum_{t=1}^T \check{u}_{0it}\right|\right|^2 = \sum_{l=1}^p \sum_{i=1}^N \sum_{t=1}^T E(\check{u}_{0it,l}^2) = O((NT)^{-2}h^{-1}). \tag{B.10}$$

This is so because

$$\begin{aligned}
&\sum_{i=1}^N \sum_{t=1}^T E(\check{u}_{0it,l}^2) = \sum_{i=1}^N \sum_{t=1}^T E[E\left(\check{u}_{0it,l}^2 | Z_{js}, Z_{it}\right)] \\
&= \frac{1}{(NTh)^2} \sum_{i=1}^N \sum_{t=1}^T \sum_{j=1}^N \sum_{s=1}^T \int E(\ddot{X}_{js,l}^2 \ddot{u}_{0js}^2 | Z_{js}, Z_{it}) K^2\left(\frac{Z_{js} - Z_{it}}{h}\right) f_z(Z_{js}) f_z(Z_{it}) dZ_{js} dZ_{it} \\
&= \frac{1}{(NT)^2 h} \sum_{j=1}^N \sum_{s=1}^T \int \left\{\sum_{i=1}^N \sum_{t=1}^T E(\ddot{X}_{js,l}^2 \ddot{u}_{0js}^2 | Z_{js}, Z_{it}) f_z(Z_{it}) dZ_{it}\right\} f_z(Z_{js}) K^2(v) dv \\
&\leq \frac{1}{(NT)^2 h} \sigma_{v,0}^2 \mathcal{K}^2 \sum_{j=1}^N \sum_{s=1}^T E(\ddot{X}_{js,l}^2 | Z_{js}) f_z(Z_{js}) = \frac{1}{(NT)^2 h} \sigma_{v,0}^2 \mathcal{K}^2 E(\ddot{X}_{js,l}^2) \\
&= O((NT)^{-2}h^{-1}).
\end{aligned} \tag{B.11}$$

Finally, applying the Markov inequality to (B.7) and (B.8) and (B.6) lead to

$$\tilde{\sigma}_v^2 = \bar{\sigma}_v^2 + O_P((NT)^{-1/2}), \tag{B.12}$$

so that $\tilde{\ell}_z^c(\delta) = \bar{\ell}_z^c(\delta) + O_P((NT)^{-1/2})$.

Now, we consider the uniform Lipschitz continuity of $\tilde{\sigma}_v^2 - \bar{\sigma}_v^2$. Let us begin with

$$\sup_{||\delta-\delta^*||<\epsilon} |\tilde{\sigma}_v^2(\delta) - \bar{\sigma}_v^2(\delta) - \{\tilde{\sigma}_v^2(\delta^*) - \bar{\sigma}_v^2(\delta^*)\}| \\ \leq \sup_{||\delta-\delta^*||<\epsilon} \left|\left|\{\tilde{\sigma}_v^2(\bar{\delta})\}^{(1)} - \{\bar{\sigma}_v(\bar{\delta})\}^{(1)}\right|\right| \cdot ||\delta - \delta^*|| = o_P(1),$$

where $\delta^* \in \Delta$ lies on an ϵ-neighborhood of δ such that $||\delta - \delta^*|| = 0$ as $\epsilon \to 0$, $\bar{\delta}$ lies on the line segment $\{\lambda\delta + (1-\lambda)\delta^*); \lambda \in (0,1)\}$ and $\{\tilde{\sigma}_v^2\}^{(1)}$ and $\{\bar{\sigma}_v^2\}^{(1)}$ denote the gradients of $\tilde{\sigma}_v^2$ and $\bar{\sigma}_v^2$, respectively. Hence, the uniform Lipschitz continuity is established by showing

$$\left|\left|\{\tilde{\sigma}_v^2(\bar{\delta})\}^{(1)} - \{\bar{\sigma}_v^2(\bar{\delta})\}^{(1)}\right|\right| = O_P(1). \tag{B.13}$$

In order to show the boundedness of (B.13), let us expand the difference of the gradients as follows

$$\{\tilde{\sigma}_v^2(\bar{\delta})\}^{(1)} - \{\bar{\sigma}_v^2(\bar{\delta})\}^{(1)} = \frac{1}{NT}\left\{\mathcal{R}_N' - \sigma_{v0}^2 \mathrm{TR}\left[\mathbb{Q}_{0N}\frac{\partial \mathbb{Q}_N^{-1}(\bar{\delta})}{\partial \bar{\delta}}\right] + \mathbf{R}_{1,N}' - \mathbf{R}_{2,N}'\right\},$$

where $\mathcal{R}_N'$, $\mathbf{R}_{1,N}'$ and $\mathbf{R}_{2,N}'$ denote gradients of $\mathcal{R}_N = u_N^\top \mathbb{Q}_N^{-1}(\bar{\delta}) u_N$, $\mathbf{R}_{1,N}(\bar{\delta}) = \sum_{i=1}^N \sum_{t=1}^T \mathbf{R}_{1,it}(\bar{\delta})$ and $\mathbf{R}_{2,N}(\bar{\delta}) = \sum_{i=1}^N \sum_{t=1}^T \mathbf{R}_{2,it}(\bar{\delta})$, respectively. Moreover, $E(\mathcal{R}_N') = \sigma_{v,0}^2 \mathrm{TR}\left[\mathbb{Q}_{0N}\frac{\partial \mathbb{Q}_N^{-1}(\bar{\delta})}{\partial \bar{\delta}}\right]$ and therefore $\frac{1}{NT}\mathcal{R}_N' = \frac{1}{NT}\sigma_{v,0}^2 \mathrm{TR}\left[\mathbb{Q}_{0N}\frac{\partial \mathbb{Q}_N^{-1}(\bar{\delta})}{\partial \bar{\delta}}\right] + O_P((NT)^{-1/2})$ by using similar arguments to (B.6). The rest of the terms can be similarly worked out as follows $E(\mathbf{R}_{1,N}') = O((NTh)^{-3/2})$ and $E(\mathbf{R}_{2,N}') = O((NTh)^{-1/2})$. The detailed derivation of these results are available upon request from the authors. Finally, (B.13) holds due to the Markov inequality.

Now let us consider the unique identification conditions of δ_0. The unique identification of δ_0 is firstly considered by showing the counter argument. Consider the Jensen's inequality below

$$\frac{1}{NT}\left\{\bar{\ell}_z^c(\delta) - \bar{\ell}_z^c(\delta_0)\right\} = \frac{1}{NT}\log|\mathbb{Q}_N \mathbb{Q}_{0N}^{-1}| - \frac{1}{2}\log\left(\frac{\mathrm{TR}[\mathbb{Q}_{0N}\bar{\mathbb{Q}}_N^\top K_N \bar{\mathbb{Q}}_N]}{NT}\right) \leq 0. \tag{B.14}$$

The equality of (B.14) holds when $\mathbb{Q}_N \mathbb{Q}_{0N}^{-1} = \mathbb{Q}_N^{-1}\mathbb{Q}_{0N} = I_{NT}$. Hence δ_0 is not uniquely identified when there is a sequence such that $\delta_N \in D_\epsilon(\delta^*)$ converges to $\delta^* \in \bar{D}_\epsilon(\delta_0) \cap \Delta$ where $D_\epsilon(\cdot)$ and $\bar{D}_\epsilon(\cdot)$ represent an open ϵ-neighborhood and its complement, respectively, and $\lim_{N\to\infty} \mathbb{Q}_N(\delta^*) \to \lim_{N\to\infty} \mathbb{Q}_{0N}$. Hence the unique identification condition requires that

$$\limsup_{N\to\infty}\left\{\max_{\delta \in \bar{D}_\epsilon(\delta_0)\cap\Delta} \bar{\ell}_z^c(\delta)\right\} \neq \limsup_{N\to\infty} \bar{\ell}_z^c(\delta_0)$$

for any δ.

B.3. Proof of Theorem 2.2(a):

The proof of Theorem 2.2(a) follows from that of Theorem 2.1; see in particular the proof of (B.12).

C. Useful lemmas

In this section, we present a set of lemmas that will be useful for the proof that follows. For the sake of clarity in the proof, we simplify the notations, so that $\ddot{X}_{0N} \equiv \ddot{X}_N$ and $\ddot{y}_{0N} \equiv \ddot{y}_N$. Also, let $\ddot{X}_{js}$ be the js-th row of $\ddot{X}_N$ and $\ddot{y}_{js}$ be the js-th element of $\ddot{y}_N = \ddot{y}_N$.

Lemma C.1. *Let Assumptions A to D hold. Then,*

$$\sup_{z\in[0,1]} \left| \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[\ddot{X}_{it}^\top \ddot{X}_{it} K_h(Z_{it} - z) - E\left\{ \ddot{X}_{it}^\top \ddot{X}_{it} K_h(Z_{it} - z) \right\} \right] \right| = O_P \left\{ h^2 + \left(\frac{\log(1/h)}{(NT)h} \right)^{1/2} \right\}$$

$$E\left\{ \ddot{X}_{it}^\top \ddot{X}_{it} K_h(Z_{it} - z) \right\} = f(z)\Omega(z) + O(h^2), \tag{C.1}$$

where $\Omega(z) = E[\ddot{X}_{it}^\top \ddot{X}_{it} | Z_{it} = z]$, *which is assumed to have bounded derivative.*

Proof of Lemma C.1: Let us consider firstly a more general case of this result

$$\begin{aligned} &\sup_{z\in[0,1]} \left| (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[K_h(Z_{it} - z)\xi_{it} - E\left\{ K_h(Z_{it} - z)\xi_{it} \right\} \right] \right| \\ &\quad = O_P \left\{ h^2 + \left(\frac{\log(1/h)}{(NT)h} \right)^{1/2} \right\}, \end{aligned} \tag{C.2}$$

where (ξ_{it}, Z_{it}) are i.i.d. random vectors, ξ_{it} are scalar random variables with $E|\xi_{it}|^s < \infty$, and $\sup_z \int |y|^s f(z, v)dv < \infty$ (where f denotes the joint density of (ξ_1, Z_1)). The proof of (C.2) can be found in various existing works, e.g. Fan and Zhang (2000a). Regarding (C.1):

$$\begin{aligned} E\left\{ \ddot{X}_{it}^\top \ddot{X}_{it} K_h(Z_{it} - z) \right\} &= E\left\{ E[\ddot{X}_{it}^\top \ddot{X}_{it} K_h(Z_{it} - z)|z] \right\} \\ &= h^{-1} \int E[\ddot{X}_{it}^\top \ddot{X}_{it}|z] f(Z_{it}) K\left(\frac{Z_{it} - z}{h} \right) dZ_{it} \\ &= h^{-1} \int E[\ddot{X}_{it}^\top \ddot{X}_{it}|z] f(z + vh) K(v) h dv \\ &= \int E[\ddot{X}_{it}^\top \ddot{X}_{it}|z] \{ f(z) + f'(z)vh + (1/2)f''(z)h^2v^2 + O(h^3) \} k(v) dv \\ &= E[\ddot{X}_{it}^\top \ddot{X}_{it}|z] f(z) + O(h^2), \end{aligned}$$

where $f(z + vh) = f(z) + f'(z)vh + (1/2)f''(z)h^2v^2 + O(h^3)$ is used for the fourth equality.

Lemma C.2. *Let Assumptions A to D hold. Then,*

$$\sup_{z\in[0,1]} \left| \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[\hat{X}_{it}^\top \hat{X}_{it} K_h(Z_{it} - z) - \ddot{X}_{it}^\top \ddot{X}_{it} K_h(Z_{it} - z) \right] \right| = O_P((NTh)^{-1}). \tag{C.3}$$

Proof of Lemma C.2: Let us represent (C.3) by using Taylor expansion as follows

$$\begin{aligned} &\sup_{z\in[0,1]} \left| \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[\hat{X}_{it}^\top \hat{X}_{it} K_h(Z_{it} - z) - \ddot{X}_{it}^\top \ddot{X}_{it} K_h(Z_{it} - z) \right] \right| \\ &\quad \leq \sup_{z\in[0,1]} \left\{ ||\hat{\delta} - \delta|| \cdot \left\| X_N^\top K_N \frac{\partial Q_N^{-1}}{\partial \delta} X_N \right\|_F \right\}, \end{aligned} \tag{C.4}$$

where

$$\frac{\partial Q_N^{-1}}{\partial \rho} = 2[I_T \otimes (I_N - \rho W_N^\top)]\{Q_{0,N} + (1+T\phi)^{-1} Q_{1,N}\}[I_T \otimes (I_N - W_N)], \tag{C.5}$$

$$\frac{\partial Q_N^{-1}}{\partial \phi} = [I_T \otimes (I_N - \rho W_N^\top)] \left\{ \frac{1}{(1+T\phi)^2} Q_{1,N} \right\} [I_T \otimes (I_N - \rho W_N)]. \tag{C.6}$$

The uniform consistency of $\hat{\delta}$ in (C.4) over $z \in [0,1]$ was already established in Theorem 2.1.

Lemma C.3. *Let Assumptions A to D hold, and* $\tilde{\Sigma}(z) = (NT)^{-1} \sum_{j=1}^N \sum_{s=1}^T \ddot{X}_{js}^\top \ddot{X}_{js} K_h(z - Z_{js})$. *Then,*

$$\sup_{z\in[0,1]} \left| \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \left[E\left\{ \ddot{X}_{it}^\top \ddot{X}_{it} K_h(Z_{it} - z) - f(z)\Omega(z) \right\} \right] \right| = O_P\left\{ h^2 + \left(\frac{\log(1/h)}{(NT)h} \right)^{1/2} \right\}, \tag{C.7}$$

$$\sup_{z\in[0,1]} \left| \tilde{\Sigma}(z) - f(z)\Omega(z) \right| = O_P\left\{ h^2 + \left(\frac{\log(1/h)}{(NT)h} \right)^{1/2} \right\}. \tag{C.8}$$

Proof of Lemma C.3: Lemma C.3 follows immediately from (C.1).

Lemma C.4. *Let Assumptions A to D hold. Then*

$$\begin{aligned}
\hat{u}_{it} &= h^{1/2}(NT)^{-1/2} \sum_{j=1}^N \sum_{s=1}^T \hat{X}_{js}^\top \{\hat{X}_{js}[\beta_0(Z_{it}) - \beta_0(Z_{js})] + \hat{u}_{js}\} K_h(Z_{it} - Z_{js}) \\
&= h^{1/2}(NT)^{-1/2} \sum_{j=1}^N \sum_{s=1}^T \ddot{X}_{js}^\top \{\ddot{X}_{js}[\beta_0(Z_{it}) - \beta_0(Z_{js})] + \ddot{u}_{js}\} K_h(Z_{it} - Z_{js}) + R_{1,it} + R_{2,it},
\end{aligned}$$

where

$$\begin{aligned}
\ddot{u}_{js} &= \ddot{y}_{js} - \ddot{X}_{js}\beta_0(Z_{it}) \\
R_{1,it} &= h^{1/2}(NT)^{-1/2} X_N^\top K_N \frac{\partial Q_N^{-1}}{\partial \delta} X_N[\beta_0(Z_{it}) - \beta_0(Z_{js})] \cdot ||\hat{\delta} - \delta|| \\
R_{2,it} &= h^{1/2}(NT)^{-1/2} X_N^\top K_N \frac{\partial Q_N^{-1}}{\partial \delta} u_N[\beta_0(Z_{it}) - \beta_0(Z_{js})] \cdot ||\hat{\delta} - \delta||.
\end{aligned}$$

In addition,

$$\frac{1}{NT} \|\hat{u}\|^2 = O_P(1), \tag{C.9}$$

where $\|\hat{u}\| = \sum_{i=1}^N \sum_{t=1}^T |\hat{u}_{it}|^2$.

Proof of Lemma C.4: By using the same argument as in Lemma C.2, i.e. the uniform consistency of $\hat{\delta}$ over $z \in [0,1]$ established in Theorem 2.1, $R_{1,it}$ and $R_{2,it}$ are $o_p(1)$ uniformly over $z \in [0,1]$ and are therefore negligible. As the results, we simply write $\hat{u}_{it} = \hat{u}_{1,it} + \hat{u}_{2,it}$, where

$$\begin{aligned}
\hat{u}_{1,it} &= h^{1/2}(NT)^{-1/2} \sum_{j=1}^N \sum_{s=1}^T \ddot{X}_{js}^\top \ddot{X}_{js}[\beta_0(Z_{it}) - \beta_0(Z_{js})] K_h(Z_{it} - Z_{js}) \\
\hat{u}_{2,it} &= h^{1/2}(NT)^{-1/2} \sum_{j=1}^N \sum_{s=1}^T \ddot{X}_{js}^\top \ddot{u}_{js} K_h(Z_{it} - Z_{js}).
\end{aligned}$$

We intend to show that

$$\begin{aligned}\frac{1}{NT}E\|\hat{u}\|^2 &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}E|\hat{u}_{it}|^2 \\ &\leq \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}E\left\{|\hat{u}_{1,it}|^2 + |\hat{u}_{2,it}|^2 + 2|\hat{u}_{1,it}|\cdot|\hat{u}_{2,it}|\right\} = O(1). \qquad \text{(C.10)}\end{aligned}$$

Firstly, we start by writing

$$\hat{u}_{1,it}^2 = \frac{h}{NT}\sum_{j\neq\imath}\sum_{s\neq\tau}q_{js,\imath\tau,it} + \frac{h}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}r_{js,it} = |\hat{u}_{1,it}|^2$$

$$E|\hat{u}_{1,it}|^2 = \frac{h}{NT}\sum_{j\neq\imath}\sum_{s\neq\tau}Eq_{js,\imath\tau,it} + \frac{h}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}Er_{js,it}, \qquad \text{(C.11)}$$

where

$$\begin{aligned}q_{js,\imath\tau,it} &= [\beta_0(Z_{it})-\beta_0(Z_{js})]^\top\ddot{X}_{js}^\top\ddot{X}_{js}\ddot{X}_{\imath\tau}^\top\ddot{X}_{\imath\tau}[\beta_0(Z_{it})-\beta_0(Z_{\imath\tau})]K_h(Z_{it}-Z_{js})K_h(Z_{it}-Z_{\imath\tau}) \\ r_{js,it} &= [\beta_0(Z_{it})-\beta_0(Z_{js})]^\top\ddot{X}_{js}^\top\ddot{X}_{js}\ddot{X}_{js}^\top\ddot{X}_{js}[\beta_0(Z_{it})-\beta_0(Z_{js})]K_h^2(Z_{it}-Z_{js}).\end{aligned}$$

We consider firstly $Eq_{js,\imath\tau,it}$. Observe that $q_{js,\imath\tau,it} = q_{js,\imath\tau,it,1} + q_{js,\imath\tau,it,2}$, where

$$\begin{aligned}q_{js,\imath\tau,it,1} &= \{\beta_0'(Z_{it})\}^\top\ddot{X}_{js}^\top\ddot{X}_{js}\ddot{X}_{\imath\tau}^\top\ddot{X}_{\imath\tau}\beta_0'(Z_{it})(Z_{js}-Z_{it})(Z_{\imath\tau}-Z_{it})K_h(Z_{it}-Z_{js})K_h(Z_{it}-Z_{\imath\tau}) \\ q_{js,\imath\tau,it,2} &= C\ddot{X}_{js}^\top\ddot{X}_{js}\ddot{X}_{\imath\tau}^\top\ddot{X}_{\imath\tau}(Z_{js}-Z_{it})^2(Z_{\imath\tau}-Z_{it})^2K_h(Z_{it}-Z_{js})K_h(Z_{it}-Z_{\imath\tau}).\end{aligned}$$

In addition, $Eq_{js,\imath\tau,it} = Eq_{js,\imath\tau,it,1} + Eq_{js,\imath\tau,it,2}$. Regarding the first term,

$$\begin{aligned}Eq_{js,\imath\tau,it,1} &= E\{E[q_{js,\imath\tau,it,1}|Z_{js}=Z_{it},Z_{\imath\tau}=Z_{it}]\} \\ &= \int\{\beta_0'(Z_{it})\}^\top\Omega(Z_{js})\Omega(Z_{\imath\tau})\beta_0'(Z_{it})(Z_{js}-Z_{it})(Z_{\imath\tau}-Z_{it}) \\ &\qquad K_h(Z_{it}-Z_{js})K_h(Z_{it}-Z_{\imath\tau})f(Z_{it})f(Z_{js})f(Z_{\imath\tau})dZ_{it}dZ_{js}dZ_{\imath\tau} \\ &= h^2\int\{\beta_0'(Z_{it})\}^\top\Omega(Z_{js})\Omega(Z_{\imath\tau})\beta_0'(Z_{it})v_1v_2K(v_1)K(v_2)dv_1dv_2f(Z_{js})f(Z_{\imath\tau})f(Z_{it})dZ_{it} \\ &= h^2\int\{\beta_0'(Z_{it})\}^\top\Omega(Z_{js})f(Z_{js})\Omega(Z_{\imath\tau})f(Z_{\imath\tau})\beta_0'(Z_{it})f(Z_{it})dZ_{it}\int\int v_1v_2K(v_1)K(v_2)dv_1dv_2 \\ &= O(h^2)o(1),\end{aligned}$$

where $\Omega(Z_{js}) = E[\ddot{X}_{js}^\top\ddot{X}_{js}|Z_{js}=Z_{it}]$ and $\Omega(Z_{\imath\tau}) = E[\ddot{X}_{\imath\tau}^\top\ddot{X}_{\imath\tau}|Z_{\imath\tau}=Z_{it}]$, by which the third and forth equality are obtained based on $Z_{js} = Z_{it} + v_1h$ and $Z_{js} = Z_{\imath\tau} + v_2h$, and

$$\begin{aligned}&\int\{\beta_0'(Z_{it})\}^\top\Omega(Z_{js})f(Z_{js})\Omega(Z_{\imath\tau})f(Z_{\imath\tau})\beta_0'(Z_{it})f(Z_{it})dZ_{it} \\ &\quad = E\{\beta_0'(Z_{it})\ddot{X}_{it}^\top\ddot{X}_{it}\ddot{X}_{it}^\top\ddot{X}_{it}\beta_0'(Z_{it})\} = O(1),\end{aligned}$$

respectively. Regarding the second term, $Eq_{js,\imath\tau,it,2} = O(h^4)$ uniformly over all pairs (i,t), $i = 1,\ldots,N$ and $t = 1,\ldots,T$, that is

$$\begin{aligned}Eq_{js,\imath\tau,it,2} &\leq CE\left\{|\ddot{X}_{js}^\top\ddot{X}_{js}\ddot{X}_{\imath\tau}^\top\ddot{X}_{\imath\tau}|\left|(Z_{js}-Z_{it})^2(Z_{\imath\tau}-Z_{it})^2K_h(Z_{it}-Z_{js})K_h(Z_{\imath\tau}-Z_{js})\right|\right\} \\ &= O(h^4)\end{aligned}$$

by using the similar argument for $Eq_{js,\imath\tau,it,1}$. Hence, $Eq_{js,\imath\tau,it} = O(h^4)$. Regarding $Er_{js,it}$, in the same spirit as the above, we can show that

$$Er_{js,it} = O(h^3) \tag{C.12}$$

uniformly over all pairs (i,t), $i = 1,\ldots,N$ and $t = 1,\ldots,T$. Hence, by applying these results to (C.11), we obtain

$$\begin{aligned} E\|\hat{u}_{1,it}\|^2 &= (NT)^{-1}h(NT)\{(NT)-1\}O(h^4) + (NT)^{-1}h(NT)O(h^3) \\ &= O((NT)h^5) - O(h^5) + O(h^4) \;=\; O(1). \end{aligned} \tag{C.13}$$

Secondly,

$$\hat{u}_{2,it}^2 = \frac{h}{NT}\left\{\sum_{j=1}^{N}\sum_{s=1}^{T}\ddot{X}_{js}^{\top}\ddot{u}_{js}K_h(Z_{it}-Z_{js})\right\}^2,$$

so that

$$\begin{aligned} E|\hat{u}_{2,it}|^2 &= \frac{h}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}E\left\{\|\ddot{X}_{js}^{\top}\ddot{u}_{js}\|^2K_h^2(Z_{it}-Z_{js})\right\} \\ &= h^{-1}(NT)^{-1}\{(NT)-1\}E\left\{\|\ddot{X}_{11}^{\top}\ddot{u}_{11}\|^2K^2\left(\frac{Z_{22}-Z_{11}}{h}\right)\right\} \\ &\quad + h^{-1}(NT)^{-1}K^2(0)E\left\{\|\ddot{X}_{11}^{\top}\ddot{u}_{11}\|^2\right\}. \end{aligned} \tag{C.14}$$

Observe that

$$\begin{aligned} &E\left\{\|\ddot{X}_{11}^{\top}\ddot{u}_{11}\|^2K^2\left(\frac{Z_{22}-Z_{11}}{h}\right)\right\} \\ &= E\left\{E\left[\|\ddot{X}_{11}^{\top}\ddot{u}_{11}\|^2K^2\left(\frac{Z_{22}-Z_{11}}{h}\right)|Z_{11}=Z_{22}\right]\right\} \\ &= \int \varrho^2(Z_{11})K^2\left(\frac{Z_{22}-Z_{11}}{h}\right)f(Z_{11})f(Z_{22})dZ_{22}dZ_{11} \quad \text{(by using } Z_{22}=Z_{11}+vh) \\ &= h\int \varrho^2(Z_{11})K^2(v)f(Z_{22})f(Z_{11})dZ_{11}dv \\ &\le Ch\int K^2(v)dv\int \varrho^2(Z_{11})f(Z_{22})f(Z_{11})dZ_{11} \;=\; Ch\mathcal{K}E\|\ddot{X}_{11}^{\top}\ddot{u}_{11}\|^2, \end{aligned}$$

where $\varrho^2(Z_{11}) = E\left\{\|\ddot{X}_{11}^{\top}\ddot{u}_{11}\|^2|Z_{11}=Z_{22}\right\}$. Such a result leads to

$$\begin{aligned} E\|\hat{u}_{2,it}\|^2 &\le (NT)^{-1}\{(NT)-1\}C\mathcal{K}E\left\{\|\ddot{X}_{11}^{\top}u_{11}\|^2\right\} + h^{-1}(NT)^{-1}K^2(0)E\left\{\|\ddot{X}_{11}^{\top}\ddot{u}_{11}\|^2\right\} \\ &= O(1). \end{aligned} \tag{C.15}$$

Finally, applications of (C.13) and (C.15) in (C.10) complete the proof.

Lemma C.5. *Let Assumptions A to D hold and $h \propto (NT)^{-1/5}$. Then*

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|\hat{\beta}(Z_{it})-\beta_0(Z_{it})\right\|^2 = O_P\left\{(NT)^{-4/5}\right\}. \tag{C.16}$$

Proof of Lemma C.5: In the spirit of Fan and Li (2001), it suffices to show that for any small probability $\epsilon > 0$ we can always find a constant $C > 0$ such that

$$\lim_{NT\to\infty}\inf \; P\left[\inf_{(NT)^{-1}\|m\|_F^2=C} Q(B_0+\{(NT)h\}^{-1/2}m) > Q(B_0)\right] = 1-\epsilon, \tag{C.17}$$

where m is as defined in Appendix A. To do so requires observing firstly that

$$\begin{aligned} &h(NT)^{-1}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{N}\sum_{s=1}^{T}\left\{\hat{y}_{js}-\hat{X}_{js}\left[\beta_0(Z_{it})+\{(NT)h\}^{-1/2}m_{it}\right]\right\}^2 K_h(Z_{it}-Z_{js}) \\ &\quad - h(NT)^{-1}\textstyle\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{N}\sum_{s=1}^{T}\left\{\hat{y}_{js}-\hat{X}_{js}\beta_0(Z_{it})\right\}^2 K_h(Z_{it}-Z_{js}). \end{aligned} \tag{C.18}$$

Furthermore, let $\alpha \equiv \{(NT)h\}^{-1/2}$ and observe that

$$\begin{aligned} &\left\{\hat{y}_{js}-\hat{X}_{js}\left[\beta_0(Z_{it})+\alpha m_{it}\right]\right\}^\top\left\{\hat{y}_{js}-\hat{X}_{js}\left[\beta_0(Z_{it})+\alpha m_{it}\right]\right\} \\ &\quad -\left\{\hat{y}_{js}-\hat{X}_{js}\beta_0(Z_{it})\right\}^\top\left\{\hat{y}_{js}-\hat{X}_{js}\beta_0(Z_{it})\right\} \\ &= \left\{\ddot{y}_{js}-\ddot{X}_{js}\left[\beta_0(Z_{it})+\alpha m_{it}\right]\right\}^\top\left\{\ddot{y}_{js}-\ddot{X}_{js}\left[\beta_0(Z_{it})+\alpha m_{it}\right]\right\} \\ &\quad -\left\{\ddot{y}_{js}-\ddot{X}_{js}\beta_0(Z_{it})\right\}^\top\left\{\ddot{y}_{js}-\ddot{X}_{js}\beta_0(Z_{it})\right\} + R_3. \end{aligned} \tag{C.19}$$

In this regard,

$$\begin{aligned} R_3 \;=\; & -\alpha y_N^\top\frac{\partial Q_N^{-1}}{\partial\delta}K_N X_N||\hat{\delta}-\delta|| + \alpha\beta_0(Z_{it})^\top X_N^\top\frac{\partial Q_N^{-1}}{\partial\delta}K_N X_N||\hat{\delta}-\delta|| \\ & + \alpha m_{it}^\top X_N^\top\frac{\partial Q_N^{-1}}{\partial\delta}K_N y_N||\hat{\delta}-\delta|| - \alpha m_{it}^\top X_N^\top\frac{\partial Q_N^{-1}}{\partial\delta}K_N X_N\beta_0(Z_{it})||\hat{\delta}-\delta|| \\ & + \alpha m_{it}^\top X_N^\top\frac{\partial Q_N^{-1}}{\partial\delta}K_N X_N m_{it}\alpha||\hat{\delta}-\delta|| \;=\; O_P\{(NT)^{-4/5}\} \end{aligned}$$

by using Theorem 2.1. Hence, the first two terms of (C.19) are the leading terms. A slight rewriting of these terms gives

$$\begin{aligned} &\left\{\ddot{y}_{js}-\ddot{X}_{js}\left[\beta_0(Z_{it})+\alpha m_{it}\right]\right\}^\top\left\{\ddot{y}_{js}-\ddot{X}_{js}\left[\beta_0(Z_{it})+\alpha m_{it}\right]\right\} \\ &\quad -\left\{\ddot{y}_{js}-\ddot{X}_{js}\beta_0(Z_{it})\right\}^\top\left\{\ddot{y}_{js}-\ddot{X}_{js}\beta_0(Z_{it})\right\} \\ &= -2\alpha m_{it}^\top\ddot{X}_{js}^\top\ddot{X}_{js}[\beta_0(Z_{js})-\beta_0(Z_{it})] \;-\; 2\alpha m_{it}^\top\ddot{X}_{js}^\top\ddot{u}_{js} \;+\; \alpha m_{it}^\top\ddot{X}_{js}^\top\ddot{X}_{js}m_{it}\alpha. \end{aligned}$$

This suggests writing

$$
\begin{aligned}
R_4 &= h(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{N}\sum_{s=1}^{T}\left\{\ddot{y}_{js}-\ddot{X}_{js}\left[\beta_0(Z_{it})+\{(NT)h\}^{-1/2}m_{it}\right]\right\}^2 K_h(Z_{it}-Z_{js}) \\
&\quad - h(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{N}\sum_{s=1}^{T}\left\{\ddot{y}_{js}-\ddot{X}_{js}\beta_0(Z_{it})\right\}^2 K_h(Z_{it}-Z_{js}). \\
&= -2h(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{N}\sum_{s=1}^{T}\{(NT)h\}^{-1/2}m_{it}^{\top}\ddot{X}_{js}^{\top}\ddot{X}_{js}[\beta_0(Z_{js})-\beta_0(Z_{it})]K_h(Z_{it}-Z_{js}) \\
&\quad - 2h(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{N}\sum_{s=1}^{T}\{(NT)h\}^{-1/2}m_{it}^{\top}\ddot{X}_{js}^{\top}\ddot{u}_{js}K_h(Z_{it}-Z_{js}) \\
&\quad + h(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{N}\sum_{s=1}^{T}\{(NT)h\}^{-1/2}m_{it}^{\top}\ddot{X}_{js}^{\top}\ddot{X}_{js}m_{it}\{(NT)h\}^{-1/2}K_h(Z_{it}-Z_{js}) \\
&= (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}m_{it}^{\top}\tilde{\Sigma}(Z_{it})m_{it} - 2(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}m_{it}^{\top}\tilde{u}_{it} \qquad \text{(C.20)}
\end{aligned}
$$

where

$$
\begin{aligned}
\tilde{\Sigma}(Z_{it}) &= (NT)^{-1}\sum_{j=1}^{N}\sum_{s=1}^{T}\ddot{X}_{js}^{\top}\ddot{X}_{js}K_h(Z_{it}-Z_{js}) \\
\tilde{u}_{it} &= h^{1/2}(NT)^{-1/2}\sum_{j=1}^{N}\sum_{s=1}^{T}\ddot{X}_{js}^{\top}\{\ddot{X}_{js}[\beta_0(Z_{js})-\beta_0(Z_{it})]+\ddot{u}_{js}\}K_h(Z_{it}-Z_{js}).
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
2(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}m_{it}^{\top}\hat{u}_{it} &\le 2(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\|m_{it}^{\top}\|\|\tilde{u}_{it}\| \\
(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}m_{it}^{\top}\tilde{\Sigma}(Z_{it})m_{it} &\ge (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\gamma}_{it}^{\min}\|m_{it}\|^2,
\end{aligned}
$$

where $\tilde{\gamma}_{it}^{\min}$ denote the smallest eigenvalue of $\tilde{\Sigma}(Z_{it})$. As the results,

$$
\begin{aligned}
R_4 &\ge (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}m_{it}^{\top}\tilde{\Sigma}(Z_{it})m_{it} - 2(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\|m_{it}^{\top}\|\|\tilde{u}_{it}\| \\
&\ge (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\gamma}_{it}^{\min}\|m_{it}\|^2 - 2(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\|m_{it}^{\top}\|\|\tilde{u}_{it}\| \\
&\ge (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\gamma}_{it}^{\min}\|m_{it}\|^2 - 2\left\{(NT)^{-1}\|m\|^2\right\}^{1/2}\left\{(NT)^{-1}\|\tilde{u}\|^2\right\}^{1/2} \ \ge\ R_5,
\end{aligned}
$$

where the third inequality is due to the Cauchy–Schwarz inequality and

$$
\begin{aligned}
R_5 &= \tilde{\gamma}^{\min}\cdot(NT)^{-1}\|m\|^2 - 2\left\{(NT)^{-1}\|m\|^2\right\}^{1/2}\left\{(NT)^{-1}\|\tilde{u}\|^2\right\}^{1/2} \\
&= \tilde{\gamma}^{\min}C^2 - 2C\left\{(NT)^{-1}\|\tilde{u}\|^2\right\}^{1/2}, \qquad \text{(C.21)} \\
C &= \{(NT)^{-1}\|m\|^2\}^{1/2}.
\end{aligned}
$$

In this regard, it is the case that $\gamma_0^{\min} > 0$. Since $(NT)^{-1}\|\tilde{u}\|^2 = O_P(1)$ using Lemma C.4, positivity of R_5 is ensured for sufficiently large C. Finally, note that $h(NT)^{-1} = (NT)^{-4/5}$.

Lemma C.6. *Let Assumptions A to E hold. Then*

$$(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|\hat{\beta}_{\lambda,a}(Z_{it}) - \beta_0(Z_{it})\right\|^2 = O_P\left\{(NT)^{-4/5}\right\}, \tag{C.22}$$

where $\hat{\beta}_{\lambda,a}(Z_{it}) = \{\hat{\beta}_{\lambda,1}(Z_{it}), \ldots, \hat{\beta}_{\lambda,D_0}(Z_{it})\}^\top$.

Proof of Lemma C.6: Given Lemma C.5 and its proof, we may begin by noting that

$$\begin{aligned} & h(NT)^{-1}\left\{Q_\lambda(B_0 + \{(NT)h\}^{-1/2}m) - Q_\lambda(B_0)\right\} \\ & \qquad = \; R_4 \; + \; O_P\left\{(NT)^{-4/5}\right\} \; + \; h(NT)^{-1}\sum_{d=1}^{D}\lambda_d\left[\|b_{0d} + \{(NT)h\}^{-1/2}v_d\| - \|b_{0d}\|\right]. \end{aligned}$$

To prove Lemma C.6 only requires showing that

$$\begin{aligned} R_6 \; &= \; h(NT)^{-1}\sum_{d=1}^{D}\lambda_d\left[\|b_{0d} + \{(NT)h\}^{-1/2}v_d\| - \|b_{0d}\|\right] \\ &= \; h(NT)^{-1}\left\{\sum_{d=1}^{D_0}\lambda_d\left[\|b_{0d} + \{(NT)h\}^{-1/2}v_d\| - \|b_{0d}\|\right] + \sum_{d=D_0+1}^{D}\lambda_d\|\{(NT)^{-1/2}v_d\|\right\} \to 0. \end{aligned}$$

With regard to the first term,

$$\begin{aligned} R_7 \; &= \; h(NT)^{-1}\sum_{d=1}^{D_0}\lambda_d\left[\|b_{0d} + \{(NT)h\}^{-1/2}v_d\| - \|b_{0d}\|\right] \; \leq \; h^{1/2}(NT)^{-3/2}\sum_{d=1}^{D_0}\lambda_d\|v_d\| \\ &\leq \; h^{1/2}(NT)^{-3/2}a_{NT}\sum_{d=1}^{D_0}\|v_d\| \; \leq \; h^{1/2}(NT)^{-1}a_{NT}\left\{(NT)^{-1}\sum_{d=1}^{D_0}\|v_d\|^2\right\}^{1/2} \\ &= \; \left\{h^{1/2}(NT)^{-1}a_{NT}\right\}C, \end{aligned}$$

which converges to zero since $\left\{h^{1/2}(NT)^{-1}a_{NT}\right\} \propto (NT)^{11/10}a_N \to 0$ under the conditions of the lemma. Furthermore, the second term can be similarly worked out. That is

$$\begin{aligned} R_8 \; &= \; \sum_{d=D_0+1}^{D}\lambda_d\|\{(NT)^{-1/2}v_d\| \; \leq \; h^{1/2}(NT)^{-3/2}\sum_{d=D_0+1}^{D}\lambda_d\|v_d\| \\ &\leq \; h^{1/2}(NT)^{-3/2}a_{NT}\sum_{d=D_0+1}^{D}\|v_d\| \; \leq \; h^{1/2}(NT)^{-1}a_{NT}\left\{(NT)^{-1}\sum_{d=D_0+1}^{D}\|v_d\|^2\right\}^{1/2} \\ &= \; \left\{h^{1/2}(NT)^{-1}a_{NT}\right\}C \to 0. \end{aligned}$$

Lemma C.7. *Let Assumptions A to E hold. Also, let*

$$\alpha_{23t} = \sum_{j=1}^{N}\sum_{s=1}^{T}\ddot{X}_{k,js}\ddot{X}_{js}[\beta_0(Z_{it}) - \hat{\beta}_\lambda(Z_{it})]K_h(Z_{it} - Z_{js}).$$

Then,

$$\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\alpha_{23t}\right)^{1/2} \leq O_P\{(NT)h^{-1/2}\}. \tag{C.23}$$

Proof of Lemma C.7: Observe that

$$\alpha_{23t}^2 \leq \|\beta_0(Z_{it}) - \hat{\beta}_\lambda(Z_{it})\|^2 \left\|\sum_{j=1}^{N}\sum_{s=1}^{T}\ddot{X}_{k,js}\ddot{X}_{js}K_h(Z_{it}-Z_{js})\right\|^2,$$

so that

$$\begin{aligned}\sum_{i=1}^{N}\sum_{t=1}^{T}\alpha_{23t}^2 &\leq (NT)\left\{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\|\beta_0(Z_{it}) - \hat{\beta}_\lambda(Z_{it})\|^2\right\} \\ &\qquad \times(NT)^2\left\|\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\ddot{X}_{k,js}\ddot{X}_{js}K_h(Z_{it}-Z_{js})\right\|^2. \\ &= (NT)O_P\left\{(NTh)^{-1}\right\}O_P\{(NT)^2\},\end{aligned}$$

where the final result is based on Lemma C.5 and since

$$\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\ddot{X}_{k,js}\ddot{X}_{js}K_h(Z_{it}-Z_{js}) = O_P(1),$$

which follows Lemma C.1.

Lemma C.8. *Let Assumptions A to E hold. Then*

$$P\left(\left\|\hat{b}_{\lambda,d}\right\| = 0\right) \to 1 \quad \textit{for any} \quad D_0 < d \leq D.$$

Proof of Lemma C.8: Consider the D-th column of $\hat{B}_\lambda$, i.e. $\hat{b}_{\lambda,D}$. Such solution must satisfy

$$0 = \frac{\partial Q_\lambda(B)}{\partial b_D}\Big|_{B=\hat{B}_\lambda} = \alpha_1 + \alpha_2, \tag{C.24}$$

where

$$Q_\lambda(B) = \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{N}\sum_{s=1}^{T}\left\{\hat{y}_{js} - \hat{X}_{js}\beta(Z_{it})\right\}^2 K_h(Z_{it}-Z_{js}) + \sum_{d=1}^{D}\lambda_d\|b_d\|$$

as in (2.22), $\alpha_1 = \lambda_D(b_D/\|b_D\|)$ and α_2 is also a $NT\times 1$ vector in which

$$\alpha_{2,it} = -2\sum_{j=1}^{N}\sum_{s=1}^{T}\hat{X}_{D,js}\{\hat{y}_{js} - \hat{X}_{js}\hat{\beta}_\lambda(Z_{it})\}K_h(Z_{it}-Z_{js}). \tag{C.25}$$

Let us first consider α_2. Observe that

$$\{\hat{y}_{js} - \hat{X}_{js}\hat{\beta}_\lambda(Z_{it})\} = \{(\hat{X}_{js}\beta_0(Z_{js}) + \hat{u}_{js}) - \hat{X}_{js}\beta_0(Z_{it}) + \hat{X}_{js}\beta_0(Z_{it}) - \hat{X}_{js}\hat{\beta}_\lambda(Z_{it})\}.$$

This leads to

$$\alpha_{2,it} = \alpha_{21,it} + \alpha_{22,it} + \alpha_{23,it},$$

where

$$\alpha_{21,it} = \sum_{j=1}^{N}\sum_{s=1}^{T} \ddot{X}_{D,js}\ddot{u}_{js}K_h(Z_{it}-Z_{js}) + R_{21,it},$$

$$\alpha_{22,it} = \sum_{j=1}^{N}\sum_{s=1}^{T} \ddot{X}_{D,js}\ddot{X}_{js}(\beta_0(Z_{js})-\beta_0(Z_{it}))K_h(Z_{it}-Z_{js}) + R_{22,it},$$

$$\alpha_{23,it} = \sum_{j=1}^{N}\sum_{s=1}^{T} \ddot{X}_{D,js}\ddot{X}_{js}(\beta_0(Z_{it})-\hat{\beta}_\lambda(Z_{it}))K_h(Z_{it}-Z_{js}) + R_{23,it}.$$

Since $R_{21,it}, R_{22,it}$ and $R_{23,it}$ are respectively defined in the same manner as $R_{1,it}$, $R_{2,it}$ and R_3, they are asymptotically negligible.

We are able to obtain the following results by omitting these negligible terms. Firstly, (C.14) suggests that $\alpha_{21,it}^2 = O\left(h^{-2}\right)$, so that

$$\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\alpha_{21,it}^2\right)^{1/2} = O_P\left(\{(NT)^2h^{-2}\}^{1/2}\right) = O_P\left((NT)h^{-1}\right).$$

Similarly, (C.12) implies $\alpha_{22,it}^2 = O(h)$, so that

$$\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\alpha_{22,it}^2\right)^{1/2} = O_P\left((NT)h^{-1/2}\right).$$

Moreover, Lemmas C.1 and C.6 point to

$$\begin{aligned}
\alpha_{23,it} &\leq \left\{\sum_{i=1}^{N}\sum_{t=1}^{T}\|\beta_0(Z_{it})-\hat{\beta}_\lambda(Z_{it})\|^2\right\}^{1/2}\left\{\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|\sum_{j=1}^{N}\sum_{s=1}^{T}\ddot{X}_{D,js}\ddot{X}_{js}K_h(Z_{it}-Z_{js})\right\|^2\right\}^{1/2} \\
&= \left[(NT)O_P((NT)^{-4/5})\cdot O_P((NT)^2)\right]^{1/2} = O_P\left((NT)h^{-1/2}\right).
\end{aligned}$$

These implies collectively that

$$\begin{aligned}
\|\alpha_{2,it}\| &= \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\alpha_{2,it}^2\right)^{1/2} \\
&\leq \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\alpha_{21,it}^2\right)^{1/2} + \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\alpha_{21,it}^2\right)^{1/2} + \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\alpha_{23,it}^2\right)^{1/2} \\
&+ \left\{2\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\alpha_{21,it}\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\alpha_{22,it}\right)\right\}^{1/2} + \left\{2\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\alpha_{21,it}\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\alpha_{23,it}\right)\right\}^{1/2} \\
&+ \left\{2\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\alpha_{22,it}\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\alpha_{23,it}\right)\right\}^{1/2} = O_P\left((NT)h^{-1/2}\right). \qquad \text{(C.26)}
\end{aligned}$$

Finally, it is the case that

$$\|\alpha_2\| = \sqrt{\alpha_{2,11}^2+\cdots+\alpha_{2,NT}^2} \leq \sqrt{\alpha_{2,11}^2}+\cdots+\sqrt{\alpha_{2,NT}^2} = O_P\left((NT)h^{-1/2}\right) \qquad \text{(C.27)}$$

since $\sqrt{\alpha_{2,it}^2} = \sqrt{\alpha_{2,it,1}^2+\cdots+\alpha_{2,it,D}^2} = \|\alpha_{2,it}\|$.

Next, we consider $\alpha_1 = \lambda_D(b_D/\|b_D\|)$. Since $b_d = (\beta_d(Z_{11}), \ldots, \beta_d(Z_{NT}))^\top$ and $\|b_d\| = \sqrt{\sum_{i=1}^N \sum_{t=1}^T \beta_d^2(Z_{it})}$, it is the case that

$$\begin{aligned}
\|\alpha_1\| &= \|\lambda_d(b_d/\|b_d\|)\| \\
&= \lambda_d \sqrt{\sum_{i=1}^N \sum_{t=1}^T \left\{ \beta_d^2(Z_{it}) / \sum_{i=1}^N \sum_{t=1}^T \beta_d^2(Z_{it}) \right\}} \\
&\geq b_n \propto O_P\left((NT)h^{-1/2}\right).
\end{aligned}$$

As the results, $P(\|\alpha_1\| > \|\alpha_2\|) \to 1$ as $(NT) \to \infty$ as such the condition in (C.24) cannot hold. This suggest that $\hat{b}_{\lambda,d}$ must be located at the place where the objective function is not differentiable, i.e. the origin. This leads to $P(\hat{b}_{\lambda,d} = 0) \to 1$ and completes the proof.

D. Proof of the results in Section 2.2

D.1. Proof of Theorem 2.2(b):

The proof of Theorem 2.2(b) follows immediately from Lemmas C.1 to C.5.

D.2. Proof of Theorem 2.3:

Firstly, by using the results of Theorem 2.1, the Taylor expansion of $\hat{\mathbb{Q}}_N^{-1}$ can be expressed as

$$\hat{\mathbb{Q}}_N^{-1} = \mathbb{Q}_{0N}^{-1} + \frac{\partial \mathbb{Q}_{0N}^{-1}}{\partial \rho}(\hat{\rho} - \rho_0) + \frac{\partial \mathbb{Q}_N^{-1}}{\partial \phi}(\hat{\phi} - \phi_0) + o_P((NT)^{-1/2}). \tag{D.1}$$

Accordingly, $\hat{\beta}(z)$ formula in (2.14) can be re-written as

$$\begin{aligned}
\hat{\beta}(z) \;=\; \beta_0(z) + \Bigg\{ &\left(\sum_{j=1}^N \sum_{s=1}^T \ddot{X}_{0js}^\top \ddot{X}_{0js} K_h(Z_{js} - z) + \dot{D}_{js}(z) O_P((NT)^{-1/2}) \right)^{-1} \\
&\times \left(\sum_{j=1}^N \sum_{s=1}^T \{ \hat{X}_{js}^\top \hat{X}_{js}(\beta_0(Z_{js}) - \beta_0(z)) + \hat{X}_{js}^\top \hat{u}_{js}^0 \} K_h(Z_{js} - z) \right) \Bigg\},
\end{aligned} \tag{D.2}$$

where $\hat{u}_{js}^0 = \hat{y}_{js} - \hat{X}_{js}\beta_0(Z_{js})$. In this regard, $\dot{D}_{js}(z) = \dot{D}_{js,\rho}(z) + \dot{D}_{js,\phi}(z)$ in which

$$\begin{aligned}
\dot{D}_{js,\rho}(z) &= \frac{1}{NT} \sum_{j=1}^N \sum_{s=1}^T \dot{X}_{0js,\rho}^\top \dot{X}_{0js,\rho} K_h(Z_{js} - z), \\
\dot{D}_{js,\phi}(z) &= \frac{1}{NT} \sum_{j=1}^N \sum_{s=1}^T \dot{X}_{0js,\phi}^\top \dot{X}_{0js,\phi} K_h(Z_{js} - z),
\end{aligned}$$

where $\dot{X}_{0N}^\top \dot{X}_{0N} = X_N^\top \frac{\partial \mathbb{Q}_{0N}^{-1}}{\partial \delta} X_N$ and $\dot{X}_{0N}^\top \dot{u}_N = X_N^\top \frac{\partial \mathbb{Q}_{0N}^{-1}}{\partial \delta} u_N$. In addition, we obtain by using the Triangular inequality and following standard nonparametric analysis

$$E||\dot{D}_{js}(z)||_F \leq E||\dot{D}_{js,\rho}(z)||_F + E||\dot{D}_{js,\phi}(z)||_F = O(1). \tag{D.3}$$

By using and (D.3), we re-write (D.2) as follows

$$\begin{aligned}\hat{\beta}(z) &= \beta_0(z) + \left\{ \left[\sum_{j=1}^{N}\sum_{s=1}^{T} \ddot{X}_{0js}^{\top}\ddot{X}_{0js}K_h(Z_{js}-z) + o_P(1) \right]^{-1} \right. \\ &\quad \left. \times \left(\sum_{j=1}^{N}\sum_{s=1}^{T} \ddot{X}_{0js}^{\top}\{\ddot{X}_{0js}(\beta_0(Z_{js})-\beta_0(z)) + \ddot{u}_{js}^{0}\}K_h(Z_{js}-z) \right) \right\} \\ &\quad + \{\mathbb{R}_{11,N}(z) + \mathbb{R}_{12,N}(z)\}O_P((NT)^{-1/2}), \end{aligned} \tag{D.4}$$

where $\ddot{u}_{js}^{0} = \ddot{y}_{0js} - \ddot{X}_{0js}\beta_0(Z_{js})$, and

$$\begin{aligned}\mathbb{R}_{11,N}(z) &= \left(\sum_{j=1}^{N}\sum_{s=1}^{T} \ddot{X}_{0js}^{\top}\ddot{X}_{0js}K_h(Z_{js}-z) \right)^{-1} \sum_{j=1}^{N}\sum_{s=1}^{T} \dot{X}_{0js}^{\top}\dot{u}_{js}^{0}K_h(Z_{js}-z), \\ \mathbb{R}_{12,N}(z) &= \left(\sum_{j=1}^{N}\sum_{s=1}^{T} \ddot{X}_{0js}^{\top}\ddot{X}_{0js}K_h(Z_{js}-z) \right)^{-1} \sum_{j=1}^{N}\sum_{s=1}^{T} \left\{ \dot{X}_{0js}^{\top}\dot{X}_{0js}(\beta_0(Z_{js})-\beta_0(z)) \right\} K_h(Z_{js}-z). \end{aligned}$$

We firstly consider the denominator of the above terms. In this regard,

$$\begin{aligned} E\left(\frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T} \ddot{X}_{0js}^{\top}\ddot{X}_{0js}K_h(Z_{js}-z) \right) &= f(z)E(\ddot{X}_{0js}^{\top}\ddot{X}_{0jz}|Z_{js}=z) \\ &\geq \inf_{||z||\leq c_N} f(z)E(\ddot{X}_{0js}^{\top}\ddot{X}_{0js}|z). \end{aligned} \tag{D.5}$$

By denoting $\inf_{||z||\leq c_N} f_z(z)E(\ddot{X}_{0js}^{\top}\ddot{X}_{0js}|Z_{js}=z) = \mathscr{D}^*$, we have

$$E||\mathbb{R}_{11}(z)|| \leq \mathscr{D}^{*-1}E\left|\left| \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T} \dot{X}_{0js}^{\top}\dot{u}_{js}^{0}K_h(Z_{js}-z) \right|\right| = O((NTh)^{-1/2}) \tag{D.6}$$

by using Triangular and Cauchy-Schwartz inequalities, and the standard nonparametric analysis. Similarly,

$$\begin{aligned} E||\mathbb{R}_{12}(z)|| &\leq \mathscr{D}^{*-1}E\left|\left| \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T} \left\{ \dot{X}_{0js}^{\top}\dot{X}_{0js}(\beta_0(Z_{js})-\beta_0(z)) \right\} K_h(Z_{js}-z) \right|\right| \\ &= O((NT)^{-1/2}h^{1/2}) + O(h^2). \end{aligned}$$

These suggest that we re-write (D.4) as follows

$$\begin{aligned}\hat{\beta}(z) &= \beta_0(z) + \left\{ \left(\sum_{j=1}^{N}\sum_{s=1}^{T} \ddot{X}_{0js}^{\top}\ddot{X}_{0js}K_h(Z_{js}-z) \right)^{-1} \right. \\ &\quad \left. \times \sum_{j=1}^{N}\sum_{s=1}^{T} \ddot{X}_{0js}^{\top}\{\ddot{X}_{0js}(\beta_0(Z_{js})-\beta_0(z)) + \ddot{u}_{0js}\}K_h(Z_{js}-z) \right\} + o_P(1). \end{aligned}$$

The rest of the proofs is straightforward as shown in the standard varying-coefficient literature. Let us present the denominator case as follows

$$E\left\{\frac{1}{NTh}\sum_{j=1}^{N}\sum_{s=1}^{T}\ddot{X}_{0js}^{\top}\ddot{X}_{0js}K\left(\frac{Z_{js}-z}{h}\right)\right\}=\mathscr{D}(z)+O(h^2),$$

where $\mathscr{D}(z)=f_z(z)E(\ddot{X}_{0js}^{\top}\ddot{X}_{0js}|z)$ and

$$\mathrm{Var}\left\{\frac{1}{NTh}\sum_{j=1}^{N}\sum_{s=1}^{T}\ddot{X}_{0js}^{\top}\ddot{X}_{0js}K\left(\frac{Z_{js}-z}{h}\right)\right\}=O((NTh)^{-1}),$$

and

$$E\left\{\frac{1}{NTh}\sum_{j=1}^{N}\sum_{s=1}^{T}\ddot{X}_{0js}^{\top}\ddot{X}_{0js}(\beta_0(Z_{js})-\beta_0(z))K\left(\frac{Z_{js}-z}{h}\right)\right\}=\mathscr{K}_2h^2\mathscr{B}.$$

Finally,

$$E\left\{\frac{1}{NTh}\sum_{j=1}^{N}\sum_{s=1}^{T}\ddot{X}_{0js}^{\top}\ddot{u}_{0js}K\left(\frac{Z_{js}-z}{h}\right)\right\}=0$$

and

$$\mathrm{Var}\left\{\frac{1}{\sqrt{NTh}}\sum_{j=1}^{N}\sum_{s=1}^{T}\ddot{X}_{0js}^{\top}\ddot{u}_{0js}K\left(\frac{Z_{js}-z}{h}\right)\right\} \quad = \quad V(z).$$

Therefore,

$$\sqrt{NTh}\left(\hat{\beta}(z)-\beta_0(z)-Bias\right)\to_D N(0,\Sigma),$$

where $Bias=\mathscr{D}^{-1}(z)\mathscr{K}_2h^2\mathscr{B}$ and $\Sigma=\mathscr{D}^{-1}(z)V(z)\mathscr{D}^{-1}(z)$.

E. Proof of results in Section 2.3

The following definitions are useful for providing proof of Theorems 2.4 and 2.5 :

$$\hat{\mathscr{M}}(z) \;=\; \begin{pmatrix}\hat{\mathscr{M}}_{aa}(z) & \hat{\mathscr{M}}_{ab}(z)\\ \hat{\mathscr{M}}_{ba}(z) & \hat{\mathscr{M}}_{bb}(z)\end{pmatrix},\ \hat{\mathscr{N}}(z) \;=\; \begin{pmatrix}\hat{\mathscr{N}}_{a}(z)\\ \hat{\mathscr{N}}_{b}(z)\end{pmatrix},\ \hat{\mathscr{D}} \;=\; \begin{pmatrix}\hat{\mathscr{D}}_{aa} & 0\\ 0 & \hat{\mathscr{D}}_{bb}\end{pmatrix},$$

where $\hat{\mathscr{D}}_{aa}$ and $\hat{\mathscr{D}}_{bb}$ are $(D_0\times D_0)$ and $(D-D_0\times D-D_0)$ diagonal block matrices, respectively, whose diagonal elements are $\lambda_j/\|\hat{b}_j\|$. In addition,

$$\begin{aligned}
\hat{\mathscr{M}}_{aa}(z) &= \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\hat{X}_{a,js}^{\top}\hat{X}_{a,js}K_h(Z_{it}-Z_{js}),\ \hat{\mathscr{M}}_{bb}(z) \;=\; \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\hat{X}_{b,js}^{\top}\hat{X}_{b,js}K_h(Z_{it}-Z_{js}),\\
\hat{\mathscr{M}}_{ab}(z) &= \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\hat{X}_{a,js}^{\top}\hat{X}_{b,js}K_h(Z_{it}-Z_{js})=\hat{\mathscr{M}}_{ba}(z),\\
\hat{\mathscr{N}}_{a}(z) &= \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\hat{X}_{a,js}^{\top}\hat{y}_{js}K_h(Z_{it}-Z_{js}) \text{ and } \hat{\mathscr{N}}_{b}(z) \;=\; \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T}\hat{X}_{b,js}^{\top}\hat{y}_{js}K_h(Z_{it}-Z_{js}).
\end{aligned}$$

Moreover, $\left[\hat{\mathscr{M}}(z)+(NT)^{-1}\hat{\mathscr{D}}\right]^{-1}=\begin{pmatrix}\hat{\Omega}_{aa} & \hat{\Omega}_{ab}\\ \hat{\Omega}_{ba} & \hat{\Omega}_{bb}\end{pmatrix}$, where

$$\begin{aligned}
\hat{\Omega}_{aa} &= \left(\hat{\mathscr{M}}_{aa}(z)+\hat{\mathscr{D}}_{aa}-\hat{\mathscr{M}}_{ab}(z)\left\{\hat{\mathscr{M}}_{bb}+\hat{\mathscr{D}}_{bb}\right\}^{-1}\hat{\mathscr{M}}_{ba}\right)^{-1},\\
\hat{\Omega}_{ab} &= -\left\{\hat{\mathscr{M}}_{aa}(z)+\hat{\mathscr{D}}_{aa}\right\}^{-1}\hat{\mathscr{M}}_{ab}(z)\hat{\Omega}_{aa},\ \hat{\Omega}_{ba} = -\left\{\hat{\mathscr{M}}_{bb}(z)+\hat{\mathscr{D}}_{bb}\right\}^{-1}\hat{\mathscr{M}}_{ba}(z)\hat{\Omega}_{bb},\\
\hat{\Omega}_{bb} &= \left(\hat{\mathscr{M}}_{bb}(z)+\hat{\mathscr{D}}_{bb}-\hat{\mathscr{M}}_{ba}(z)\left\{\hat{\mathscr{M}}_{aa}+\hat{\mathscr{D}}_{aa}\right\}^{-1}\hat{\mathscr{M}}_{ab}\right)^{-1}.
\end{aligned}$$

An example for the use of the above definitions is to rewrite the penalized estimators as

$$\hat{\beta}_{\lambda}(z)=\left[\hat{\mathscr{M}}(z)+(NT)^{-1}\hat{\mathscr{D}}\right]^{-1}\hat{\mathscr{N}}(z). \tag{E.1}$$

E.1. Proof of Theorem 2.4:

The penalized estimators of $\beta_{0,b}(z)=\{\beta_{0,D_0+1}(z),\ldots,\beta_{0,D}(z)\}^{\top}$, i.e. the coefficient vector associated with the irrelevant regressors, can be expressed as

$$\hat{\beta}_{\lambda,b}(z)=\hat{\Omega}_{ba}(z)\hat{\mathscr{N}}_a(z)+\hat{\Omega}_{bb}(z)\hat{\mathscr{N}}_b(z).$$

We note firstly that both $\hat{\mathscr{N}}_a(z)$ and $\hat{\mathscr{N}}_b(z)$ are uniformly bounded in a similar fashion to Lemmas C.1 and C.2. Hence, to prove that $\hat{\beta}_{\lambda,b}(z)\to 0$ as $NT\to\infty$ uniformly on $z\in[0,1]$ only requires showing that every elements of $\hat{\Omega}_{ba}$ and $\hat{\Omega}_{bb}$ converge to zero in the same manner. To this end, we note that the diagonal elements of $\hat{\mathscr{D}}_{bb}$ are $\lambda_d/\|\hat{b}_d\|$ for $(D_0+1)\le d\le D$, and

$$\sup_{z\in[0,1]}\|\hat{\beta}(z)\|\to 0,$$

which is in accordance with Theorem 2.2. Hence,

$$\min\left\|\hat{\mathscr{D}}_{bb}\right\|=\left\|\operatorname{diag}\left\{\frac{b_N}{\left\|\hat{b}_{D_0+1}\right\|},\ldots,\frac{b_N}{\left\|\hat{b}_D\right\|}\right\}\right\|\to\infty \tag{E.2}$$

due to Assumption E1. This completes the proof.

E.2. Proof of Theorem 2.5

The penalized estimator of $\beta_{0,a}(z)=\{\beta_{0,1}(z),\ldots,\beta_{0,D_0}(z)\}^{\top}$, i.e. the coefficient vector associated with the relevant regressors, can be expressed as

$$\hat{\beta}_{\lambda,a}(z)=\hat{\Omega}_{ab}(z)\hat{\mathscr{N}}_b(z)+\hat{\Omega}_{aa}(z)\hat{\mathscr{N}}_a(z),$$

whereas the unpenalized counterpart is

$$\hat{\beta}_a(z)=\hat{\Phi}_{ab}(z)\hat{\mathscr{N}}_b(z)+\hat{\Phi}_{aa}(z)\hat{\mathscr{N}}_a(z),$$

where

$$\begin{aligned}
\hat{\Phi}_{aa} &= \left(\hat{\mathscr{M}}_{aa}(z)-\hat{\mathscr{M}}_{ab}(z)\left\{\hat{\mathscr{M}}_{bb}\right\}^{-1}\hat{\mathscr{M}}_{ba}\right)^{-1},\ \hat{\Phi}_{ab} = -\left\{\hat{\mathscr{M}}_{aa}(z)\right\}^{-1}\hat{\mathscr{M}}_{ab}(z)\hat{\Phi}_{aa},\\
\hat{\Phi}_{bb} &= \left(\hat{\mathscr{M}}_{bb}(z)-\hat{\mathscr{M}}_{ba}(z)\left\{\hat{\mathscr{M}}_{aa}\right\}^{-1}\hat{\mathscr{M}}_{ab}\right)^{-1},\ \hat{\Phi}_{ba} = -\left\{\hat{\mathscr{M}}_{bb}(z)\right\}^{-1}\hat{\mathscr{M}}_{ba}(z)\hat{\Phi}_{bb}.
\end{aligned}$$

Hence, the difference between these estimators is

$$\hat{\beta}_{\lambda,a}(z) - \hat{\beta}_a(z) = \{\hat{\Omega}_{ab}(z) - \hat{\Phi}_{ab}(z)\}\hat{\mathcal{N}}_b(z) + \{\hat{\Omega}_{aa}(z) - \hat{\Phi}_{aa}(z)\}\hat{\mathcal{N}}_a(z). \tag{E.3}$$

This implies that proving the theorem requires showing that every elements of $\hat{\Omega}_{ab}$ and $\hat{\Omega}_{aa}$ converge to zero. In the other words, the convergence of $\max_{z\in(0,1)} \|\hat{\beta}_{\lambda,a}(z) - \hat{\beta}_a(z)\|$ depends entirely on that of

$$\max \|\hat{\mathcal{D}}_{aa}\| = \left\| \text{diag} \left\{ \frac{a_N}{\|\hat{b}_1\|}, \ldots, \frac{a_N}{\|\hat{b}_{D_0}\|} \right\} \right\|. \tag{E.4}$$

Hence, the claimed result is obtained immediately by noting Assumption E1 and Theorem 2.2.

E.3. Proof of Theorem 2.6

An arbitrary model $\mathcal{S}_\lambda$ may be correctly-fitted, under-fitted or over-fitted. Accordingly, we can create three mutually exclusive sets, $\mathbb{R}_0 = \{\lambda \in \mathbb{R}^D : \mathcal{S}_\lambda = \mathcal{S}_T\}$, $\mathbb{R}_- = \{\lambda \in \mathbb{R}^D : \mathcal{S}_\lambda \not\supset \mathcal{S}_T\}$ and $\mathbb{R}_+ = \{\lambda \in \mathbb{R}^D : \mathcal{S}_\lambda \supset \mathcal{S}_T, \mathcal{S}_\lambda \neq \mathcal{T}\}$, which belong to correctly-fitted, under-fitted and over-fitted, respectively. Also, let λ_{NT} denote a reference tuning parameter that satisfies the conditions of Assumption E1. This can be obtained, for example, by setting $\lambda_0 = (NT)^{-3/2}\log(NT)$. Moreover, we can deduce from the proof of Theorem 2.1

$$R\hat{S}S_F \to_P R\tilde{S}S_F \quad \text{whereas} \quad R\tilde{S}S_F \to_P \sigma^2_{v,0}, \tag{E.5}$$

and from Lemmas C.2 and C.3

$$\hat{\Sigma}(z) \to_P \tilde{\Sigma}(z) \quad \text{whereas} \quad \tilde{\Sigma}(z) \to_P f(z)\Omega(z). \tag{E.6}$$

Furthermore,

$$\|\hat{\beta}(Z_{it}) - \hat{\beta}_\lambda(Z_{it})\|^2 \geq \left\| \|\hat{\beta}(Z_{it})\|^2 - \|\hat{\beta}_\lambda(Z_{it})\|^2 \right\|^2$$

due to the reverse triangle inequality.

We consider first the case of under-fitting, i.e. $\lambda \in \mathbb{R}_-$. Recall and rewrite

$$\begin{aligned} R\hat{S}S_\lambda &= (NT)^{-2} \sum_{i=1}^N \sum_{t=1}^T \sum_{j=1}^N \sum_{s=1}^T \{\hat{y}_{js} - \hat{X}_{js}\hat{\beta}_\lambda(Z_{it})\}^2 K_h(Z_{it} - Z_{js}) \\ &= R\hat{S}S_F + \hat{R}_\lambda, \end{aligned} \tag{E.7}$$

where $\hat{\beta}_\lambda(z) = \{\hat{\beta}_{\lambda,1}(z), \ldots, \hat{\beta}_{\lambda,D}(z)\}^\top$, and

$$\begin{aligned} R\hat{S}S_F &= (NT)^{-2} \sum_{i=1}^N \sum_{t=1}^T \sum_{j=1}^N \sum_{s=1}^T \{\hat{y}_{js} - \hat{X}_{js}\hat{\beta}(Z_{it})\}^2 K_h(Z_{it} - Z_{js}), \\ \hat{R}_\lambda &= (NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T \{\hat{\beta}(Z_{it}) - \hat{\beta}_\lambda(Z_{it})\}^\top \hat{\Sigma}(Z_{it}) \{\hat{\beta}(Z_{it}) - \hat{\beta}_\lambda(Z_{it})\}. \end{aligned}$$

In this regard, these suggest that (i) $R\hat{S}S_F \to_P \sigma_{v,0}^2$, (ii) for $\hat{R}_\lambda$ we concentrate directly on

$$\begin{aligned}
&(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\{\hat{\beta}(Z_{it})-\hat{\beta}_\lambda(Z_{it})\}^\top\tilde{\Sigma}(Z_{it})\{\hat{\beta}(Z_{it})-\hat{\beta}_\lambda(Z_{it})\}\\
&\geq\ \tilde{\gamma}^{\min}\left\{(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\|\hat{\beta}(Z_{it})-\hat{\beta}_\lambda(Z_{it})\|^2\right\}\\
&\geq\ \tilde{\gamma}^{\min}\left\{(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|\|\hat{\beta}(Z_{it})\|^2-\|\hat{\beta}_\lambda(Z_{it})\|^2\right\|^2\right\}\\
&\geq\ \tilde{\gamma}^{\min}\left\{(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\|\hat{\beta}_1(Z_{it})\|^2\right\}\\
&\to_P\quad \gamma_0^{\min}E\{\beta_{0,1}^2(Z_{it})\},
\end{aligned}\tag{E.8}$$

where the third inequality is obtained by assuming that the first coefficient is selected as being irrelevant, i.e. $\hat{\beta}_{\lambda,1}(Z_{it})=0$. Therefore,

$$R\hat{S}S_\lambda=\sigma_v^2+\gamma_0^{\min}E\{\beta_{0,1}^2(Z_{it})\}\tag{E.9}$$

in probability. We may similarly define

$$R\hat{S}S_{\lambda_{NT}}=R\hat{S}S_F+\hat{R}_{\lambda_{NT}},$$

where

$$\hat{R}_{\lambda_{NT}}=(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\{\hat{\beta}(Z_{it})-\hat{\beta}_{\lambda_{NT}}(Z_{it})\}^\top\hat{\Sigma}(Z_{it})\{\hat{\beta}(Z_{it})-\hat{\beta}_{\lambda_{NT}}(Z_{it})\},$$

but focus instead on

$$\begin{aligned}
&(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\{\hat{\beta}(Z_{it})-\hat{\beta}_{\lambda_{NT}}(Z_{it})\}^\top\tilde{\Sigma}(Z_{it})\{\hat{\beta}(Z_{it})-\hat{\beta}_{\lambda_{NT}}(Z_{it})\}\\
&\leq\ \tilde{\gamma}^{\max}\left\{(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\|\hat{\beta}(Z_{it})-\hat{\beta}_{\lambda_{NT}}(Z_{it})\|^2\right\}\\
&\leq\ \tilde{\gamma}^{\max}\left\{(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\|\hat{\beta}(Z_{it})-\beta_0(Z_{it})\|^2\right\}\\
&\qquad+\tilde{\gamma}^{\max}\left\{(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\|\beta_0(Z_{it})-\hat{\beta}_{\lambda_{NT}}(Z_{it})\|^2\right\},
\end{aligned}$$

which converges to zero in probability according to Lemmas C.5 and C.6. Therefore,

$$\begin{aligned}
\inf_{\lambda\in\mathbb{R}_-}\{BIC_\lambda-BIC_{\lambda_{NT}}\}\ &=\ \inf_{\lambda\in\mathbb{R}_-}(RSS_\lambda-RSS_{\lambda_{NT}})+(df_\lambda-df_{\lambda_{NT}})\left\{\frac{\log\{(NT)h\}}{(NT)^{4/5}}\right\}\\
&>\ 0\ \text{ in probability.}
\end{aligned}\tag{E.10}$$

Now, we consider the case where $\lambda\in\mathbb{R}_+$, so that an arbitrary model $\mathcal{S}_\lambda$ is over-fitted. In addition, let $\hat{B}_{S_\lambda}=(\hat{\beta}_{S_\lambda}(Z_11),\ldots,\hat{\beta}_{S_\lambda}(Z_11))^\top$ denote an unpenalised estimate, which belongs to

the over-fitted model S_λ. Similarly to (E.7), we may define

$$\begin{aligned} R\hat{S}S_{S_\lambda} &= (NT)^{-2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{N}\sum_{s=1}^{T}\{\hat{y}_{js}-\hat{X}_{js}\hat{\beta}_{S_\lambda}(Z_{it})\}^2K_h(Z_{it}-Z_{js}) \\ &= R\hat{S}S_F+\hat{R}_{S_\lambda}, \end{aligned}$$

where

$$\hat{R}_{S_\lambda}=(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\{\hat{\beta}(Z_{it})-\hat{\beta}_{S_\lambda}(Z_{it})\}^\top\hat{\Sigma}(Z_{it})\{\hat{\beta}(Z_{it})-\hat{\beta}_{S_\lambda}(Z_{it})\}.$$

In this regard, we deduce from the over-fitting nature and the consistency of the unpenalised estimator, i.e. Lemma C.5, that $\hat{R}_\lambda \geq \hat{R}_{S_\lambda}$, so that $RSS_\lambda \geq RSS_{S_\lambda}$. In addition,

$$\begin{aligned} \log RSS_\lambda &\geq \log RSS_{S_\lambda} \\ \log RSS_\lambda - \log RSS_F &\geq \log RSS_{S_\lambda} - \log RSS_F. \end{aligned}$$

Moreover,

$$\begin{aligned} &\log RSS_{S_\lambda} - \log RSS_F \\ &\quad = \log\left\{1+\frac{1}{(NT)\hat{\sigma}_v^2}\sum_{i=1}^{n}\sum_{t=1}^{T}\{\hat{\beta}(Z_{it})-\hat{\beta}_{S_\lambda(Z_{it})}\}^\top\hat{\Sigma}(Z_{it})\{\hat{\beta}(Z_{it})-\hat{\beta}_{S_\lambda}(Z_{it})\}\right\}. \end{aligned}$$

In accordance with (E.6), we can concentrate directly on

$$\begin{aligned} &\log\left\{1+\frac{1}{(NT)\hat{\sigma}_v^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\{\hat{\beta}(Z_{it})-\hat{\beta}_{S_\lambda(Z_{it})}\}^\top\tilde{\Sigma}(Z_{it})\{\hat{\beta}(Z_{it})-\hat{\beta}_{S_\lambda}(Z_{it})\}\right\} \\ &\quad \geq -\frac{1}{\hat{\sigma}_v^2}\cdot\frac{1}{(NT)}\sum_{i=1}^{N}\sum_{t=1}^{T}\{\hat{\beta}(Z_{it})-\hat{\beta}_{S_\lambda}(Z_{it})\}^\top\tilde{\Sigma}(Z_{it})\{\hat{\beta}(Z_{it})-\hat{\beta}_{S_\lambda}(Z_{it})\} \\ &\quad \geq -\frac{1}{\hat{\sigma}_v^2}\left(\tilde{\gamma}^{\max}\left\{(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\|\hat{\beta}(Z_{it})-\beta_0(Z_{it})\|^2\right\}\right. \\ &\qquad\qquad \left. +\,\tilde{\gamma}^{\max}\left\{(NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\|\beta_0(Z_{it})-\hat{\beta}_{S_\lambda}(Z_{it})\|^2\right\}\right) \\ &\quad = -|O_P\{(NT)^{-4/5}\}|. \end{aligned} \tag{E.11}$$

The first inequality is since $\log(1+b)\geq-\log(1+b)\geq -b$ for $b\geq 0$. The second inequality is due to the triangle inequality, whereas the third inequality is in accordance with Lemma C.5. It can be similarly shown that

$$\log RSS_{\lambda_{NT}} - \log RSS_F \geq -|O_P\{(NT)^{-4/5}\}|,. \tag{E.12}$$

Hence, we are able to deduce from (E.11) and (E.12)

$$\inf_{\lambda\in\mathbb{R}_+}(RSS_\lambda - RSS_{\lambda_{NT}}) \geq -|O_P\{(NT)^{-4/5}\}|. \tag{E.13}$$

Moreover,

$$\begin{aligned}\inf_{\lambda\in\mathbb{R}_-}\{BIC_\lambda - BIC_{\lambda_{NT}}\} &= \inf_{\lambda\in\mathbb{R}_-}(RSS_\lambda - RSS_{\lambda_{NT}}) + (df_\lambda - df_{\lambda_{NT}})\left\{\frac{\log\{(NT)h\}}{(NT)^{4/5}}\right\} \\ &> 0 \text{ in probability.} \end{aligned} \tag{E.14}$$

In order to obtain the result in (E.14), observe that (i) $P(df_{\lambda_{NT}} = D_0) \to 1$ which is an implication of Assumption E1, (ii) since $\lambda \in \mathbb{R}_+$ and S_λ is an over-fitted model, we must have $P(df_\lambda \geq D_0 + 1) \to 1$, and (iii) under Assumption B2 we have $\log((NT)h) \propto \log(NT) \to \infty$ and so $df_\lambda - df_{\lambda_{NT}} \geq 1$ with probability tending to one.

F. *Proof of Corollary 2.1*

Similarly to Theorem 2.2(b), Corollary 2.1 follows immediately from Lemmas C.1.

G. *Proof of results in Section 2.5*

Proof of Corollary 2.2:

The proof of Corollary 2.2 relies heavily on Theorem 2.4. We commence by noting that the penalized estimators under the local quadratic approximation, i.e. $\hat{\beta}_\lambda^{(m+1)}(z)$, differs from $\hat{\beta}_\lambda(z)$ only by replacing the diagonal matrix $\hat{\mathfrak{D}}$ with

$$\hat{\mathfrak{D}}^{(m)} = \text{diag}\left\{\frac{\lambda_1}{\|\hat{b}_{\lambda,1}^{(m)}\|}, \ldots, \frac{\lambda_K}{\|\hat{b}_{\lambda,D}^{(m)}\|}\right\}. \tag{G.1}$$

We know that $\|\hat{b}_{\lambda,d}^{(m)}\| = \|\hat{b}_{\lambda,d}\|$ as $m \to \infty$, for every $1 \leq d \leq D$, by using the results in Hunter and Li (2005). By Theorem 2.4, we also know that $P\left(\left\|\hat{b}_{\lambda,d}\right\| = 0\right) \to 1$ for any $(D_0+1) \leq d \leq D$ and $P\left(\left\|\hat{b}_{\lambda,d}\right\| \neq 0\right) \to 1$ for $1 \leq d \leq D_0$. Hence, it must be the case that $\|\hat{b}_{\lambda,d}^{(m)}\|$ converges to 0 for every $D_0 < d \leq D$, while converging to a positive number for every $d \leq D_0$.

Next we partition $\hat{\mathfrak{D}}^{(m)}$ into sub-matrices $\hat{\mathfrak{D}}_{aa}^{(m+1)}$, i.e. upper $D_0 \times D_0$ diagonal sub-matrix, and $\hat{\mathfrak{D}}_{bb}^{(m+1)}$, i.e. lower $(D - D_0) \times (D - D_0)$ diagonal sub-matrix. By the definitions in (2.25) and (G.1), it must be the case that all the diagonal elements of $\hat{\mathfrak{D}}_{aa}^{(m+1)}$ converge to some finite number, whereas those of $\hat{\mathfrak{D}}_{bb}^{(m+1)}$ diverge to infinity when $m \to \infty$.

Finally, since these conclusions are similar to those drawn for $\hat{\mathfrak{D}}_{aa}$ and $\hat{\mathfrak{D}}_{bb}$, the rest of the proof closely follows that of Theorem 2.4.

Proof of Corollary 2.3:

The proof of Corollary 2.3 follows that of Corollary 2.2 and Theorem 2.5.

References

Aziz, F., McCrone, P., Boyle, S., Knapp, M., 2003. Financing mental health services in london: central funding and local expenditure .

Baltagi, B.H., Bresson, G., Pirotte, A., 2012. Forecasting with spatial panel data. Computational Statistics & Data Analysis 56, 3381–3397.

Baltagi, B.H., Song, S.H., Koh, W., 2003. Testing panel data regression models with spatial error correlation. Journal of econometrics 117, 123–150.

Baltagi, B.H., et al., 2008. Econometric analysis of panel data. volume 4. Springer.

Cai, Z., Fan, J., Li, R., 2000. Efficient estimation and inferences for varying-coefficient models. Journal of the American Statistical Association 95, 888–902.

Cliff, A.D., 1973. Spatial autocorrelation. Technical Report.

Cliff, A.D., Ord, J.K., 1981. Spatial processes: models & applications. Taylor & Francis.

Fan, J., Farmen, M., Gijbels, I., 1998. Local maximum likelihood estimation and inference. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60, 591–608.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association 96, 1348–1360.

Fan, J., Zhang, J.T., 2000a. Two-step estimation of functional linear models with applications to longitudinal data. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62, 303–322.

Fan, J., Zhang, W., 1999. Statistical estimation in varying coefficient models. The annals of Statistics 27, 1491–1518.

Fan, J., Zhang, W., 2000b. Simultaneous confidence bands and hypothesis testing in varying-coefficient models. Scandinavian Journal of Statistics 27, 715–731.

Fan, J., Zhang, W., 2008. Statistical methods with varying coefficient models. Statistics and its Interface 1, 179.

Feng, G., Gao, J., Peng, B., Zhang, X., 2017. A varying-coefficient panel data model with fixed effects: theory and an application to us commercial banks. Journal of econometrics 196, 68–82.

Gao, J., Xia, K., Zhu, H., 2020. Heterogeneous panel data models with cross-sectional dependence. Journal of Econometrics 219, 329–353.

Hu, T., Xia, Y., 2012. Adaptive semi-varying coefficient model selection. Statistica Sinica , 575–599.

Hunter, D., Li, R., 2005. Variable selection using mm algorithms. Annals of statistics , 1617.

Kapoor, M., Kelejian, H.H., Prucha, I.R., 2007. Panel data models with spatially correlated error components. Journal of Econometrics 140, 97–130.

Kelejian, H.H., Prucha, I.R., 1999. A generalized moments estimator for the autoregressive parameter in a spatial model. International economic review 40, 509–533.

Lee, L.f., 2004. Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. Econometrica 72, 1899–1925.

Lee, L.f., Yu, J., 2010. Estimation of spatial autoregressive panel data models with fixed effects. Journal of econometrics 154, 165–185.

Liu, S.F., Yang, Z., 2015. Asymptotic distribution and finite sample bias correction of qml estimators for spatial error dependence model. Econometrics 3, 376–411.

Magnus, J.R., Muris, C., 2010. Specification of variance matrices for panel data models. Econometric Theory 26, 301–310.

Manski, C.F., 1993. Identification problems in the social sciences. Sociological Methodology , 1–56.

McCrone, P., Jacobson, B., 2004. Indicators of mental health activity in london: Adjusting for sociodemographic need. London: London Development Centre for Mental Health .

Moscone, F., Knapp, M., Tosetti, E., 2007. Mental health expenditure in england: a spatial panel approach. Journal of Health Economics 26, 842–864.

Okui, R., Takahide, Y., 2018. Kernel estimation for panel data with heterogeneous dynamics. Working Paper .

Robinson, P.M., 2011. Asymptotic theory for nonparametric regression with spatial data. Journal of Econometrics 165, 5–19.

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

Su, L., Yang, Z., 2015. Qml estimation of dynamic panel data models with spatial errors. Journal of Econometrics 185, 230–258.

Sun, Y., Carroll, R.J., Li, D., 2009. Semiparametric estimation of fixed-effects panel data varying coefficient models, in: Nonparametric econometric methods. Emerald Group Publishing Limited.

Wang, H., Leng, C., 2007. Unified lasso estimation by least squares approximation. Journal of the American Statistical Association 102, 1039–1048.

Wang, H., Xia, Y., 2009. Shrinkage estimation of the varying coefficient model. Journal of the American Statistical Association 104, 747–757.

Xia, Y., Zhang, W., Tong, H., 2004. Efficient estimation for semivarying-coefficient models. Biometrika 91, 661–681.

Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association , 1418–1429.

Zou, H., Li, R., 2007. One-step sparse estimates in nonconcave penalized likelihood models. Annals of statistics 36, 1509.