

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

**Working Paper**

## Varying coefficient model with correlated error components and application to disparities between mental health service by councils in England

Cardiff Economics Working Papers, No. E2022/1

**Provided in Cooperation with:**

Cardiff Business School, Cardiff University

*Suggested Citation:* Wongs-arta, Pipat; Kim, Namhyun; Xia, Yingcun; Moscone, Francesco (2022) : Varying coefficient model with correlated error components and application to disparities between mental health service by councils in England, Cardiff Economics Working Papers, No. E2022/1, Cardiff University, Cardiff Business School, Cardiff

This Version is available at:

<https://hdl.handle.net/10419/261229>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

# 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](http://business.cardiff.ac.uk)

This working paper is produced for discussion purpose only. These working papers are expected to be published in due course, in revised form, and should not be quoted or cited without the author's written permission.

Cardiff Economics Working Papers are available online from:

<http://econpapers.repec.org/paper/cdfwpaper/> and

[business.cardiff.ac.uk/research/academic-sections/economics/working-papers](http://business.cardiff.ac.uk/research/academic-sections/economics/working-papers)

Enquiries: [EconWP@cardiff.ac.uk](mailto:EconWP@cardiff.ac.uk)

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

4 Pipat Wongs-art<sup>a,\*</sup>, Namhyun Kim<sup>b</sup>, Yingcun Xia<sup>c</sup>, Francesco Moscone<sup>d</sup>

5 <sup>a</sup>Cardiff Business School, Cardiff University, United Kingdom

6 <sup>b</sup>Exeter Business School, University of Exeter, United Kingdom

7 <sup>c</sup>Department of Statistics&Applied Probability, National University of Singapore, Singapore

8 <sup>d</sup>Brunel Business School, Brunel University London, United Kingdom

---

9 **Abstract**

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

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

23 *JEL:* C14; C51; C52; G12; G17

---

\*Corresponding authors

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

24 **1. Introduction**

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

34 Spatial models for panel data are important tools in economics, regional science and  
35 geography in analysing a wide range of empirical issues. By far the most widely used  
36 spatial models are variants of that originated by Cliff (1973) (see also Cliff and Ord  
37 (1981)). Spatial panel data models with spatial autoregressive (SAR) disturbances are  
38 considered in Baltagi et al. (2003), Kapoor et al. (2007), Liu and Yang (2015), and Su and  
39 Yang (2015) among others. Gao et al. (2020) consider an alternative assumption on the  
40 dependence that is shown to be closely related to the SAR. Moreover, varying coefficient  
41 models are a useful extension of classical linear models and have been the main focus of  
42 many methodological studies in the literature (see e.g. Fan and Zhang (1999), Cai et al.  
43 (2000) and Xia et al. (2004)). Varying coefficient panel data models have also attracted  
44 considerable attention in the past two decades. For example, Sun et al. (2009) propose a  
45 panel data varying coefficient model by imposing a widely-used identification restriction  
46 such that the sum of the fixed effects is zero, whereas Rodriguez-Poo and Soberon (2014)  
47 propose to use the first difference to remove the fixed effects. Furthermore, Feng et al.  
48 (2017) consider varying coefficient panel data models where all covariates are categorical.

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

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

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

99 The rest of the paper is organised as follows. In Section 2.1, we propose a varying-  
100 coefficient panel data model that includes spatially correlated error components. In Sec-  
101 tions 2.2, we illustrate the model’s estimation procedure and establish its relevant asymp-  
102 totic properties, whereas in Section 2.3 we introduce the alternative variable selection  
103 procedure. In Section 3, we conduct extensive simulation exercises in order to exam-  
104 ine the finite sample performance and robustness of our proposed procedures, while we  
105 present the application of our model and methods to analyse municipal disparities in men-  
106 tal health service spending by councils in England in Section 4. Finally, Section 5 presents  
107 conclusions and further discussion. Technical proofs are provided in the Appendix.

## 108 2. Varying coefficient model with spatially correlated error components

109 We begin by introducing the model specification and basic assumptions.

### 110 2.1. Model specification

Let  $y_{it} \in \mathbb{R}^1$  be a response of interest,  $X_{it} = (X_{it,1}, \dots, X_{it,D}) \in \mathbb{R}^D$  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  $i = 1, \dots, N$  and  $t = 1, \dots, T$ , we assume that we observe  $y_{it}$  generated by

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

where  $\beta_0(z) = \{\beta_{1,0}(z), \dots, \beta_{D,0}(z)\}^\top \in \mathbb{R}^D$  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

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

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

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

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

$$y_N = (B_0 \circ X_N)e_D + u_N, \quad (2.2)$$

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

$$u_N = (I_T \otimes \rho_0 W_N)u_N + \varepsilon_N, \quad (2.3)$$

where  $\otimes$  signifies the Kronecker product,  $W_N$  is an  $N \times N$  spatial weights matrix which is nonstochastic,  $\rho_0$  is a scalar auto-regressive parameter and  $\varepsilon_N$  is an  $NT \times 1$  vector of innovations. Moreover,  $\varepsilon_N$  follows a classical one-way error component model (see e.g. Baltagi et al. (2008))

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

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

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

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

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

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

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

125 **Assumption A5.** The unit specific error components  $\{\mu_i\}$ ,  $i = 1, 2, \dots, N$  are i.i.d.  
 126 across  $i$  with zero mean, variance  $\sigma_{\mu,0}^2$  and  $E(|\mu_i|^{4+\varsigma}) < \infty$  for some  $\varsigma = 2\varrho$ , where  
 127  $0 < \varrho \leq 2$ .

128 **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

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

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

134 number of neighbours, or in the other words, if  $W_N$  is sparse. Alternatively, they may  
 135 be satisfied when  $W_N$  is not sparse if its elements decline with a distance measure that  
 136 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

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

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

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

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

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

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

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

## 143 2.2. Estimation procedure

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

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

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

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

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

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



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

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

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

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

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  $K_N = \text{diag}\{K_h(Z_{11} - z), \dots, K_h(Z_{N1} - z), K_h(Z_{12} - z), \dots, K_h(Z_{NT} - z)\}$ , 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

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

149 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  
 150 estimator of  $\delta_0 = (\phi_0, \rho_0)^\top$ . To establish its asymptotic properties requires a number of  
 151 additional conditions. We introduce first some conditions of standard nature on the kernel  
 152 function and the bandwidth.

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

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

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

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

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

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

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

168 Assumptions B1 and B2 are primitive and used regularly in nonparametric studies (see  
169 e.g. Okui and Takahide (2018)), whereas Assumption C1 ensures that the observed index  
170 values are sufficiently dense on the support. This implies that maximal distance between  
171 two consecutive index variables is only of the order  $O_P\left(\frac{\log NT}{NT}\right)$ . In addition, Assumption  
172 C2 replaces assumptions in spatial panel regression models. For example, it is required  
173 in those studies that elements of  $X_N$  are uniformly bounded constants for all  $N$  and  
174  $\lim_{NT \rightarrow \infty} (NT)^{-1} X_N^\top Q_N^{-1} X_N$  exists and is nonsingular for all  $\delta \in \Delta$ , (e.g. Assumption 6 of  
175 Lee (2004)). For an arbitrary index value  $z \in [0, 1]$ , let  $z^*$  be its nearest neighbor among the  
176 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  
177 a smoothness condition on the functional coefficient, which implies that  $\|\beta_0(z) - \beta_0(z^*)\| =$   
178  $O_P\left(\frac{\log NT}{NT}\right)$  (see e.g. Xia et al. (2004)). This is an order substantially smaller than the  
179 optimal nonparametric convergence rate, which is  $(NT)^{-2/5}$ .

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

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

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

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)

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

184 where  $\bar{\sigma}_v^2 = (1/NT)\sigma_{v0}^2 \text{TR}[\mathbb{Q}_{0N} \bar{\mathbb{Q}}_N^\top K_N \bar{\mathbb{Q}}_N] \left[ \sum_{j=1}^N \sum_{s=1}^T K_h(Z_{js} - z) \right]^{-1}$ .

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

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

185 *for any  $\delta$ , where  $\bar{D}_\epsilon(\delta_0)$  is the complement of  $\epsilon$ -neighbourhood of  $\delta_0$ . Then,  $\delta_0$  is uniquely*  
 186 *identifiable and  $\hat{\delta} = \delta_0 + O_P((NT)^{-1/2})$  as  $N \rightarrow \infty$ .*

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

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

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

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

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

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

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

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

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

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

195 *where  $\text{Bias} = \mathfrak{D}^{-1}(z) \mathfrak{K}_2 h^2 \mathfrak{B}$  and  $\Sigma = \mathfrak{D}^{-1}(z) V(z) \mathfrak{D}^{-1}(z)$ .*

196 *2.3. SAREC-KLASSO method*

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

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

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

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

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

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

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

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

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

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

where  $\theta = (\sigma_v^2, \phi, \rho)^\top \in \Theta$  for which  $\Theta$  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

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

214 enables the formulation of the concentrated log-likelihood

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

215 Now, let  $\hat{\delta}_\lambda$  denotes quasi-maximum likelihood estimates of  $\delta_0$ . Then, the penalized  
216 estimate of  $B_0 = \{\beta_0(Z_{11}), \dots, \beta_0(Z_{N1}), \beta_0(Z_{12}), \dots, \beta_0(Z_{NT})\}^\top$  is

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

in which

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

where  $\lambda = (\lambda_1, \dots, \lambda_D)^\top \in \mathbb{R}^D$  are the tuning parameters,  $b_d \in \mathbb{R}^{TN \times 1}$  is the  $d$ th 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

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

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

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

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

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

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

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

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

**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

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

227 Theorem 2.4 suggests that the true model can be consistently selected as long as the  
 228 tuning parameters satisfy the conditions listed in Assumption E1. Moreover, since it is  
 229 associated with  $D_0$ ,  $\hat{\beta}_a(z)$  can be viewed as the oracle estimators. Theorem 2.5 suggests  
 230 that the asymptotically optimal nonparametric convergence rate can be achieved as long  
 231 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.  $\lambda_1, \dots, \lambda_D$ , 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

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

and

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

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

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

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

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

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

236 Let  $\hat{B}_{\hat{\lambda}}$  denote a penalised estimator in (2.21), which corresponds to  $\hat{\lambda} = \text{argmin}_\lambda \text{BIC}_\lambda$ ,  
 237 and  $\mathcal{S}_{\hat{\lambda}}$  represent the model identified by  $\hat{B}_{\hat{\lambda}}$ . Theorem 2.6 below states that the turning  
 238 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*

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

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

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

#### 246 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

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

247 where  $C_d$  is a constant.

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

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

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

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

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

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

249 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,  
 250 the null hypothesis is rejected when the test statistic exceeds the asymptotic critical value  
 251  $c_\alpha = -\log(-0.5 \log \alpha)$ .

## 252 2.5. Local quadratic approximation of the penalty function

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

$$\begin{aligned} & \sum_{i=1}^N \sum_{t=1}^T \sum_{j=1}^N \sum_{s=1}^T \left\{ \hat{y}_{js} - \hat{X}_{j's}^\top \beta(Z_{it}) \right\}^2 K_h(Z_{it} - Z_{js}) + \sum_{d=1}^D \lambda_d \frac{\|b_d\|^2}{\|\hat{b}_{\lambda,d}^{(m)}\|} \\ &= \sum_{i=1}^N \sum_{t=1}^T \left\{ \sum_{j=1}^N \sum_{s=1}^T \left\{ \hat{y}_{js} - \hat{X}_{j's}^\top \beta(Z_{it}) \right\}^2 K_h(Z_{it} - Z_{js}) + \sum_{d=1}^D \lambda_d \frac{\beta_d^2(Z_{it})}{\|\hat{b}_{\lambda,d}^{(m)}\|} \right\}, \end{aligned} \quad (2.34)$$



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

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

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

260 We next study the dynamics of  $\hat{\beta}_\lambda^{(m+1)}(z)$  as  $m \rightarrow \infty$ . The results are presented as  
 261 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*

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

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

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

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

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

### 264 3. Simulation studies

265 In this section, we present a set of simulation exercises that examine the finite-sample  
 266 performance of the procedures introduced in the previous sections. These are (3.1) spatial  
 267 estimation, which involves estimation of  $\delta_0 = (\phi_0, \rho_0)^\top$ , using the concentrated likelihood;  
 268 (3.2) nonparametric estimation of coefficient functions  $\beta_0(z) = \{\beta_{0,1}(z), \dots, \beta_{0,D}(z)\}^\top$   
 269 based on the oracle, unpenalised and the penalised estimators; (3.3) variable selection,  
 270 i.e. relevant versus irrelevant variables, based on the SAREC-KLASSO and KLASSO  
 271 methods; and (3.4) hypothesis testing of coefficient constancy, i.e  $H_0 : \beta_{0,d}(z) = C_d$  versus  
 272  $H_1 : \beta_{0,d}(z) \neq C_d$ , where  $1 \leq d \leq \hat{D}$  and  $C_d$  is a constant.

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

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

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

Table 1: Spatial estimation: Model I

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

275 The difference between Models I and II lies in the fact that the former includes zero  
276 constant-coefficients, whereas the latter includes two, i.e. 0.5 and 0.7. Hence, Model II  
277 is an example of the semi-varying coefficient models. We set  $X_{it,1} = 1$ , and generate  
278  $(X_{it,2}, \dots, X_{it,7})^\top$  from multivariate normal distribution by setting  $cov(X_{it,j_1}, X_{it,j_2}) =$   
279  $0.5^{|j_1 - j_2|}$  for any  $2 \leq j_1, j_2 \leq 7$ . We also generate  $Z_{it}$  from uniform distribution  $U[0, 1]$ .  
280 Furthermore, the disturbance follows the SAR and EC processes explained in Section 2.1  
281 with  $\rho_0 = 0.3$  and  $\sigma_{\mu,0}^2 = \sigma_{v,0}^2 = 1$ . Regarding the required spatial weight matrix, we follow  
282 Kelejian and Prucha (1999) and employ matrices that differ in their degree of sparseness.  
283 Particularly, we construct what known in the literature as the “ $P$ -ahead-and- $P$ -behind”  
284 spatial association. For example,  $P = 1$  leads to the “1-ahead-and-1-behind” matrix,  
285 whose  $i$ th row has nonzero elements in positions  $i + 1$  and  $i - 1$ , so that the  $i$ th element  
286 is directly related only to two other elements, namely the ones in front and behind it. We  
287 construct three spatial weight matrices based on  $P = 2$ ,  $P = 5$  or  $P = 8$ , which lead to 4,  
288 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

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

289 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  
 290 alleviate a serious computational burden caused by repeated evaluations of this  $TN \times TN$   
 291 matrix during the optimisation process. Other necessary computational parameters are  
 292 selected as follows. For each simulation repetition, we select the optimal bandwidth based  
 293 on the method of leaving-one-out cross validation within the context of the unpenalised  
 294 estimation since the asymptotic theory for such selection is already well developed in the  
 295 literature (see e.g. Lee and Yu (2010)). The bandwidth selected in this step is also used in  
 296 the penalized estimation. In addition, the optimal shrinkage parameter is selected based  
 297 on the BIC criterion defined in (2.6), whereas the total number of iteration of the iterative  
 298 algorithm in Section 2.5 is set at 15. In the simulation exercises that follow, a total of  
 299 200 repetitions are conducted for each of the model setups. Tables 1 to 3 summarise the  
 300 simulation-results obtained. Below, we discuss a number of important findings.

### 301 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  $\rho_0$  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  $\phi_0 = \sigma_{\mu,0}^2/\sigma_{v,0}^2$ . 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

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

302 where *bias* refers to the difference between the median of the estimates and  $\rho_0$ , *IQ* is the  
 303 inter-quantile range  $c_1 - c_2$  in which  $c_1$  and  $c_2$  are the 0.75 and 0.25 quantiles, respectively.

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

Table 2: Spatial estimation: Model II

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

311 the penalised estimation are able to provide the much more accurate estimates of the  
312 coefficient functions (we will discuss this further below). At  $N = 300$ ,  $\hat{\phi}_\lambda$  performs almost  
313 as well as the oracle-based counterpart. Moreover, an increase in  $P$ , which leads to a  
314 higher number of nonzero elements in a give row of the weighting matrix, renders less  
315 accurate estimation of both the spatial parameter and the variance ratio. However, that  
316 of the former seems to be affected more significantly. Finally, similar results are obtained  
317 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)

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

319 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

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

### 330 3.3. Variable selection

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

339 These results show that the performance of our procedure is not affected by the fact  
340 that Model II contains constant coefficients. Such a finding paves way for identifying  
341 constant coefficients in semivarying coefficient models using the procedure introduced in  
342 Section 2.4. A higher number of nonzero coefficients leads to better finite sample perfor-  
343 mance at smaller  $N$ . Nonetheless, the results for the two models converge when  $N$  increases  
344 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

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

### 349 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:

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

350 Otherwise, the remaining details are as previously specified.

351 Table 5 summarises the simulation results. The table shows percentages of correct  
352 rejections and non-rejections (out of 150 replications) obtained by applying the Fan and  
353 Zhang's (2000) testing procedure *with* and *without* spatial error dependence being ad-  
354 dressed and the random effect being utilised in order to obtain efficiency gain. Before  
355 discussing our findings, it is important to note that, in Model III,  $\beta_{0,1}(z)$  demonstrates a  
356 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	<i>with</i>	<i>without</i>	<i>with</i>	<i>without</i>	<i>with</i>	<i>without</i>
$H_0 : \beta_1(z) = C_1$	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	<i>with</i>	<i>without</i>	<i>with</i>	<i>without</i>	<i>with</i>	<i>without</i>
$H_0 : \beta_1(z) = C_1$	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.

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

#### 363 4. Public mental health expenditure in England

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

Table 6: Our data and its sources

Symbols	Descriptions
<i>vote</i>	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 <a href="http://www.electionscentre.co.uk">www.electionscentre.co.uk</a>
<i>tph</i>	Population-standardised total public health by local authority Source: Reported in the DCLG's Revenue Outturn, Social Care and Public Health data available at <a href="http://www.ons.gov.uk">www.ons.gov.uk</a>
<i>mhs</i>	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 <a href="http://www.ons.gov.uk">www.ons.gov.uk</a>
<i>nuc</i>	Claimants of unemployment-related benefits on Benefits Agency Administrative System Source: Regional labour market Claimant Count by unitary and local authority available at <a href="http://www.ons.gov.uk">www.ons.gov.uk</a>
<i>pmp</i>	Percentage of male population by local authority Source: Estimates of the population for the UK available at <a href="http://www.ons.gov.uk">www.ons.gov.uk</a>
<i>pu14</i>	Percentage of population under 14 year of age Source: Estimates of the population for the UK available at <a href="http://www.ons.gov.uk">www.ons.gov.uk</a>
<i>smr</i>	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 <a href="https://data.gov.uk">https://data.gov.uk</a>
<i>noj</i>	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 <a href="https://data.gov.uk">https://data.gov.uk</a>
<i>plp</i>	Percentage of households headed by lone parent by local authority Source: Estimated number of households by household types, local authorities in England available at <a href="http://www.ons.gov.uk">www.ons.gov.uk</a>
<i>mhp</i>	Median house price paid by local authority Source: Median house prices for administrative geographies available at <a href="http://www.ons.gov.uk">www.ons.gov.uk</a>
<i>mww</i>	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 <a href="http://www.ons.gov.uk">www.ons.gov.uk</a>
<i>psq</i>	Population density defined as population per square kilometre Source: Estimates of the population for the UK available at <a href="http://www.ons.gov.uk">www.ons.gov.uk</a>

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



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

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

391 The study in this section focuses on 151 councils in England, which have social services  
392 responsibility, out of 333 local authorities. However, two local authorities, namely City of  
393 London and Isles of Scilly, are excluded from our analysis due their unusual socio-economic  
394 and demographic characteristics. On the time dimension, we focus on the observation  
395 period between 2016/17 and 2019/20, which reflects our interest on the impact of the  
396 2013 government’s public health reform and the reduction in its spending on the public  
397 health grant during the period. These lead therefore to  $N = 149$  and  $T = 4$ . Below we  
398 begin by first establishing the empirical model.

#### 399 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

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

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

403 *4.1.1. Regressors and covariate*

Let us begin with  $X = (X_1, \dots, X_D)$  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

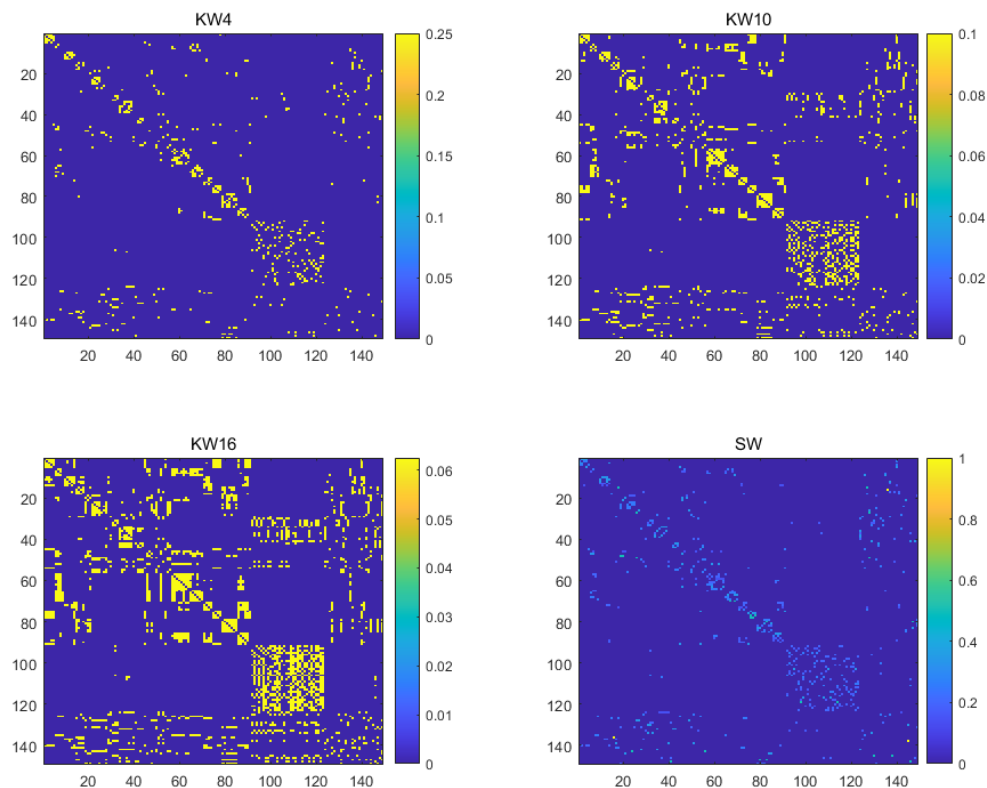
$$mhs = \beta_{0,1}(Z) + X^* \beta_0^*(Z) + u, \quad (4.2)$$

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

415 Moreover, it is important to note that the English local authorities have considerable  
 416 autonomy under the reformed system to (a) allocate resources from central government  
 417 among various local services (e.g. education, housing, leisure, community resources and  
 418 social services), and (b) prioritise particular areas and client groups in line with local  
 419 interpretations of need. Therefore, their actual spending will likely reflect local policies  
 420 and preference rather than the standard spending assessment by the central government.  
 421 Hence, it is implausible to assume, for example, that the percentage of people aged under  
 422 14 (%POPu14) would have the same effect on *mhs* across all the local authorities. In  
 423 the light of this argument, we will focus our study on two strategies. Firstly, we define  
 424 the covariate  $Z$  in order to take into account a political influence. The basis of this idea  
 425 is from a hypothesis that there might be councils that decide (based on their political  
 426 beliefs, for example) to give more weight in terms of resources to the elderly while others  
 427 to the youths. In the practical analysis, the covariate is represented by the percentage  
 428 of voters with right-wing ideology, *vote* hereafter. More specifically, it is the percentage  
 429 of voters that have voted for the Conservative and UK Independence Parties in the local  
 430 government elections. Secondly, the covariate  $Z$  is defined as total public health spending  
 431 by local authority standardized by the total population in each authority, *tph* hereafter.  
 432 The purpose of this is to take into consideration a type of Engel’s law, which might  
 433 be in operation within the context of MHS. To understand this idea more clearly, let  
 434 us recall the Engel’s law in economics which suggests that the poorer a family is, the  
 435 larger the budget share it spends on nourishment. Within our panel data model, the  
 436 varying coefficient specification helps to highlight local authorities’ views about each of

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

Figure 1: Dependence implied by weight matrices under consideration



442 *4.1.2. Spatial Error Dependence (SED) versus Spatial Lag Dependence (SLD)*

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

Table 7: Descriptive statistics

	Mean	StD	Min	Max
<i>tph</i>	65.793	24.372	29.739	172.647
<i>mhs</i>	13.265	7.163	0.100	53.710
<i>nuc</i>	5,169.98	4,228.38	105.00	48,145.00
<i>pmp</i>	0.495	0.009	0.473	0.553
<i>pu14</i>	0.172	0.020	0.135	0.247
<i>smr</i>	967.366	131.461	583.100	1345.800
<i>noj</i>	202,926	175,994	19,000	2,130,000
<i>plp</i>	0.107	0.028	0.042	0.216
<i>mhp</i>	274,133	173,263	105,000	1,425,000
<i>mww</i>	469.094	77.396	332.100	784.400
<i>psq</i>	2823.152	3367.706	63.000	16425.320

450 its usefulness is diminished by its inability to disentangle or to identify these effects, i.e.  
451 the so-called reflection problem posed by Manski (1993). Within the context of our model,  
452 there are enough reasons to believe that a more relevant type of dependence is the SED. In  
453 (4.2), since  $Z$  represents the authorities specific socio-demographic/economic attributes,  
454  $\beta_{0,1}(Z)$  can help to indirectly model the contextual and correlated effects. Furthermore,  
455 the non-linear interactions, which are modelled via our varying coefficient specification,  
456 can help to capture these effects even more effectively. In addition, measurement errors  
457 that spill across grid boundaries, for example, can easily lead to the SED. Otherwise,  
458 there may exist unobservable latent variables that might be unaccounted for in the model.  
459 For instance, closure of a large psychiatric hospital, which serves patients from various  
460 municipalities, clearly has an impact on social care sector across a wide territory. Such a  
461 closure of hospitals, which has been one of the most prominent features of mental health  
462 policy in the UK for some years, will substantially increase the need for social care ser-  
463 vices across a wide area and ultimately influencing expenditure. Another example would  
464 be the provision of high-secure and medium-secure units for people with forensic needs,  
465 often organised at multi-regional level in line with nationally agreed population catchment  
466 areas. Their funding may be a NHS hospital, NHS trust, or other independent provider's  
467 responsibility, but there will be again a social care shared (spatial) resource effect of not  
468 providing these services. Other sources of unobserved spatial concentration could be sug-  
469 gested as the high psychiatric hospital admission in two or more neighbouring authorities,  
470 which may be caused by noise pollution from airports. Noise has been the major environ-  
471 mental issue in the field of aviation, primarily impacting residential communities close to  
472 airports by affecting community annoyance, sleep deprivation, and mental health issues.

473 *4.1.3. Spatial weight matrices*

474       Regarding the SAR in (2.3), how and to what extent the MHS by a local authority  
475 depends on that of the others are captured by elements in the matrix  $W_N$ . In spatial  
476 econometrics in general, a weight matrix is constructed based either on geographical or  
477 socio-economic/demographic distances of individuals. Within the context of our model,  
478 we argue that geographical-based weight matrices are more effective due to a number  
479 of reasons. Firstly, they are exogenous to the model and also time-non-varying. These  
480 conditions are required, but cannot be guaranteed when adopting weights based on socio-  
481 economic/demographic distance metrics. Moreover, our varying-coefficient specification  
482 enables modelling non-linear interactions between the risk factors of mental health need  
483 and authorities specific socio-economic/demographic attributes. Hence, it is reasonable to  
484 assume that socio-economic/demographic interactions of MHS spending are fully captured  
485 within the model. In the current section, we consider a similar set of weight matrices to  
486 that used in Section 3, so that we can analyse if and how estimation results change with  
487 weight matrices that differ in their degree of sparseness. In particular, we consider weights  
488 matrices, which are constructed based on (i) the  $k$ -nearest neighbours criteria, where  $k$  is 4,  
489 10, or 16, and (ii) sphere of influence. Figure 1 depicts spatial dependence implied by these  
490 weight matrices, which are referred to hereafter as K4, K10, K16 and SW, respectively.

491 *4.2. Empirical analysis*

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

493 *4.2.1. Basic data exploration*

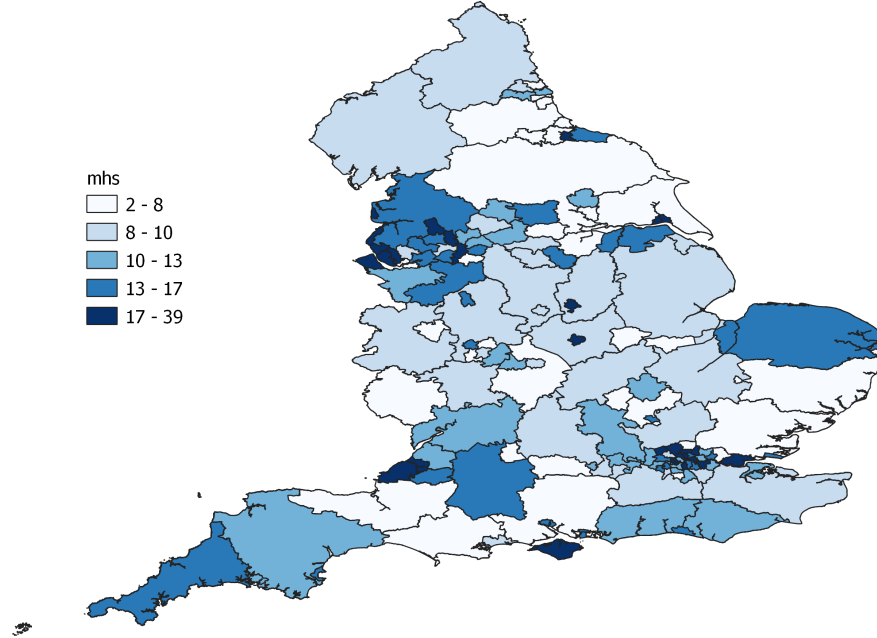
494       Table 7 presents descriptive statistics, which describe basic features of the data used in  
495 our study. Figure 2 presents the average *mhs* (i.e. per capita measure of standardised MHS  
496 for persons age between 18 and 64) for all the local authorities over 2016/17 to 2019/20.  
497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

499 *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

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

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,

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

$\alpha = 0.1$ ,  $d_N$  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

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

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

- 501
- In both panels of the table, the estimates of the autoregressive parameter increase

502 as higher number of nearest neighbours being taken into consideration. Moreover,

503 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 vary across

504 different weight matrices used. The selected number of relevant variables are 5 and

505 3 when  $z$  is defined as the percentage of right-wing voters, *vote*, and total public

506 health, *tph*, respectively. Similarly, the outcomes of the coefficient constancy test

507

Table 8: Estimation results

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

$W$	$\rho$	$\phi$	$\hat{K}$	$inct$	$nuc$	$pmp$	$pu14$	$smr$	$noj$	$plp$	$mhp$	$mwu$	$psq$
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

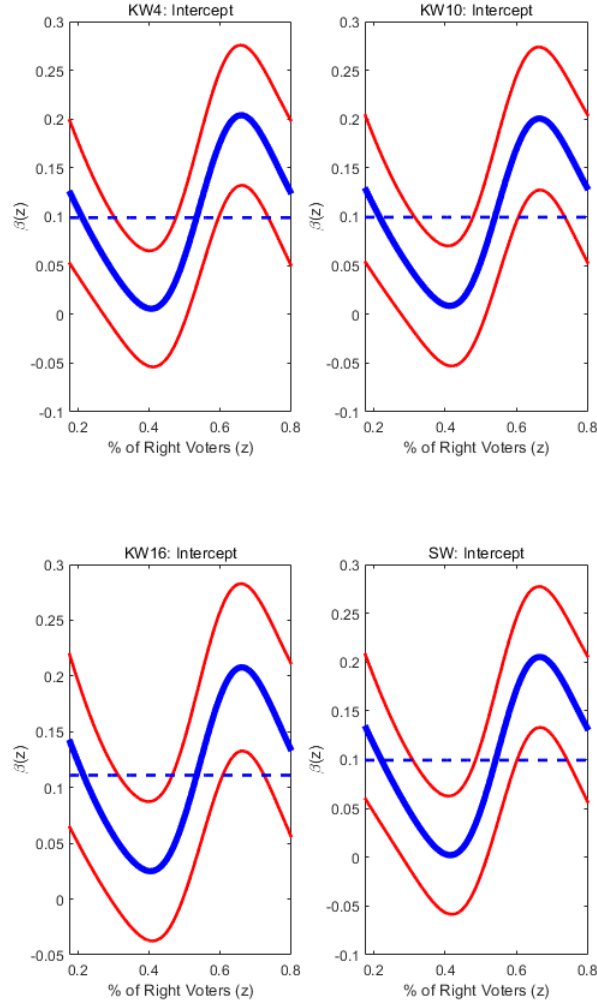
(b)  $z$  is defined as total public health ( $tph$ )

$W$	$\rho$	$\phi$	$\hat{K}$	$inct$	$nuc$	$pmp$	$pu14$	$smr$	$noj$	$plp$	$mhp$	$mwu$	$psq$
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.

- 508 remain largely unchanged across different weight matrices used. However, without  
509 taking into consideration the potential SAR and error component structure, the  
510 selection suggests that all (but one) variables in each of the panels are relevant.  
511 Such a finding is in significant contrast to that based on the SAREC-KLASSO  
512 method. In addition, the test statistics of the coefficient constancy test are much  
513 larger compared to those associated with  $K4$ ,  $K10$ ,  $K16$ ,  $SI$  and  $K0$ .
- 514 • By applying the SAREC-KLASSO method, we have found that the intercept and  
515 two other regressors are relevant in explaining the disparities between mental health  
516 spending by councils in England, namely median weekly wage ( $mwu$ ) and population  
517 per square kilometre ( $psq$ ). While the effects of these regressors depends non-linearly  
518 on total public health ( $tph$ ), they are independent of the percentage of right-wing  
519 voters ( $vote$ ). These findings are also supported by the confidence bands drawn  
520 in Figures 3 to 8. These confidence bands are drawn at the 10% significance level  
521 and suggest consentaneously that estimates of the functional coefficients for the  
522 *intercept*,  $mwu$  and  $psq$  are statistically significant. Furthermore, the dependence  
523 of their effects on  $tph$  and independence from  $vote$ , which we highlighted earlier, are  
524 also confirmed.
  - 525 • Interestingly, the percentage of male population ( $pmp$ ) and standardized mortality

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



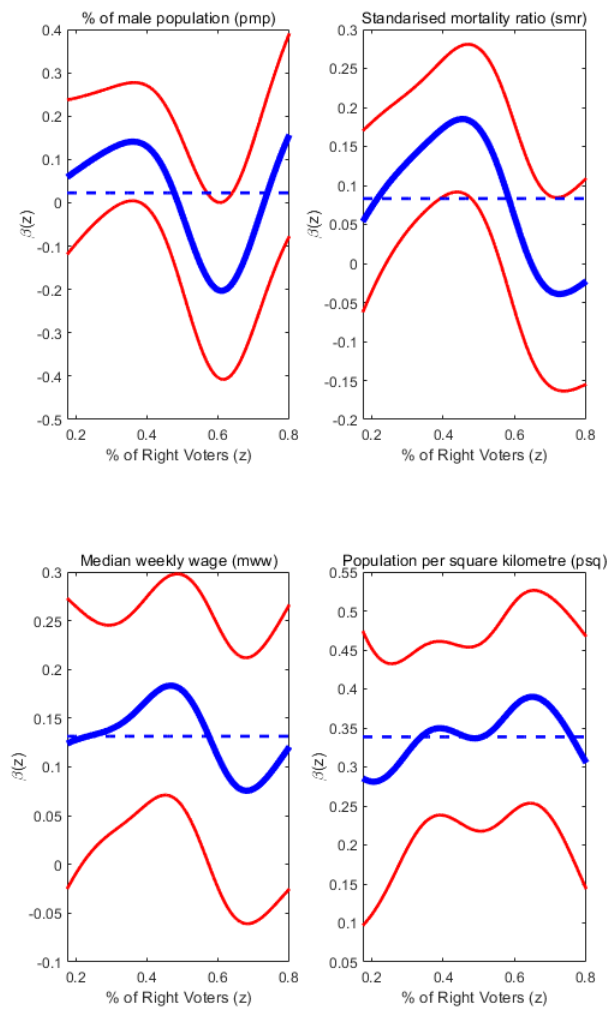
Note: The red solid curves are 90% confidence bands defined in (4.3). The blue broken line is  $\hat{C}_j$  in (4.4).

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



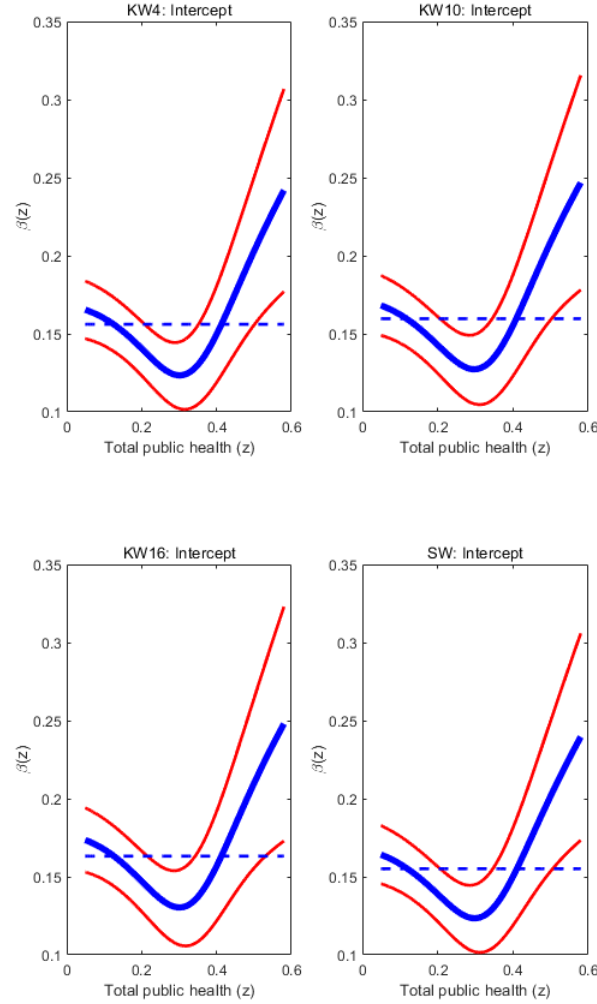
effects of  $smr$  on  $mhs$  indeed are dependent of  $vote$ .

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).

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

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

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

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

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

550 • From the figures, it is clear that these estimates (and their associated confidence  
551 bands) are quite consistent across the weight matrices used. As the results, our  
552 discussion will only concentrate on Figures 4 and 9, which are based on *KW4*, for  
553  $Z$  defined by *vote* and by *tph*, respectively.

554 • The impact of population density on spending is positive and significant as we ex-  
555 pected since we anticipate higher mental health expenditure in inner-city areas that  
556 are more densely populated.

557 • Furthermore, we find strong evidence that *mwv* should have a demand-side inter-  
558 pretation instead since councils with higher median weekly earnings seem to spend  
559 more on mental health services. Such a result is consistent with that reported in  
560 Moscone et al. (2007) for their spatial error model. The estimate of the functional  
561 coefficient suggests that impact of *mwv* on *mhs* increases between low to medium  
562 *tph* (i.e. *mwv* plays a similar role to the luxury goods in the Engel curve literature),  
563 but decreases between medium to high *mhs* (i.e. *mwv* plays a similar role to the  
564 inferior goods in the Engel curve literature).

565 • Moreover, independently to *vote* and *tph*, the percentage of male population does  
566 not seem to have a significant effect on the mental health service spending across  
567 councils. However, by conditioning its effect on the UK political spectrum, it seems  
568 that *pmp* has a positive (negative) impact on *mhs* in councils that are dominated  
569 by central-left (central-right) politics.

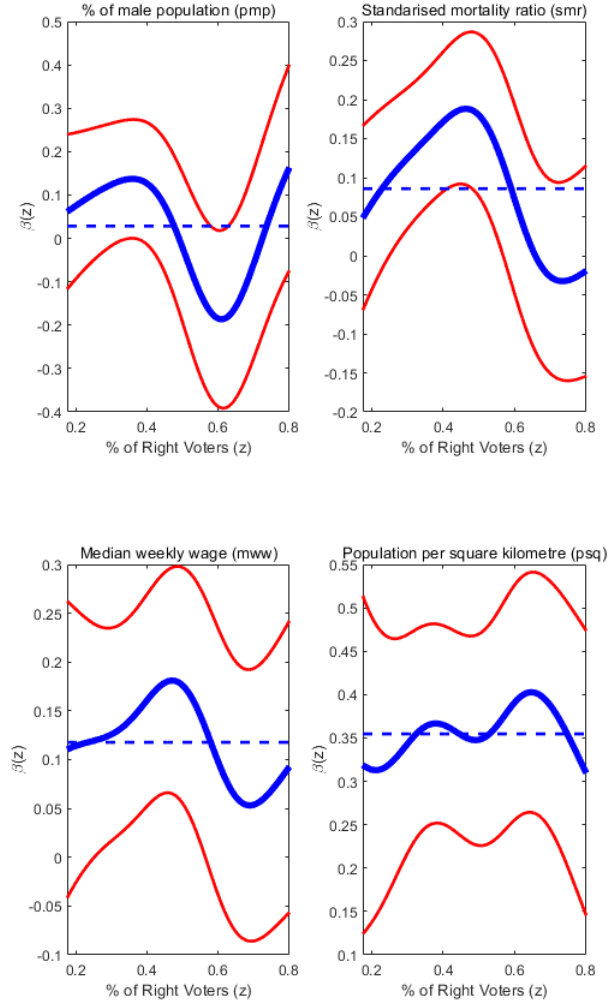
570 • Similarly, independently to *vote* and *tph*, the standardised mortality ratio does not  
571 seem to have a significant effect on the mental health service spending across councils.  
572 However, by conditioning its effect on the UK political spectrum, it seems that *smr*  
573 has a positive impact on *mhs* in councils that are dominated by central-left politics.

574

## 575 5. Conclusions

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

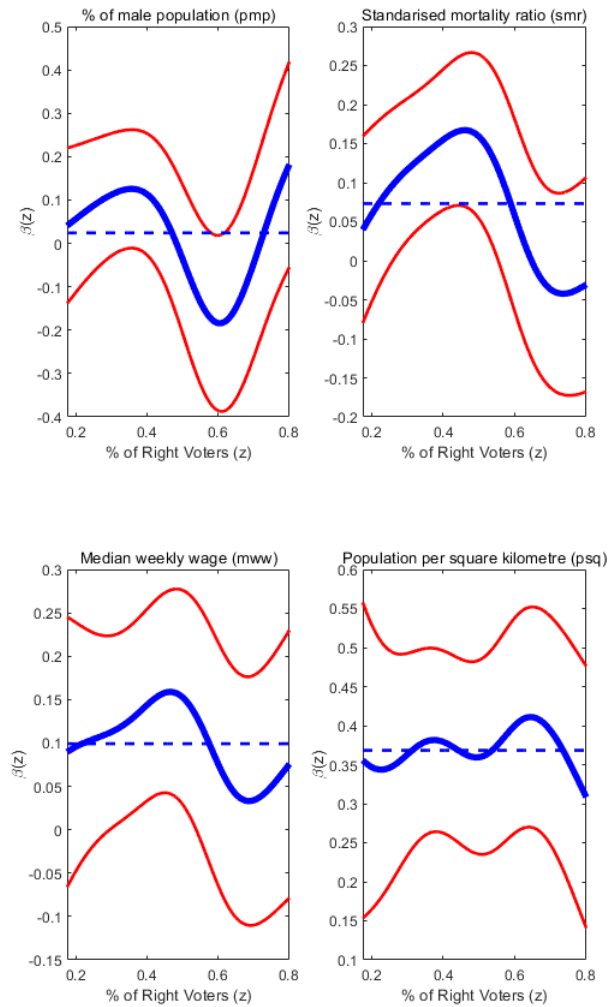
Figure 6: Estimates coefficient functions based on KW10:  $Z$  represents *vote*



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

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

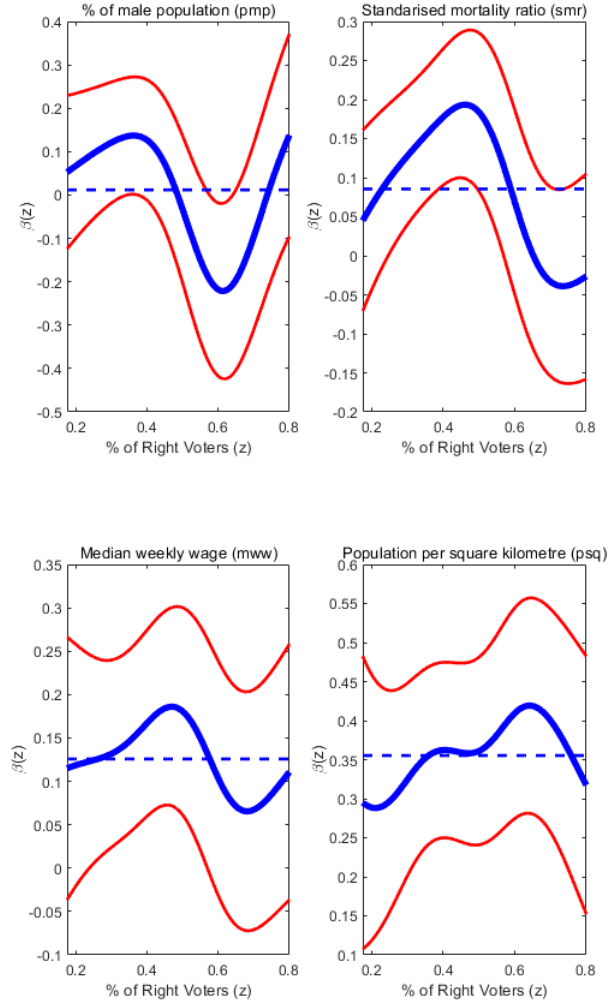
Figure 7: Estimates coefficient functions based on KW16:  $Z$  represents *vote*



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

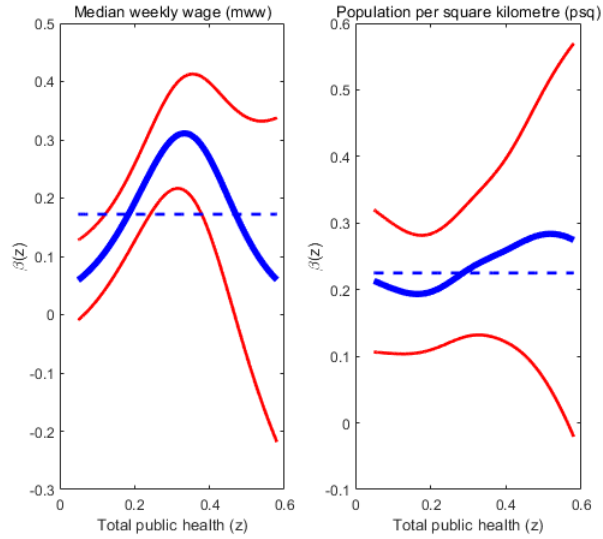
590 and illustrated their practical applicability by applying them to analyse municipal dispar-  
 591 ities in MHS spending by councils in England. Specifically, we studied the interaction  
 592 between a set of demand and supply factors of mental health needs and local authorities  
 593 specific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

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



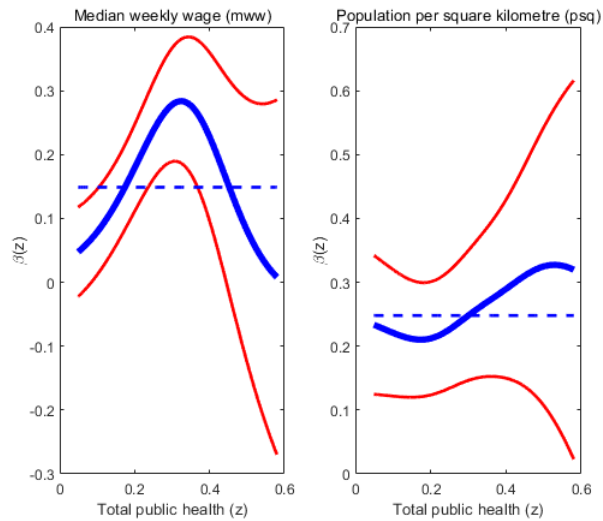
Note: The red solid curves are 90% confidence bands defined in (4.3). The blue broken line is  $\hat{C}_j$  in (4.4).

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



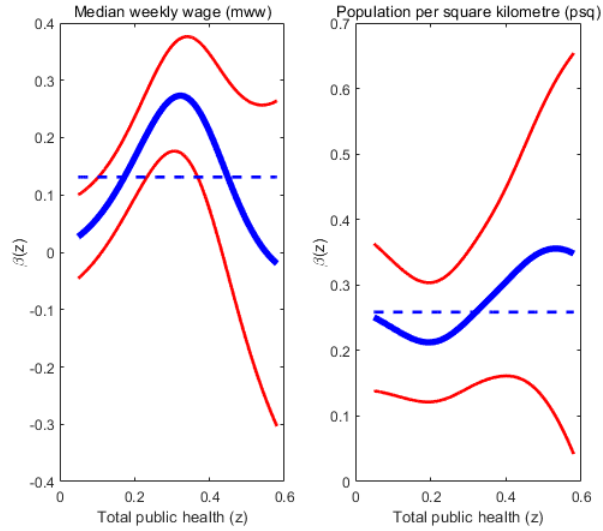
Note: The red solid curves are 90% confidence bands defined in (4.3). The blue broken line is  $\hat{C}_j$  in (4.4).

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



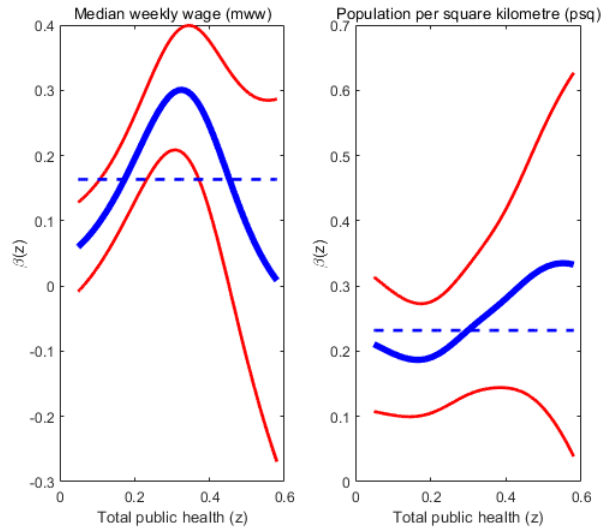
Note: The red solid curves are 90% confidence bands defined in (4.3). The blue broken line is  $\hat{C}_j$  in (4.4).

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).



595 **6. Appendices**

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

601 *A. Definitions*

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

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

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  $(i, t) \equiv it = \{11, 21, \dots, N1, 12, \dots, N2, 13, \dots, NT\}$ .  
 Let  $m = (m_{it,d}) \in \mathbb{R}^{NT \times D}$  denote an arbitrary  $NT \times D$  matrix with rows

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

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

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

and

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

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

603 *B.1. Proof of Lemma 2.1:*

604 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

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

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

606 *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

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

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

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

607 established in Lemma 2.1. Given the unique identification of  $\delta_0$ , consistency of  $\hat{\delta} = (\hat{\phi}, \hat{\rho})^\top$  as  
 608 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  $\Delta$ .  
 609 This can be established in two steps, namely (i) establishing point-wise convergence of  $\frac{1}{NT}\tilde{\ell}_z^c(\delta)$  to  
 610  $\frac{1}{NT}\bar{\ell}_z^c(\delta)$ , 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

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

611 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  $\tilde{u}_{js} =$   
 612  $\ddot{y}_{js} - \ddot{X}_{js}\tilde{\beta}_{it} = \ddot{X}_{js}(\beta_0 - \tilde{\beta}_{it}) + \ddot{u}_{0js}$ , 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

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

613 Moreover,

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

614 and

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

where

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

In this regard, we note firstly that

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

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

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

and

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

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

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

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

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

617 This is so because

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

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

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

618 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

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

where  $\delta^* \in \Delta$  lies on an  $\epsilon$ -neighborhood of  $\delta$  such that  $\|\delta - \delta^*\| = 0$  as  $\epsilon \rightarrow 0$ ,  $\bar{\delta}$  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

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

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

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

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

Now let us consider the unique identification conditions of  $\delta_0$ . The unique identification of  $\delta_0$  is firstly considered by showing the counter argument. Consider the Jensen's inequality below

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

The equality of (B.14) holds when  $\mathbb{Q}_N \mathbb{Q}_{0N}^{-1} = \mathbb{Q}_N^{-1} \mathbb{Q}_{0N} = I_{NT}$ . Hence  $\delta_0$  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  $\epsilon$ -neighborhood and its complement, respectively, and  $\lim_{N \rightarrow \infty} \mathbb{Q}_N(\delta^*) \rightarrow \lim_{N \rightarrow \infty} \mathbb{Q}_{0N}$ . Hence the unique identification condition requires that

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

626 for any  $\delta$ .

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

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

630 *C. Useful lemmas*

631 In this section, we present a set of lemmas that will be useful for the proof that follows. For  
 632 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,  
 633 let  $\ddot{X}_{j_s}$  be the  $j_s$ -th row of  $\ddot{X}_N$  and  $\ddot{y}_{j_s}$  be the  $j_s$ -th element of  $\ddot{y}_N = \ddot{y}_N$ .

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

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

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

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

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

$$\sup_{z \in [0,1]} \left| (NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T [K_h(Z_{it} - z) \xi_{it} - E \{K_h(Z_{it} - z) \xi_{it}\}] \right|$$

$$= O_P \left\{ h^2 + \left( \frac{\log(1/h)}{(NT)h} \right)^{1/2} \right\}, \quad (\text{C.2})$$

636 where  $(\xi_{it}, Z_{it})$  are i.i.d. random vectors,  $\xi_{it}$  are scalar random variables with  $E|\xi_{it}|^s < \infty$ , and  
 637  $\sup_z \int |y|^s f(z, v) dv < \infty$  (where  $f$  denotes the joint density of  $(\xi_1, Z_1)$ ). The proof of (C.2) can be  
 638 found in various existing works, e.g. Fan and Zhang (2000a). Regarding (C.1):

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

639 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,*

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

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

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

641 where

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

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

642 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,*

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

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

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

644 **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

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

In addition,

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

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

646 *Proof of Lemma C.4:* By using the same argument as in Lemma C.2, i.e. the uniform consistency  
647 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  $z \in [0, 1]$   
648 and are therefore negligible. As the results, we simply write  $\hat{u}_{it} = \hat{u}_{1,it} + \hat{u}_{2,it}$ , where

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

649 We intend to show that

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

Firstly, we start by writing

$$\begin{aligned} \hat{u}_{1,it}^2 &= \frac{h}{NT} \sum_{j \neq i} \sum_{s \neq \tau} q_{js,\nu\tau,it} + \frac{h}{NT} \sum_{j=1}^N \sum_{s=1}^T r_{js,it} = |\hat{u}_{1,it}|^2 \\ E |\hat{u}_{1,it}|^2 &= \frac{h}{NT} \sum_{j \neq i} \sum_{s \neq \tau} E q_{js,\nu\tau,it} + \frac{h}{NT} \sum_{j=1}^N \sum_{s=1}^T E r_{js,it}, \end{aligned} \quad (\text{C.11})$$

650 where

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

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

$$\begin{aligned} q_{js,\nu\tau,it,1} &= \{\beta'_0(Z_{it})\}^\top \ddot{X}_{js}^\top \ddot{X}_{js} \ddot{X}_{\nu\tau}^\top \ddot{X}_{\nu\tau} \beta'_0(Z_{it}) (Z_{js} - Z_{it})(Z_{\nu\tau} - Z_{it}) K_h(Z_{it} - Z_{js}) K_h(Z_{it} - Z_{\nu\tau}) \\ q_{js,\nu\tau,it,2} &= C \ddot{X}_{js}^\top \ddot{X}_{js} \ddot{X}_{\nu\tau}^\top \ddot{X}_{\nu\tau} (Z_{js} - Z_{it})^2 (Z_{\nu\tau} - Z_{it})^2 K_h(Z_{it} - Z_{js}) K_h(Z_{it} - Z_{\nu\tau}). \end{aligned}$$

652 In addition,  $E q_{js,\nu\tau,it} = E q_{js,\nu\tau,it,1} + E q_{js,\nu\tau,it,2}$ . Regarding the first term,

$$\begin{aligned} E q_{js,\nu\tau,it,1} &= E \{ E [q_{js,\nu\tau,it,1} | Z_{js} = Z_{it}, Z_{\nu\tau} = Z_{it}] \} \\ &= \int \{\beta'_0(Z_{it})\}^\top \Omega(Z_{js}) \Omega(Z_{\nu\tau}) \beta'_0(Z_{it}) (Z_{js} - Z_{it})(Z_{\nu\tau} - Z_{it}) \\ &\quad K_h(Z_{it} - Z_{js}) K_h(Z_{it} - Z_{\nu\tau}) f(Z_{it}) f(Z_{js}) f(Z_{\nu\tau}) dZ_{it} dZ_{js} dZ_{\nu\tau} \\ &= h^2 \int \{\beta'_0(Z_{it})\}^\top \Omega(Z_{js}) \Omega(Z_{\nu\tau}) \beta'_0(Z_{it}) v_1 v_2 K(v_1) K(v_2) dv_1 dv_2 f(Z_{js}) f(Z_{\nu\tau}) f(Z_{it}) dZ_{it} \\ &= h^2 \int \{\beta'_0(Z_{it})\}^\top \Omega(Z_{js}) f(Z_{js}) \Omega(Z_{\nu\tau}) f(Z_{\nu\tau}) \beta'_0(Z_{it}) f(Z_{it}) dZ_{it} \int \int v_1 v_2 K(v_1) K(v_2) dv_1 dv_2 \\ &= O(h^2) o(1), \end{aligned}$$

653 where  $\Omega(Z_{js}) = E[\ddot{X}_{js}^\top \ddot{X}_{js} | Z_{js} = Z_{it}]$  and  $\Omega(Z_{\nu\tau}) = E[\ddot{X}_{\nu\tau}^\top \ddot{X}_{\nu\tau} | Z_{\nu\tau} = Z_{it}]$ , by which the third and  
654 fourth equality are obtained based on  $Z_{js} = Z_{it} + v_1 h$  and  $Z_{\nu\tau} = Z_{it} + v_2 h$ , and

$$\begin{aligned} &\int \{\beta'_0(Z_{it})\}^\top \Omega(Z_{js}) f(Z_{js}) \Omega(Z_{\nu\tau}) f(Z_{\nu\tau}) \beta'_0(Z_{it}) f(Z_{it}) dZ_{it} \\ &= E \{ \beta'_0(Z_{it}) \ddot{X}_{it}^\top \ddot{X}_{it} \ddot{X}_{it}^\top \ddot{X}_{it} \beta'_0(Z_{it}) \} = O(1), \end{aligned}$$

655 respectively. Regarding the second term,  $E q_{js,\nu\tau,it,2} = O(h^4)$  uniformly over all pairs  $(i, t)$ ,  $i =$   
656  $1, \dots, N$  and  $t = 1, \dots, T$ , that is

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

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

$$Er_{j_s, it} = O(h^3) \quad (\text{C.12})$$

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

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

Secondly,

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

659 so that

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

660 Observe that

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

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

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

662 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*

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

*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

$$\liminf_{NT \rightarrow \infty} P \left[ \inf_{(NT)^{-1} \|m\|_F^2 = C} Q(B_0 + \{(NT)h\}^{-1/2} m) > Q(B_0) \right] = 1 - \epsilon, \quad (\text{C.17})$$

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

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

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

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

665 In this regard,

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

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

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

668 This suggests writing

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

669 where

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

670 Moreover,

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

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

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

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

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

673 In this regard, it is the case that  $\gamma_0^{\min} > 0$ . Since  $(NT)^{-1}\|\hat{u}\|^2 = O_P(1)$  using Lemma C.4, positivity  
674 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*

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

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

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

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

677 To prove Lemma C.6 only requires showing that

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

678 With regard to the first term,

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

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

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

**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,

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

*Proof of Lemma C.7:* Observe that

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

681 so that

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

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

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

682 which follows Lemma C.1.

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

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

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

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

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  $\alpha_2$  is also a  $NT \times 1$  vector in which

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

Let us first consider  $\alpha_2$ . 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

$$\begin{aligned}\alpha_{21,it} &= \sum_{j=1}^N \sum_{s=1}^T \ddot{X}_{D,j_s} \ddot{u}_{j_s} K_h(Z_{it} - Z_{j_s}) + R_{21,it}, \\ \alpha_{22,it} &= \sum_{j=1}^N \sum_{s=1}^T \ddot{X}_{D,j_s} \ddot{X}_{j_s} (\beta_0(Z_{j_s}) - \beta_0(Z_{it})) K_h(Z_{it} - Z_{j_s}) + R_{22,it}, \\ \alpha_{23,it} &= \sum_{j=1}^N \sum_{s=1}^T \ddot{X}_{D,j_s} \ddot{X}_{j_s} (\beta_0(Z_{it}) - \hat{\beta}_\lambda(Z_{it})) K_h(Z_{it} - Z_{j_s}) + R_{23,it}.\end{aligned}$$

683 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  $R_3$ ,  
684 they are asymptotically negligible.

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

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

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

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

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

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

686 These implies collectively that

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

Finally, it is the case that

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

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

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

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

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

693 *D. Proof of the results in Section 2.2*

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

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

696 *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

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

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

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

698 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

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

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

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

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

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

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

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

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

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

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

703 Similarly,

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

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

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

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

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

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

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

705 and

$$E \left\{ \frac{1}{NT h} \sum_{j=1}^N \sum_{s=1}^T \ddot{X}_{0js}^\top \ddot{X}_{0js} (\beta_0(Z_{js}) - \beta_0(z)) K \left( \frac{Z_{js} - z}{h} \right) \right\} = \mathfrak{K}_2 h^2 \mathfrak{B}.$$

Finally,

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

706 and

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

Therefore,

$$\sqrt{NT h} \left( \hat{\beta}(z) - \beta_0(z) - \text{Bias} \right) \rightarrow_D N(0, \Sigma),$$

707 where  $\text{Bias} = \mathfrak{D}^{-1}(z) \mathfrak{K}_2 h^2 \mathfrak{B}$  and  $\Sigma = \mathfrak{D}^{-1}(z) V(z) \mathfrak{D}^{-1}(z)$ .

708 *E. Proof of results in Section 2.3*

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

$$\hat{\mathcal{M}}(z) = \begin{pmatrix} \hat{\mathcal{M}}_{aa}(z) & \hat{\mathcal{M}}_{ab}(z) \\ \hat{\mathcal{M}}_{ba}(z) & \hat{\mathcal{M}}_{bb}(z) \end{pmatrix}, \hat{\mathcal{N}}(z) = \begin{pmatrix} \hat{\mathcal{N}}_a(z) \\ \hat{\mathcal{N}}_b(z) \end{pmatrix}, \hat{\mathcal{D}} = \begin{pmatrix} \hat{\mathcal{D}}_{aa} & 0 \\ 0 & \hat{\mathcal{D}}_{bb} \end{pmatrix},$$

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

$$\hat{\mathcal{M}}_{aa}(z) = \frac{1}{NT} \sum_{j=1}^N \sum_{s=1}^T \hat{X}_{a,js}^\top \hat{X}_{a,js} K_h(Z_{it} - Z_{js}), \hat{\mathcal{M}}_{bb}(z) = \frac{1}{NT} \sum_{j=1}^N \sum_{s=1}^T \hat{X}_{b,js}^\top \hat{X}_{b,js} K_h(Z_{it} - Z_{js}),$$

$$\hat{\mathcal{M}}_{ab}(z) = \frac{1}{NT} \sum_{j=1}^N \sum_{s=1}^T \hat{X}_{a,js}^\top \hat{X}_{b,js} K_h(Z_{it} - Z_{js}) = \hat{\mathcal{M}}_{ba}(z),$$

$$\hat{\mathcal{N}}_a(z) = \frac{1}{NT} \sum_{j=1}^N \sum_{s=1}^T \hat{X}_{a,js}^\top \hat{y}_{js} K_h(Z_{it} - Z_{js}) \text{ and } \hat{\mathcal{N}}_b(z) = \frac{1}{NT} \sum_{j=1}^N \sum_{s=1}^T \hat{X}_{b,js}^\top \hat{y}_{js} K_h(Z_{it} - Z_{js}).$$



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

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

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

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

713 *E.1. Proof of Theorem 2.4:*

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

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

We note firstly that both  $\hat{\mathcal{N}}_a(z)$  and  $\hat{\mathcal{N}}_b(z)$  are uniformly bounded in a similar fashion to Lemmas C.1 and C.2. Hence, to prove that  $\hat{\beta}_{\lambda,b}(z) \rightarrow 0$  as  $NT \rightarrow \infty$  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  $\hat{\mathcal{D}}_{bb}$  are  $\lambda_d / \|\hat{b}_d\|$  for  $(D_0 + 1) \leq d \leq D$ , and

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

which is in accordance with Theorem 2.2. Hence,

$$\min \left\| \hat{\mathcal{D}}_{bb} \right\| = \left\| \text{diag} \left\{ \frac{b_N}{\|\hat{b}_{D_0+1}\|}, \dots, \frac{b_N}{\|\hat{b}_D\|} \right\} \right\| \rightarrow \infty \quad (\text{E.2})$$

714 due to Assumption E1. This completes the proof.

715 *E.2. Proof of Theorem 2.5*

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

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

whereas the unpenalized counterpart is

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

716 where

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

Hence, the difference between these estimators is

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

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

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

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

718 *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  $\lambda_{NT}$  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

$$R\hat{S}S_F \rightarrow_P R\tilde{S}S_F \quad \text{whereas} \quad R\tilde{S}S_F \rightarrow_P \sigma_{v,0}^2, \quad (\text{E.5})$$

and from Lemmas C.2 and C.3

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

Furthermore,

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

719 due to the reverse triangle inequality.

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

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

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

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

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

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

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

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

723 but focus instead on

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

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

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

725 Now, we consider the case where  $\lambda \in \mathbb{R}_+$ , so that an arbitrary model  $\mathcal{S}_\lambda$  is over-fitted. In  
726 addition, let  $\hat{B}_{S_\lambda} = (\hat{\beta}_{S_\lambda}(Z_{11}), \dots, \hat{\beta}_{S_\lambda}(Z_{11}))^\top$  denote an unpenalised estimate, which belongs to

727 the over-fitted model  $S_\lambda$ . Similarly to (E.7), we may define

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

where

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

728 In this regard, we deduce from the over-fitting nature and the consistency of the unpenalised  
729 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,

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

730 Moreover,

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

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

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

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)

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

732 Moreover,

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

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

737 *F. Proof of Corollary 2.1*

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

739 *G. Proof of results in Section 2.5*

740 *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  $\hat{\mathcal{D}}$  with

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

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

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

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

751 *Proof of Corollary 2.3:*

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

753 **References**

- 754 Aziz, F., McCrone, P., Boyle, S., Knapp, M., 2003. Financing mental health services in london:  
755 central funding and local expenditure .
- 756 Baltagi, B.H., Bresson, G., Pirotte, A., 2012. Forecasting with spatial panel data. *Computational*  
757 *Statistics & Data Analysis* 56, 3381–3397.
- 758 Baltagi, B.H., Song, S.H., Koh, W., 2003. Testing panel data regression models with spatial error  
759 correlation. *Journal of econometrics* 117, 123–150.
- 760 Baltagi, B.H., et al., 2008. *Econometric analysis of panel data*. volume 4. Springer.
- 761 Cai, Z., Fan, J., Li, R., 2000. Efficient estimation and inferences for varying-coefficient models.  
762 *Journal of the American Statistical Association* 95, 888–902.
- 763 Cliff, A.D., 1973. *Spatial autocorrelation*. Technical Report.
- 764 Cliff, A.D., Ord, J.K., 1981. *Spatial processes: models & applications*. Taylor & Francis.
- 765 Fan, J., Farmen, M., Gijbels, I., 1998. Local maximum likelihood estimation and inference. *Journal*  
766 *of the Royal Statistical Society: Series B (Statistical Methodology)* 60, 591–608.
- 767 Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle proper-  
768 ties. *Journal of the American statistical Association* 96, 1348–1360.
- 769 Fan, J., Zhang, J.T., 2000a. Two-step estimation of functional linear models with applications to  
770 longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*  
771 62, 303–322.
- 772 Fan, J., Zhang, W., 1999. Statistical estimation in varying coefficient models. *The annals of*  
773 *Statistics* 27, 1491–1518.
- 774 Fan, J., Zhang, W., 2000b. Simultaneous confidence bands and hypothesis testing in varying-  
775 coefficient models. *Scandinavian Journal of Statistics* 27, 715–731.
- 776 Fan, J., Zhang, W., 2008. Statistical methods with varying coefficient models. *Statistics and its*  
777 *Interface* 1, 179.
- 778 Feng, G., Gao, J., Peng, B., Zhang, X., 2017. A varying-coefficient panel data model with fixed  
779 effects: theory and an application to us commercial banks. *Journal of econometrics* 196, 68–82.
- 780 Gao, J., Xia, K., Zhu, H., 2020. Heterogeneous panel data models with cross-sectional dependence.  
781 *Journal of Econometrics* 219, 329–353.
- 782 Hu, T., Xia, Y., 2012. Adaptive semi-varying coefficient model selection. *Statistica Sinica* , 575–  
783 599.
- 784 Hunter, D., Li, R., 2005. Variable selection using mm algorithms. *Annals of statistics* , 1617.

- 785 Kapoor, M., Kelejian, H.H., Prucha, I.R., 2007. Panel data models with spatially correlated error  
786 components. *Journal of Econometrics* 140, 97–130.
- 787 Kelejian, H.H., Prucha, I.R., 1999. A generalized moments estimator for the autoregressive pa-  
788 rameter in a spatial model. *International economic review* 40, 509–533.
- 789 Lee, L.f., 2004. Asymptotic distributions of quasi-maximum likelihood estimators for spatial au-  
790 toregressive models. *Econometrica* 72, 1899–1925.
- 791 Lee, L.f., Yu, J., 2010. Estimation of spatial autoregressive panel data models with fixed effects.  
792 *Journal of econometrics* 154, 165–185.
- 793 Liu, S.F., Yang, Z., 2015. Asymptotic distribution and finite sample bias correction of qml esti-  
794 mators for spatial error dependence model. *Econometrics* 3, 376–411.
- 795 Magnus, J.R., Muris, C., 2010. Specification of variance matrices for panel data models. *Econo-  
796 metric Theory* 26, 301–310.
- 797 Manski, C.F., 1993. Identification problems in the social sciences. *Sociological Methodology* , 1–56.
- 798 McCrone, P., Jacobson, B., 2004. Indicators of mental health activity in london: Adjusting for  
799 sociodemographic need. London: London Development Centre for Mental Health .
- 800 Moscone, F., Knapp, M., Tosetti, E., 2007. Mental health expenditure in england: a spatial panel  
801 approach. *Journal of Health Economics* 26, 842–864.
- 802 Okui, R., Takahide, Y., 2018. Kernel estimation for panel data with heterogeneous dynamics.  
803 Working Paper .
- 804 Robinson, P.M., 2011. Asymptotic theory for nonparametric regression with spatial data. *Journal  
805 of Econometrics* 165, 5–19.
- 806 Rodriguez-Poo, J.M., Soberon, A., 2014. Direct semi-parametric estimation of fixed effects panel  
807 data varying coefficient models. *The Econometrics Journal* 17, 107–138.
- 808 Su, L., Yang, Z., 2015. Qml estimation of dynamic panel data models with spatial errors. *Journal  
809 of Econometrics* 185, 230–258.
- 810 Sun, Y., Carroll, R.J., Li, D., 2009. Semiparametric estimation of fixed-effects panel data varying  
811 coefficient models, in: *Nonparametric econometric methods*. Emerald Group Publishing Limited.
- 812 Wang, H., Leng, C., 2007. Unified lasso estimation by least squares approximation. *Journal of the  
813 American Statistical Association* 102, 1039–1048.
- 814 Wang, H., Xia, Y., 2009. Shrinkage estimation of the varying coefficient model. *Journal of the  
815 American Statistical Association* 104, 747–757.
- 816 Xia, Y., Zhang, W., Tong, H., 2004. Efficient estimation for semivarying-coefficient models.  
817 *Biometrika* 91, 661–681.

- 818 Zou, H., 2006. The adaptive lasso and its oracle properties. *Journal of the American Statistical*  
819 *Association* , 1418–1429.
- 820 Zou, H., Li, R., 2007. One-step sparse estimates in nonconcave penalized likelihood models. *Annals*  
821 *of statistics* 36, 1509.