



# City Research Online

## City St George's, University of London

**Citation:** Wongsa-art, P., Kim, N., Xia, Y. & Moscone, F. (2024). Varying coefficient panel data models and methods under correlated error components: Application to disparities in mental health services in England. *Regional Science and Urban Economics*, 106, 104009. doi: 10.1016/j.regsciurbeco.2024.104009

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

**Permanent repository link:** <https://openaccess.city.ac.uk/id/eprint/33168/>

**Link to published version:** <https://doi.org/10.1016/j.regsciurbeco.2024.104009>

**Copyright and Reuse:** Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to [City Research Online policy](#).

# Varying Coefficient Panel Data Models and Methods Under Correlated Error Components: Application to Disparities in Mental Health Services in England

Pipat Wongsart  
City University of London

Namhyun Kim  
University of Exeter

Yingcun Xia\*  
National University of Singapore  
University of Electronic Science and Technology of China

Francesco Moscone  
Brunel University London  
Ca' Foscari University of Venice

May 15, 2024

## Abstract

The contribution of this paper is twofold. Firstly, it introduces novel regression models that combine two important areas of the methodological development in panel data analysis, namely a varying coefficient specification and spatial error dependence. The former allows relatively flexible nonlinear interactions; the latter enables spatial correlations of the disturbance and thus differ significantly from the other random effect models in the literature. To estimate the model, a new estimation procedure is established that can be viewed as a generalization of the quasi-maximum likelihood method for a spatial panel data model to the well-known conditional local likelihood procedure. Novel inference methods, particularly variable selection and hypothesis testing of the parameter constancy, are introduced and are shown to be effective under the complex spatial error dependence. Equally importantly, this paper makes a substantial contribution to the understanding of financing and expenditure for health and social care. In particular, we empirically analyze and explain the effects of political ideologies on the local fiscal policy in England, especially the expenditure on mental health services.

*Keywords:* Varying coefficient panel data models, spatial error dependence, conditional local maximum likelihood, variable selection, hypothesis testing of the parameter constancy, expenditure on mental health services

---

\*The author gratefully acknowledges National Natural Science Foundation of China (72033002) and MOE AcRF of Singapore (A-8000021-00-00)

# 1 Introduction

To analyze important phenomena in economics, many researchers have recognized the need to exploit the rich information available in panel data sets. While Moscone et al. (2007) and Meng et al. (2021) are excellent examples of studies in health economics, Fingleton (2009), and Ihlanfeldt and Mayock (2010) are in urban economics. In consequence, we have witnessed extensive development in various methodological aspects of panel data analysis in recent years (see Sarafidis and Wansbeek (2021) for an excellent survey). This paper contributes to this development in two important areas, namely the varying-coefficient (VC) specification and spatial error dependence (SED). Equally importantly, this paper makes a substantial contribution to the understanding of financing and expenditure for health and social care. In particular, we empirically analyze and explain the effects of political ideologies on the local fiscal policy in England, especially the expenditure on mental health services (MHS). We elaborate on these points below.

Even though the VC specification is the main focus of many methodological studies, recently varying-coefficient panel data (VCPD) models have also attracted much attention (see e.g., Feng et al. (2017) and Dong et al. (2021)). However, recent extensions of the VCPD models to spatial econometrics have concentrated mainly on the spatial lag dependence (SLD). Some well-known examples of these works are Sun and Malikov (2018), Zhang and Shen (2015), and Liang et al. (2022) who estimate the SLD panel data model with functional coefficients. Although the SLD specification is useful for modeling the endogenous and contextual effects, which are important topics in the social-interaction literature, its significance is diminished by the inability to disentangle these effects. This is often referred to as the “reflection problem” first discussed by Manski (1993) (see also discussion in Anselin (2009)). Within the context of our VCPD model, we argue that the VC specifications can help to model both these effects (see Section 4 for more detail).

In our view, it is as important to study the SED as to study the SLD. Measurement errors that spill across grid boundaries and unobservable latent variables (unaccounted for in a model) are only two examples of many phenomenons that can lead to the SED. If they are not addressed, these problems can hinder effective inferences, particularly variable selection and specification testing. Moscone et al. (2007) provide the empirical illustration

of this issue in applied health economics. In addition, Fingleton (2009) and Kim et al. (2023) show that these problems are common in studies of urban and environmental economics, respectively.

This paper establishes a novel estimation procedure and inference methods for VC short panel regression models with first-order Spatial Autoregressive (SAR) disturbance terms, whose innovations have an error component structure. Our specification of the disturbance allows for spatial interactions of both the error and unit-specific components and, therefore differs significantly from the other random effects models in the literature (see e.g., Baltagi et al. (2012), and Liu and Yang (2015)). Furthermore, our estimation procedure can be viewed as a generalization of the quasi-maximum likelihood (QML) method for spatial panel data models (see e.g., Lee and Yu (2010), and Liu and Yang (2015)) to the conditional local kernel-weighted QML method. Many previous studies considered the conditional likelihood methods for estimating VC models (see Fan et al. (1998), Cai et al. (2000), and Fan and Zhang (2008) for details), but no existing work has done so within the context of a panel data regression that allows for spatial interactions. In these aspects, our model and methods synergize directly with the geographically weighted regression (GWR) models with SLD (i.e., the so-called GWR-SAR model) discussed in Li et al. (2019). As well as topics in health economics, these models have been applied extensively to problems in applied urban economics (see e.g., Tomal (2020), and Tomal and Helbich (2023)).

Moreover, we establish a novel procedure for variable selection for our spatial VCPD model. In this regard, Wang and Xia (2009) introduced the so-called Kernel Least Absolute Shrinkage and Selection Operator (KLASSO) technique. We show that this technique is ineffective when applied to panel data (especially under spatial interactions) and suggest an alternative procedure. We also extend our procedure to handle the selection of a more complex specification known as the semi-varying coefficient model in the literature. To ensure the theoretical validity of our procedures, we establish a set of asymptotic results based on the standard regularity conditions that are often used in the semi- and nonparametric literature, particularly Mack and Silverman (1982), Robinson (1988), Fan and Li (1999), Fan and Zhang (2000), and Robinson (2011). In conjunction with these results, we conduct an extensive simulation exercise to illustrate the finite-sample performance and robustness of our proposed procedures.

In recent years, the mental health issue has become the emerging health policy priorities in the UK from the point of view of healthcare (see Walker et al. (2019) for a comprehensive discussion) but the local authorities' expenditure and financing are also the topics of heated debate. This paper empirically studies these issues by applying our spatial VCPD model and inference methods to analyze the funding decisions by the English local authorities during the UK's fiscal squeeze. To understand this paper's contribution to the study of these topics, it is useful to first note the difference between "*variations in*" and "*disparities in*" MHS expenditure by the local authorities in England. A number of studies have empirically investigated determinants of the local authorities' MHS expenditure, particularly Aziz et al. (2003), McCrone and Jacobson (2004), Moscone et al. (2007), and Moscone and Tosetti (2010). Their empirical models are based on a traditional reduced-form demand/supply framework, whereby the MHS expenditure is explained by a set of risk factors of mental health needs such as population density, percentage of people aged under 14, and percentage of households headed by lone parent. In consequence, these models explain variations in the MHS expenditure by a local government on the basis of the mental health risk factors in which they merely analyze the marginal effects of those risk factors. On the contrary, we analyze the municipal disparities in local governments' MHS expenditures by allowing the marginal effects of those risk factors to depend on the political ideology (left-wing or right-wing) towards which voters in the respective local authorities are leaning. An intuitive example behind this hypothesis is the phenomenon in which the marginal effects of the percentage of households headed by lone parent on MHS spending are larger in left-leaning local authorities.

Within the context of the regression analysis, the most common modeling strategy is to include a slope dummy variable or an interaction term in order to incorporate the interaction effects. It is argued that the VC specification considered in this paper is able to offer a much more general framework for analyzing these disparities. Moreover, this paper illustrates how the spatial dependence brought about by measurement errors that spills across boundaries and/or spatially correlated unobservable latent variables can be captured by the SED specification. In the context of the MHS spending in the UK, one example of a factor that may lead to these types of spatial interactions is the closure of a large psychiatric hospital, which serves patients from various municipalities. Another

example is the high number of psychiatric hospital admissions in two or more neighboring authorities, which may be caused by aviation impacting residential communities close to airports by affecting community annoyance, sleep deprivation, and mental health issues.

The remainder of this paper is structured as follows. Section 2 introduces the VCPD model with the SED specification, then establishes the estimation procedure and inference methods. Section 3 presents and discusses an extensive simulation exercise examining the finite sample performances and robustness of our proposed procedures. Section 4 presents an empirical study of the determinants of MHS spending by local councils in England, and whether different political preferences of residents within the local authorities bring about the disparity that can lead to unequal access to MHS. Section 5 concludes the paper. Appendix A provides mathematical proofs of the main results in this paper, and Appendix B discusses important technical points, which cannot be included in the main sections. Tables and figures are presented in Appendix C. Finally, additional discussion and results that cannot be accommodated within the paper are presented in the supplementary material.

## 2 Statistical model and methods

This section introduces the model specifications, then discusses the estimation procedure and inference methods. Proofs and other technical discussions are in the appendices.

### 2.1 Model specification

Let  $y_{it} \in \mathbb{R}^1$  be a response of interest, and let  $X_{it}^\top = \{X_{it,1}, \dots, X_{it,D}\}^\top \in \mathbb{R}^D$  and  $Z_{it} \in [0, 1]$ , which are referred to hereafter as the “regressors” and “covariate”, respectively. Moreover, let  $\beta_0(z) = \{\beta_{1,0}(z), \dots, \beta_{D,0}(z)\}^\top \in \mathbb{R}^D$  be a vector of smooth nonparametric functions in  $z$  and let  $u_{it} \in \mathbb{R}^1$  denote the error term of which  $E(u_{it}|X_{it}, Z_{it}) = 0$  almost surely (a.s.). This paper assumes that  $y_{it}$  is generated by

$$y_{it} = X_{it}\beta_0(Z_{it}) + u_{it}, \quad i = 1, \dots, N \text{ and } t = 1, \dots, T. \quad (2.1)$$

Here,  $T$  is regarded as fixed hence, our asymptotic theory relies on  $N$  diverging.

To specify the SED, we define  $u_N = (u_{11}, u_{21}, \dots, u_{N1}, u_{12}, \dots, u_{N2}, u_{13}, \dots, u_{NT})^\top$ , whose elements are grouped by time periods rather than spatial units as is commonly

done in the panel data regression literature. In addition,  $y_N$  denotes an  $NT \times 1$  vector of  $y_{it}$  and  $X_N = (X_{11}^\top, X_{21}^\top, \dots, X_{N1}^\top, X_{12}^\top, \dots, X_{N2}^\top, X_{13}^\top, \dots, X_{NT}^\top)^\top \in \mathbb{R}^{NT \times D}$  with a similar grouping as above. Accordingly, the model in (2.1) can be expressed in matrix notation as follows

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

where  $B_0 = \{\beta_0(Z_{11}), \beta_0(Z_{21}), \dots, \beta_0(Z_{N1}), \beta_0(Z_{12}), \dots, \beta_0(Z_{NT})\}^\top \in \mathbb{R}^{NT \times D}$ ,  $e_D$  is a  $D \times 1$  vector of 1s, and “ $\circ$ ” denotes the Hadamard product. The  $NT \times 1$  vector of SAR disturbances  $u_N$  is specified below

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

where  $\otimes$  signifies the Kronecker product,  $W_N$  is an  $N \times N$  spatial weights matrix, which is assumed to be nonstochastic, and  $\rho_0$  is a scalar autoregressive parameter. Moreover,  $\varepsilon_N$  is an  $NT \times 1$  vector of innovations assumed to follow a classical one-way error component model as follows

$$\varepsilon_N = (e_T \otimes I_N)\alpha_N + v_N, \quad (2.4)$$

where  $\alpha_N$  denotes the vector of unit-specific error component,  $e_T$  is a  $T \times 1$  vector of 1s, and  $v_N$  is an  $NT \times 1$  vector of independently and identically distributed idiosyncratic errors.

Collectively, (2.2) to (2.4) specify the proposed VCPD model with a first-order SAR disturbance term, whose innovation has an error component structure. In this regard, we maintain a set of standard regularity assumptions as follows.

**Assumption A1.** (i)  $W_N$  is row-normalized, in which the elements in a given row sum up to one, and is a nonstochastic spatial weights matrix with zero diagonal elements. (ii) Let  $S_N(\rho) = I_N - \rho W_N$  for an arbitrary  $\rho \in \mathcal{P}$ , where  $\mathcal{P}$  is a compact parameter space, and  $\rho_0 \in (-1, 1)$  is in the interior of  $\mathcal{P}$ , and  $S_N(\rho)$  is invertible for all  $\rho \in \mathcal{P}$ . (iii)  $W_N$  and  $S_N^{-1}(\rho)$  are uniformly bounded in both row and column sums in their absolute value.

**Assumption A2.** For all  $it = 11, \dots, NT$ , (i) the idiosyncratic error component  $v_{it} \in v_N$  has a zero mean, a variance of  $\sigma_{v,0}^2$  and  $E(|v_{it}|^{2m}) < \infty$ , where  $m > 2$ ; (ii) the unit-specific error component  $\alpha_i \in \alpha_N$  has a zero mean, a variance of  $\sigma_{\alpha,0}^2$  and  $E(|\alpha_i|^{2m}) < \infty$ ; and (iii) the processes  $\{v_{it}\}$  and  $\{\alpha_i\}$  are independent of each other.

These assumptions directly concern our model specifications and are therefore explained below in some detail. Assumption A1 (i) implies that no unit is a neighbor to itself. Although the elements of  $W_N$  are assumed to be independent of  $t$ , the number of neighbors that a given unit has may depend on the number of cross-sectional units,  $N$ . Furthermore, Assumption A1 (iii) restricts the extent of the associations between the cross-sectional units. In practice, these conditions are satisfied, given that each unit is associated only with a limited number of neighbors implying that the weighting matrix  $W_N$  is sparse. Alternatively, when  $W_N$  is not sparse, the condition of Assumption A1 (iii) is satisfied if its elements decline with a distance measure that increases sufficiently rapidly as the sample size increases. The conditions in Assumptions A1 and A2 are standard for a random effects model, and are often used in the spatial econometric literature (see, e.g., Kapoor et al. (2007) and Baltagi et al. (2012)). Finally, these assumptions ensure that (2.3) can be rewritten as follows

$$u_N = [I_T \otimes (I_N - \rho_0 W_N)^{-1}] \varepsilon_N. \quad (2.5)$$

Expressions (2.4) and (2.5) suggest that our model specifications imply spatial interaction for both the error components,  $v_{it}$ , and the unit specific error components,  $\alpha_i$ . This differs significantly from the other available random effects models in the literature (see, e.g., Baltagi et al. (2003) and Baltagi et al. (2012)), and has the important benefit of obtaining the particular pattern of the variance matrices of the overall disturbances, for which the inverse can be relatively easier to compute.

## 2.2 Estimation procedure

The disturbances in (2.5) are such that the variance-covariance matrix  $E[u_N u_N^\top]$  is as follows

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

in which  $\Omega_{\varepsilon,N}^0 = \sigma_{v,0}^2 Q_{0,N} + \sigma_{1,0}^2 Q_{1,N} \equiv E[\varepsilon_N \varepsilon_N^\top]$ , where  $\sigma_{1,0}^2 = \sigma_{v,0}^2 + T \sigma_{\alpha,0}^2$ ,

$$Q_{0,N} = (I_T - (J_T/T)) \otimes I_N \text{ and } Q_{1,N} = (J_T/T) \otimes I_N$$

with  $J_T = e_T e_T^\top$  denoting a  $T \times T$  matrix of ones. In this regard,  $Q_{0,N}$  and  $Q_{1,N}$  are the standard transformation matrices used in the error component literature, which are symmetric, idempotent and orthogonal to each other.



Alternatively, the variance-covariance matrix of  $u_N$  can be written as  $\Omega_{u,N}^0 = \sigma_{v,0}^2 \mathbb{Q}_N^0$ ,  
 where  $\mathbb{Q}_N^0 = [I_T \otimes (I_N - \rho_0 W_N)^{-1}] \{Q_{0,N} + (1 + \phi_0 T) Q_{1,N}\} [I_T \otimes (I_N - \rho_0 W_N^\top)^{-1}]$  and  $\phi_0 =$   
 $\sigma_{\alpha,0}^2 / \sigma_{v,0}^2$ . In this regard,  $(\mathbb{Q}_N^0)^{-1} = \bar{\mathbb{Q}}_N^{0\top} \bar{\mathbb{Q}}_N^0$  for which

$$\bar{\mathbb{Q}}_N^0 = \left\{ Q_{0,N} + (1 + T\phi_0)^{-1/2} Q_{1,N} \right\} [I_T \otimes (I_N - \rho_0 W_N)]$$

because of the orthogonality of  $Q_{0,N}$  and  $Q_{1,N}$ . In other words,  $\mathbb{Q}_N^0 = (1/\sigma_{v,0}^2) E[u_N u_N^\top]$ .  
 By defining  $\ddot{X}_{0N} = \bar{\mathbb{Q}}_N^0 X_N$  and  $\ddot{u}_{0N} = \bar{\mathbb{Q}}_N^0 u_N$ , the transformation is

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

which can be viewed as the Cochrane-Orcutt-type random effects of the generalized least  
 squares transformations of (2.2), where  $\ddot{y}_{0N} = \bar{\mathbb{Q}}_N^0 y_N - \{\bar{\mathbb{Q}}_N^0 (B_0 \circ X_N) e_D - (B_0 \circ \bar{\mathbb{Q}}_N^0 X_N) e_D\}$   
 to incorporate the VC specification.

For a given vector of parameters  $\delta = (\phi, \rho)^\top \in \Delta$ , where  $\Delta$  denotes a compact parameter  
 space which is a necessary condition to establish the consistency of an QML estimator (see  
 Amemiya (1985), and Newey and McFadden (1994) for a comprehensive treatment of the  
 extremum estimation), these transformations enable the construction of the conditional  
 local kernel-weighted quasi-log-likelihood as follows

$$\begin{aligned} \ell_z(\beta, \sigma_v^2, \delta) &= -\frac{1}{2} \log\{2\pi\sigma_v^2\} \sum_{j=1}^N \sum_{s=1}^T K_h(Z_{js} - z) - \frac{1}{2NT} \log\{|\mathbb{Q}_N|\} \sum_{j=1}^N \sum_{s=1}^T K_h(Z_{js} - z) \\ &\quad - \frac{1}{2\sigma_v^2} \sum_{j=1}^N \sum_{s=1}^T \{\ddot{y}_{js} - \ddot{X}_{js}\beta\}^2 K_h(Z_{js} - z), \end{aligned} \quad (2.6)$$

where  $\ddot{X}_{js}$  and  $\ddot{y}_{js}$  are the  $js$ -th rows of  $\ddot{X}_N$  and  $\ddot{y}_N$  which are arbitrary expressions of  $\ddot{X}_{0N}$   
 and  $\ddot{y}_{0N}$ , respectively. Similarly,  $\mathbb{Q}_N$  and  $\bar{\mathbb{Q}}_N$  are arbitrary expressions of the terms defined  
 above, whereas  $K_h(\cdot) = K(\cdot/h)/h$ , where  $K(\cdot)$  and  $h$  denote the kernel function and the  
 associated bandwidth parameter, respectively.

Observe that the local likelihood function in (2.6) is maximized at

$$\tilde{\beta}(z; \delta) = \left[ \sum_{j=1}^N \sum_{s=1}^T \ddot{X}_{js}^\top \ddot{X}_{js} K_h(Z_{js} - z) \right]^{-1} \sum_{j=1}^N \sum_{s=1}^T \ddot{X}_{js}^\top \ddot{y}_{js} K_h(Z_{js} - z) \quad (2.7)$$

and

$$\tilde{\sigma}_v^2(z; \delta) = \left[ \sum_{j=1}^N \sum_{s=1}^T K_h(Z_{js} - z) \right]^{-1} \left[ \sum_{j=1}^N \sum_{s=1}^T \{\ddot{y}_{js} - \ddot{X}_{js}\tilde{\beta}(z; \delta)\}^2 K_h(Z_{js} - z) \right]. \quad (2.8)$$

These suggest that we can formulate the concentrated log-likelihood, particularly  $\tilde{\ell}_z^c(\delta) \equiv \max_{\beta, \sigma_v^2} \ell_z(\beta, \sigma_v^2, \delta)$ , then the global one is as follows

$$\tilde{\mathcal{L}}^c(\delta) = -\frac{1}{2} \sum_{i=1}^N \sum_{t=1}^T \left[ \log(2\pi\tilde{\sigma}_v^2) + \frac{\log |\mathbb{Q}_N|}{NT} + 1 \right] \sum_{j=1}^N \sum_{s=1}^T K_h(Z_{js} - Z_{it}), \quad (2.9)$$

where

$$\tilde{\sigma}_v^2 = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \tilde{\sigma}_v^2(Z_{it}; \delta),$$

which complies with Assumption A2(i). Accordingly,  $\hat{\delta}$  that maximizes  $\tilde{\mathcal{L}}^c(\delta)$  in (2.9) is the QML estimator of  $\delta_0$ .

Establishing the uniform consistency of the proposed QML estimator requires the following additional assumptions.

**Assumption B1.** (i) The higher-order kernel function  $K(u)$  is absolutely continuous and integrable on its compact support, where the order of  $K(u)$  is  $\varrho > 2$ . Additionally,  $K^2(u)$  and  $(K'(u))^2$ , where  $K'(u)$  denotes the first derivative of  $K(u)$ , are finite and integrable on its support. (ii) The bandwidth parameter is any monotonic sequence of  $N$  such that  $\lim_{N \rightarrow \infty} h \rightarrow 0$  and  $\lim_{N \rightarrow \infty} N^{2\eta-1}h \rightarrow \infty$ , where  $\frac{1}{2} + \frac{1}{4\varrho} < \eta < 1 - \frac{1}{m}$ .

**Assumption B2.** For all  $it = 11, \dots, NT$ , (i)  $E|Y_{it}|^{2m} < \infty$ , and (ii)  $E||X_{it}||^{2m} < \infty$ .

**Assumption B3.** For all  $it = 11, \dots, NT$ , (i)  $\Omega(z) = E(X_{it}^\top X_{it} | Z_{it} = z)$  is nonsingular and its higher-order derivative is bounded, and  $E(||X_{it}||^{2m} | Z_{it} = z) < \infty$ ; and (ii) the higher-order derivative of  $\sigma_v^2(z) = E(v_{it}^2 | Z_{it} = z)$  is also bounded.

**Assumption B4.** For all  $it = 11, \dots, NT$ , the higher-order derivative of the marginal density  $f(z)$  of  $Z_{it}$  is bounded and  $\inf_{z \in [0,1]} f(z) > 0$ .

**Assumption B5.** For all  $d = 1, \dots, D$ , the higher-order derivative of  $\beta_{0,d}(z)$  is continuous.

Assumptions B1 to B5 are the standard regularity conditions often used in the semi- and nonparametric literature, particularly by Mack and Silverman (1982), Robinson (1988), Fan and Li (1999), Cai et al. (2000), Fan and Zhang (2000), Xia et al. (2004), and Robinson (2011). Since these are mainly technical conditions, their justification and discussion are presented in Appendix B.

Furthermore, the identification condition of the QLM estimation also requires the expected local log-likelihood function to have an identifiably unique maximizer that converges to  $\delta_0$  as  $N \rightarrow \infty$ . In this regard, Lemma A.1 in Appendix A shows that the lower bound of  $E(\tilde{\ell}_z^c(\delta))$  is

$$\bar{\ell}_z^c(\delta) \equiv -\frac{1}{2} \left[ \log\{2\pi\} + \log \left\{ \sigma_v^2(z) \frac{\text{tr}[\mathbb{Q}_{0N} \bar{\mathbb{Q}}_N^\top \bar{\mathbb{Q}}_N]}{NT} \right\} + \frac{\log |\mathbb{Q}_N|}{NT} + 1 \right] f(z), \quad (2.10)$$

where  $\text{tr}[\cdot]$  denotes the trace of a matrix. Accordingly, we assume the global identification condition as follows

$$\limsup_{N \rightarrow \infty} \left\{ \max_{\delta \in \bar{D}_\epsilon(\delta_0) \cap \Delta} \bar{\ell}_z^c(\delta) \right\} \neq \limsup_{N \rightarrow \infty} \bar{\ell}_z^c(\delta_0) \quad (2.11)$$

for any  $\delta$ , where  $\bar{D}_\epsilon(\delta_0)$  is the complement of the  $\epsilon$ -neighborhood of  $\delta_0$ . Such a condition is in line with the extremum estimation literature (see Amemiya (1985), Newey and McFadden (1994), White (1996), Lee (2004), and Liu and Yang (2015)). Theorem 2.1 below establishes the uniform consistency of our proposed QML estimator.

**Theorem 2.1.** *Let Assumptions A and B hold. Under the identification condition in (2.11),  $\delta_0$  is uniquely identifiable and  $\sup_{z \in [0,1]} \|\hat{\delta} - \delta_0\| = O_P((NT)^{-1/2})$  as  $N \rightarrow \infty$ .*

Subsequently, an estimation of the unknown  $\beta_0(z)$  and  $\sigma_{v,0}^2(z)$  can be formulated as follows

$$\hat{\beta}(z; \hat{\delta}) = \left[ \sum_{j=1}^N \sum_{s=1}^T \hat{X}_{js}^\top \hat{X}_{js} K_h(Z_{js} - z) \right]^{-1} \sum_{j=1}^N \sum_{s=1}^T \hat{X}_{js}^\top \hat{y}_{js} K_h(Z_{js} - z), \quad (2.12)$$

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

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

Under the assumption that the variance of the idiosyncratic errors does not depend on the location of  $z$ , particularly Assumption A2(i),  $\sigma_{v,0}^2$  can be estimated by

$$\hat{\sigma}_v^2 = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \hat{\sigma}_v^2(Z_{it}; \hat{\delta}). \quad (2.13)$$

For the sake of notational simplicity, let  $\hat{\beta}(z) \equiv \hat{\beta}(z; \hat{\delta})$  and  $\hat{\sigma}_v^2(z) \equiv \hat{\sigma}_v^2(z; \hat{\delta})$ , which are hereafter referred to as unpenalized estimators, whose uniform consistency is established in Theorem 2.2.

256 **Theorem 2.2.** *Under Assumptions A and B, we have*

$$\sup_{z \in [0,1]} \|\hat{\beta}(z) - \beta_0(z)\| = O_P \left( h^q + \left( \frac{\log(1/h)}{(NT)h} \right)^{1/2} \right)$$

257 *and*

$$\sup_{z \in [0,1]} |\hat{\sigma}_v^2(z) - \sigma_{v,0}^2(z)| = O_P \left( h^q + \left( \frac{\log(1/h)}{(NT)h} \right)^{1/2} \right),$$

258 *where  $q$  denotes the order of the kernel function, as  $N \rightarrow \infty$ .*

## 259 2.3 SAREC-KLASSO method

260 This section generalizes the variable selection technique considered in Wang and Xia (2009)  
 261 (see also Hu and Xia (2012)) to the VCPD model discussed in Section 2.1, in which  
 262 spatial interactions are involved in both the error components. This generalized method  
 263 is referred to hereafter as the SAR-error-component kernel LASSO or SAREC-KLASSO  
 264 method. For all  $it = 11, \dots, NT$ , we assume that there is an integer  $D_0 \leq D$  for which  
 265  $0 < E\{\beta_{d,0}^2(Z_{it})\} < \infty$  for any  $d \leq D_0$  and  $E\{\beta_{d,0}^2(Z_{it})\} = 0$  for  $D_0 + 1 \leq d \leq D$  without  
 266 loss of generality. This implies that there are  $D_0$  regressors that are truly relevant, whereas  
 267 the rest are not. Correspondingly, let us also define  $X_{ita}^\top = \{X_{it,1}, \dots, X_{it,D_0}\}^\top \in \mathbb{R}^{D_0}$  and  
 268  $X_{itb}^\top = \{X_{it,D_0+1}, \dots, X_{it,D}\}^\top \in \mathbb{R}^{D-D_0}$ .

269 Let  $B = \{b_1, \dots, b_{D_0}, b_{D_0+1}, \dots, b_D\}$ , where  $b_d = \{\beta_d(Z_{11}; \delta), \dots, \beta_d(Z_{NT}; \delta)\}^\top \in \mathbb{R}^{NT}$  is  
 270 the  $d$ -th column of  $B$ . In consequence, the last  $(D - D_0)$  columns of  $B$ , which are associated  
 271 with  $X_{itb}$ , should be 0, and the task of variable selection is equivalent to identifying these  
 272 sparse columns in the matrix,  $B$ . Following the group LASSO idea of Yuan and Lin (2006)  
 273 (see also Wang and Xia (2009)), this identification can be achieved by utilizing the penalized  
 274 estimation as follows

$$\tilde{B}_\lambda = \underset{B \in \mathbb{R}^{NT \times D}}{\operatorname{argmin}} \tilde{Q}_\lambda(B; \delta)$$

275 in which

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

276 where  $\lambda = (\lambda_1, \dots, \lambda_D)^\top \in \mathbb{R}^D$  is the vector of the tuning parameters and  $\|\cdot\|$  signifies the  
 277 usual Euclidean norm.

278 In this regard, the solutions for the penalized estimation are

$$\tilde{B}_\lambda = \{\tilde{\beta}_\lambda(Z_{11}), \dots, \tilde{\beta}_\lambda(Z_{N1}), \tilde{\beta}_\lambda(Z_{12}), \dots, \tilde{\beta}_\lambda(Z_{NT})\}^\top \equiv (\tilde{b}_{\lambda,1}, \dots, \tilde{b}_{\lambda,D_0}, \tilde{b}_{\lambda,D_0+1}, \dots, \tilde{b}_{\lambda,D}).$$

279 Accordingly, the  $it$ -th row of  $\tilde{B}_\lambda$  is defined as the transpose of

$$\tilde{\beta}_\lambda(Z_{it}; \delta) = \left[ \sum_{j=1}^N \sum_{s=1}^T \ddot{X}_{js}^\top \ddot{X}_{js} K_h(Z_{js} - Z_{it}) + \mathfrak{D} \right]^{-1} \sum_{j=1}^N \sum_{s=1}^T \ddot{X}_{js}^\top \ddot{y}_{js} K_h(Z_{js} - Z_{it}), \quad (2.14)$$

280 where  $\mathfrak{D} = \text{diag}(\lambda_1/\|b_1\|, \dots, \lambda_D/\|b_D\|)$ . In addition,

$$\begin{aligned} \tilde{\sigma}_{\lambda,v}^2(Z_{it}; \delta) &= \left[ \sum_{j=1}^N \sum_{s=1}^T K_h(Z_{js} - Z_{it}) \right]^{-1} \\ &\quad \times \left[ \sum_{j=1}^N \sum_{s=1}^T \{\ddot{y}_{js} - \ddot{X}_{js} \tilde{\beta}_\lambda(Z_{it}, \delta)\}^2 K_h(Z_{js} - Z_{it}) \right]. \end{aligned} \quad (2.15)$$

281 These estimators are essentially the penalized counterparts of those in (2.7) and (2.8).

282 Consequently, they can be used for constructing the concentrated log-likelihood in a similar  
283 manner to the method in Section 2.2. In particular, we have

$$\tilde{\mathcal{L}}_\lambda^c(\delta) = -\frac{1}{2} \sum_{i=1}^N \sum_{t=1}^T \left[ \log(2\pi \tilde{\sigma}_{\lambda,v}^2) + \frac{\log |\mathbb{Q}_N|}{NT} + 1 \right] \sum_{j=1}^N \sum_{s=1}^T K_h(Z_{js} - Z_{it}) \quad (2.16)$$

284 which is the penalized counterpart of (2.9). In this regard, maximizing (2.16) leads to an  
285 alternative QML estimator of  $\delta_0$ , which is denoted hereafter as  $\hat{\delta}_\lambda$  which signifies that its  
286 computation is based on the penalized estimators of (2.14) and (2.15).

287 Finally, the penalized estimator of  $B_0$  is

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

288 in which

$$\hat{Q}_\lambda(B; \hat{\delta}_\lambda) = \sum_{i=1}^N \sum_{t=1}^T \sum_{j=1}^N \sum_{s=1}^T \left\{ \hat{y}_{js} - \hat{X}_{js} \beta_{it} \right\}^2 K_h(Z_{js} - Z_{it}) + \sum_{d=1}^D \lambda_d \|b_d\|. \quad (2.18)$$

289 In other words, the  $it$ -th row of  $\hat{B}_\lambda$  is defined as the transpose of

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

To discuss the asymptotic properties of the penalized estimators requires maintaining some additional conditions on the amount of shrinkage being applied (see Wang and Xia (2009) and Hu and Xia (2012) for more details) as follows.

**Assumption C1.** For  $a_N = \max\{\lambda_d : 1 \leq d \leq D_0\}$  and  $b_N = \min\{\lambda_d : D_0 + 1 \leq d \leq D\}$ , assume that  $N^{\frac{4\varrho+1}{4\varrho}} a_N \rightarrow 0$  and  $N^{\frac{4\varrho+1}{4\varrho}} b_N \rightarrow \infty$ , as  $N \rightarrow \infty$ .

Let  $\hat{\beta}_\lambda(z) \equiv \hat{\beta}_\lambda(z; \hat{\delta}_\lambda)$  for notational convenience and  $\hat{\beta}_\lambda(z) = \{\hat{\beta}_{\lambda,a}(z), \hat{\beta}_{\lambda,b}(z)\}^\top$ , where  $\hat{\beta}_{\lambda,a}(z) \equiv \{\hat{\beta}_{\lambda,1}(z), \dots, \hat{\beta}_{\lambda,D_0}(z)\}^\top$  and  $\hat{\beta}_{\lambda,b}(z) \equiv \{\hat{\beta}_{\lambda,D_0+1}(z), \dots, \hat{\beta}_{\lambda,D}(z)\}^\top$ . Corollary 2.1 suggests that the true model can be consistently selected as long as the tuning parameters satisfy Assumption C1. Since it is associated with  $D_0$ ,  $\hat{\beta}_a(z)$  can be viewed as the oracle estimators. In this regard, Corollary 2.2 suggests that the uniform convergence of the penalized estimator can be achieved as long as Assumption C1 is satisfied.

**Corollary 2.1.** Let Assumptions A to C hold. Then  $P\left(\sup_{z \in [0,1]} \|\hat{\beta}_{\lambda,b}(z)\| = 0\right) \rightarrow 1$  as  $N \rightarrow \infty$ .

**Corollary 2.2.** Under Assumptions A to C, as  $N \rightarrow \infty$ ,

$$\sup_{z \in [0,1]} \|\hat{\beta}_{\lambda,a}(z) - \hat{\beta}_a(z)\| = o_P\left(h^\varrho + \left(\frac{\log(1/h)}{(NT)h}\right)^{1/2}\right),$$

where

$$\hat{\beta}_a(z) = \left[ \sum_{j=1}^N \sum_{s=1}^T \hat{X}_{jsa}^\top \hat{X}_{jsa} K_h(Z_{js} - z) \right]^{-1} \sum_{j=1}^N \sum_{s=1}^T \hat{X}_{jsa}^\top \hat{y}_{js} K_h(Z_{js} - z).$$

To apply our method in practice, there remains some implementational issues that must be clarified. These are (i) selection of the shrinkage parameters and (ii) a local quadratic approximation of the penalty function.

### Selection of shrinkage parameters

Selecting of up to  $D$  shrinkage parameters, namely  $\lambda_1, \dots, \lambda_D$ , is not straightforward. To overcome this difficulty, we follow an idea used in various studies, which is to specify

$$\lambda_d = \frac{\lambda_0}{(NT)^{-1/2} \|\hat{b}_d\|} \quad (2.19)$$

for which  $\lambda_0 > 0$  and  $\hat{b}_d$  is the  $d$ -th column of  $\hat{B} = \{\hat{\beta}(Z_{11}), \dots, \hat{\beta}(Z_{NT})\}^\top \in \mathbb{R}^{NT \times D}$  (see e.g., Zou (2006) and Wang and Leng (2007)). To provide some technical properties of  $\lambda_d$ , we put forward the following statements, which are obtained directly from the results of Section 2.2,

$$(NT)^{-1/2} \|\hat{b}_d\| = \left\{ (NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T \hat{\beta}_d^2(Z_{it}) \right\}^{1/2} \rightarrow_P \{E[\beta_d^2(Z_{it})]\}^{1/2} \text{ for } 1 \leq d \leq D_0 \quad (2.20)$$

and

$$\left\{ (NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T \hat{\beta}_d^2(Z_{it}) \right\}^{1/2} = O_P\{(NT)^{\frac{1-2\varrho}{4\varrho}}\} \text{ for } D_0 + 1 \leq d \leq D. \quad (2.21)$$

Whereas (2.20) suggests  $\lambda_d$  converges to a positive constant for  $1 \leq d \leq D_0$ , (2.21) implies that  $\lambda_d$  diverges for  $D_0 + 1 \leq d \leq D$ . Hence, to maintain  $(NT)^{\frac{4\varrho+1}{4\varrho}} a_N \rightarrow 0$  and  $(NT)^{\frac{4\varrho+1}{4\varrho}} b_N \rightarrow \infty$  we require that  $\lambda_0(NT)^{\frac{4\varrho+1}{4\varrho}} \rightarrow 0$  and  $\lambda_0(NT)^{\frac{\varrho+1}{2\varrho}} \rightarrow \infty$ , respectively.

Clearly, the specification in (2.19) helps us to reduce the original  $D$ -dimensional problem of  $\lambda \in \mathbb{R}^D$  into a univariate problem of selecting  $\lambda_0 > 0$ . In practice, such a selection is done by minimizing the following BIC-type criterion

$$\text{BIC}_\lambda = \log\{RSS_\lambda\} + df \times \frac{\log\{(NT)h\}}{(NT)h}, \quad (2.22)$$

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

$$RSS_\lambda = (NT)^{-2} \sum_{i=1}^N \sum_{t=1}^T \sum_{j=1}^N \sum_{s=1}^T \left\{ \hat{y}_{js} - \hat{X}_{js} \hat{\beta}_\lambda(Z_{it}) \right\}^2 K_h(Z_{js} - Z_{it}). \quad (2.23)$$

Let  $\hat{B}_{\hat{\lambda}}$  denote a penalized estimator in (2.17), which corresponds to  $\hat{\lambda} = \underset{\lambda}{\operatorname{argmin}} \text{BIC}_\lambda$ ,  $\mathcal{S}_{\hat{\lambda}}$  represent the model identified by  $\hat{B}_{\hat{\lambda}}$ , and  $\mathcal{S}_T = \{1, \dots, D_0\}$  denote the true model. Corollary 2.3 below states that the estimate of the tuning parameter  $\hat{\lambda}$  selected by the BIC criterion is able to consistently identify the true model.

**Corollary 2.3.** *Let Assumptions A to C hold. Then, as  $N \rightarrow \infty$ ,*

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

*Remark 2.1.* A final point to clarify regarding the variable selection procedure is the use of  $\hat{y}_{it}$  and  $\hat{X}_{it}$  in the calculation of  $RSS_\lambda$  in (2.23). In this regard, the conceptual discussion in Section 2.2 suggests that we can rely on the following steps: (i) computing

the spatial estimates of  $\delta = (\phi, \rho)^\top$  by maximizing the concentrated log-likelihood under the unpenalized estimation in (2.7) and (2.8), (ii) computing  $\hat{y}_{it}$  and  $\hat{X}_{it}$ , and then (iii) applying the SAREC-KLASSO method discussed in this section.

### Local quadratic approximation of the penalty function

This paper extends a local quadratic approximation of the penalty function in the spirit of Hunter and Li (2005) to the VCPD model in Section 2.1. More specifically, in practice, the computation follows an iterative algorithm where the loss function in (2.18) is approximated locally by

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

where  $\hat{B}_\lambda^{(m)} = \left\{ \hat{\beta}_\lambda^{(m)}(Z_{11}), \hat{\beta}_\lambda^{(m)}(Z_{21}), \dots, \hat{\beta}_\lambda^{(m)}(Z_{NT}) \right\}^\top \equiv (\hat{b}_{\lambda,1}^{(m)}, \hat{b}_{\lambda,2}^{(m)}, \dots, \hat{b}_{\lambda,D}^{(m)})$ , denoting the estimates obtained in the  $m$ -th iteration. The resulting minimizer, which hereafter is denoted by  $\hat{B}_\lambda^{(m+1)}$ , is such that its  $it$ -th row is defined as the transpose of

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

where  $\hat{\mathcal{D}}^{(m)} = \text{diag}(\lambda_1/\|\hat{b}_{\lambda,1}^{(m)}\|, \dots, \lambda_D/\|\hat{b}_{\lambda,D}^{(m)}\|)$ . We discuss the evolution of  $\hat{\beta}_\lambda^{(m+1)}(z)$  as  $m \rightarrow \infty$  in Corollaries 2.4 and 2.5 below.

**Corollary 2.4.** *Let Assumptions A to C hold and  $\hat{\beta}_{\lambda,b}^{(m+1)}(z) = (\hat{\beta}_{\lambda,D_0+1}^{(m+1)}(z), \dots, \hat{\beta}_{\lambda,D}^{(m+1)}(z))^\top$ . Then,  $P \left( \sup_{z \in [0,1]} \|\hat{\beta}_{\lambda,b}^{(m+1)}(z)\| = 0 \right) \rightarrow 1$  as  $N \rightarrow \infty$ .*

**Corollary 2.5.** *Let Assumptions A to C hold and  $\hat{\beta}_{\lambda,a}^{(m+1)}(z) = (\hat{\beta}_{\lambda,1}^{(m+1)}(z), \dots, \hat{\beta}_{\lambda,D_0}^{(m+1)}(z))^\top$ . Then,  $\sup_{z \in [0,1]} \|\hat{\beta}_{\lambda,a}^{(m+1)}(z) - \hat{\beta}_a(z)\| = o_P \left( h^q + \left( \frac{\log(1/h)}{(NT)h} \right)^{1/2} \right)$  as  $N \rightarrow \infty$ .*



## 2.4 Constant coefficients in a semi-varying coefficient model

After performing the variable selection, it is useful to identify the constant coefficient(s) among those deemed to be associated with the relevant regressors. Importantly, such an identification helps us to generalize the scope of our method from the VCPD specification to a more general one often referred to as the “semi-varying coefficient”. Below we discuss the procedure of testing the hypothesis of a constant coefficient within the context of the spatial VCPD model. Let  $\hat{D}$  denote the number of relevant regressors identified by  $\hat{B}_{\hat{\lambda}}$ , and  $C_d$  be a constant. Then the hypotheses can be written as follows

$$H_0 : \beta_{d,0}(z) = C_d \text{ versus } H_1 : \beta_{d,0}(z) \neq C_d, \quad 1 \leq d \leq \hat{D}.$$

Before discussing the test statistic, we introduce a result that concerns two unpenalized estimators, namely  $\tilde{\beta}(z; \delta_0)$  and  $\hat{\beta}(z; \hat{\delta})$ , which associate with the true spatial parameters and their QML estimates, respectively. Recall that these are the estimators defined in (2.7) and (2.12), respectively. Corollary 2.6 below ensures that the difference between these two unpenalized estimators is uniformly negligible. This result has at least two important implications.

**Corollary 2.6.** *Under the conditions of Theorem 2.2,*

$$\sup_{z \in [0,1]} \|\hat{\beta}(z; \hat{\delta}) - \tilde{\beta}(z; \delta_0)\| = o_P \left( h^q + \left( \frac{\log(1/h)}{(NT)h} \right)^{1/2} \right) \text{ as } N \rightarrow \infty.$$

Firstly, the uniform convergence that can be achieved by the former is also plausible for the latter. This suggests that most inference methods for  $\tilde{\beta}(z; \delta_0)$  are also applicable for  $\hat{\beta}(z; \hat{\delta})$ . In this section, we make use a result obtained in Fan and Zhang (2000), which states the asymptotic distribution of the maximized normalized-deviations between the estimated coefficient functions and the true functions. In particular:

$$P \left\{ (-2 \log h)^{1/2} \left( \sup_{z \in [0,1]} \|\{\widehat{\text{var}}(\hat{\beta}(z))\}^{-1/2} (\hat{\beta}(z) - \beta_0(z) - \widehat{\text{bias}}(\hat{\beta}(z)))\| - d_N \right) < x \right\} \rightarrow \exp\{-2 \exp(-x)\}, \quad (2.24)$$

where  $d_N = (-2 \log h)^{1/2} + \frac{1}{(-2 \log h)^{1/2}} \log \left\{ \frac{1}{4\kappa\pi} \int \{K'(u)\}^2 du \right\}$  with  $\kappa = \int K^2(u) du$ ,

$$\widehat{\text{var}}(\hat{\beta}(z)) = \left\{ \hat{X}_N^\top K_N \hat{X}_N \right\}^{-1} \hat{X}_N^\top K_N^2 \hat{X}_N \left\{ \hat{X}_N^\top K_N \hat{X}_N \right\}^{-1} \hat{\sigma}_v^2$$

such that  $K_N = \text{diag}\{K_h(Z_{11} - z), \dots, K_h(Z_{N1} - z), K_h(Z_{12} - z), \dots, K_h(Z_{NT} - z)\}$  and  $\hat{\sigma}_v^2$  is defined in (2.13), and

$$\widehat{\text{bias}}(\hat{\beta}(z)) \approx \left\{ \hat{X}_N^\top K_N \hat{X}_N \right\}^{-1} \hat{X}_N^\top K_N \hat{a}_N$$

for  $\hat{a}_{it} = \left\{ \hat{\beta}^{(\varrho+1)}(z)(Z_{it} - z) + \frac{\hat{\beta}^{(\varrho+2)}(z)(Z_{it} - z)^2}{2} \right\} \hat{X}_{it}$ .

The result in (2.24) can be used for finding the null distribution of the test statistic. That is

$$T_d = (-2 \log h)^{1/2} \left[ \sup_{z \in [0,1]} \left| \left\{ \widehat{\text{var}}(\hat{\beta}_d(z)) \right\}^{-1/2} (\hat{\beta}_d(z) - \hat{C}_d - \widehat{\text{bias}}(\hat{\beta}_d(z))) \right| - d_N \right], \quad (2.25)$$

where  $\widehat{\text{var}}(\hat{\beta}_d(z)) = e_{j,d}^\top \widehat{\text{var}}(\hat{\beta}(z)) e_{j,d}$ ,  $\widehat{\text{bias}}(\hat{\beta}_d(z)) \approx e_{j,d}^\top \widehat{\text{bias}}(\hat{\beta}(z))$ , for which  $e_{j,d}$  denotes the unit vector of length  $d$  with 1 at position  $j$ , and  $\hat{C}_d = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \hat{\beta}_d(Z_{it})$ . In this regard, the  $H_0$  is rejected when the test statistic exceeds the asymptotic critical value of  $c_\alpha = -\log(-0.5 \log \alpha)$ . A similar argument and procedure were used in Fan and Zhang (2000), and Wang and Xia (2009).

Furthermore, by using the procedure introduced in Section 2.2, under the conditions of Theorem 2.2, Corollary 2.6 ensures that the asymptotic behaviors of the coefficient constancy test statistic, in Fan and Zhang (2000) and Fan and Zhang (2008), are also valid for ours. In particular, the power of our test is nontrivial and approaching one as the number of observations increases, while the size approaches the significance level. These asymptotic behaviors are stated more formally in Appendix B in order to keep the discussion straightforward, and are empirically illustrated in the next section.

### 3 Simulation studies

This section presents a simulation exercise that examines finite-sample performances of the procedures considered in the previous section. These procedures are (i) estimation of the spatial parameters  $\delta_0 = (\phi_0, \rho_0)^\top$  based on the concentrated likelihood, (ii) estimation of the coefficient functions  $\beta_0(z) = \{\beta_{0,1}(z), \dots, \beta_{0,D}(z)\}^\top$  by applying the unpenalized and penalized estimations, (iii) variable selection by employing the KLASO and SAREC-KLASO methods, and (iv) hypothesis testing of the coefficient constancy.

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

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

The difference between these models lies in the fact that the former includes no constant coefficients, whereas the latter includes two nonzero constant coefficients. Therefore, Model II is an example of what known in the literature as the semi-varying coefficient model.

For each component in these models,  $X_{it,1}$  is set to 1, whereas  $(X_{it,2}, \dots, X_{it,7})^\top$  and  $Z_{it}$  are generated from the multivariate normal distribution with  $\text{Cov}(X_{it,d_1}, X_{it,d_2}) = 0.5^{|d_1-d_2|}$  for any  $2 \leq d_1, d_2 \leq 7$ , and from the uniform distribution  $U[0, 1]$ , respectively. Moreover, the disturbance satisfies the spatial interactions implied by the specifications discussed in Section 2.1 with  $\rho_0 = 0.3$  and  $\sigma_{\alpha,0}^2 = \sigma_{v,0}^2 = 1$ . Regarding the spatial weight matrices, we use matrices that differ in their degree of sparseness. As in Kelejian and Prucha (1999), we refer to these as “ $P$ -ahead-and- $P$ -behind”, where the spatial association is constructed with three values of  $P$ , namely  $P = 2$ ,  $P = 5$ , and  $P = 8$ , which lead to 4, 10, and 16 nonzero elements in a given row, respectively.

In the computation, we compute the inverse and determinant of matrix  $\mathbb{Q}_N$  based on  $(\mathbb{Q}_N)^{-1} = (1/T)J_T \otimes C_1^{-1} + \{I_T - (1/T)J_T\} \otimes C_2^{-1}$  and  $|\mathbb{Q}_N| = |C_1||C_2|^{T-1}$ , respectively, where  $C_1 = (1 + \phi T)C_2$  and  $C_2 = \{(I_N - \rho W_N)^\top (I_N - \rho W_N)\}^{-1}$ . These follow the results of Magnus and Muris (2010) and can help to alleviate the serious computational burden caused by repeated evaluations of a  $NT \times NT$  matrix during the optimization process. Moreover, we select the optimal bandwidth by applying the leave-one-unit-out cross-validation method. Even though the selection is done within the context of the unpenalized estimation, the selected bandwidth is also used in the penalized estimation. A similar leave-one-unit-out cross-validation method is also used in Li et al. (2011), Chen et al. (2012), and Liang et al. (2022). Furthermore, the optimal shrinkage parameter is selected based on the BIC criterion introduced in (2.22), whereas the total number of iterations of the iterative algorithm in Section 2.3 is set at 15. In the simulation exercise that follows, a total of 200 repetitions are conducted for each of the model setups.

The results of simulation are summarized in Tables 1 to 3 in Appendix C. Below, we discuss important findings for each of the procedures in turn.

## Spatial Estimation: SAR 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  $\rho_0$  computed by maximizing the concentrated log-likelihood under the (i) oracle, (ii) unpenalized, and (iii) penalized estimations, respectively. Similarly,  $\hat{\phi}_{or}$ ,  $\hat{\phi}_{un}$ , and  $\hat{\phi}_{\lambda}$  are those of the variance ratios of  $\phi_0 = \sigma_{\alpha,0}^2/\sigma_{v,0}^2$ . For the sake of comparison, two measures of accuracy, namely the mean absolute error (MAE) and root mean squared error (RMSE), are considered. Although the RMSE closely resembles a standard definition that is often seen in the literature, it is based on quantiles, which exist with certainty, rather than moments (see e.g., Kapoor et al. (2007)). In particular, we compute

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

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

The results in the tables show that  $\hat{\rho}_{or}$  and  $\hat{\rho}_{un}$  perform almost equally well when  $N$  is small. Although MAE and RMSE for  $\hat{\rho}_{\lambda}$  converge to zero as  $N$  increases, the estimator does not perform as well as the oracle and unpenalized based counterparts at small  $N$ . However, all the three estimators of the spatial parameter perform almost equally well at larger  $N$ . Regarding the estimators of the variance ratio, it is clear that  $\hat{\phi}_{or}$  performs the best. Unlike the estimators of the spatial parameter, here,  $\hat{\phi}_{\lambda}$  performs much better than its unpenalized counterpart. These results are not surprising, since the oracle and penalized estimations provide much more accurate estimates of the coefficient functions, as discussed below. At  $N = 300$ ,  $\hat{\phi}_{\lambda}$  performs almost as well as the oracle-based counterpart. Furthermore, an increase in  $P$ , which leads to a higher number of nonzero elements in a given row of the weighting matrix, provides less accurate estimation of both the spatial parameter and variance ratio. However, the former seems to be more significantly affected. Finally, similar results are obtained for both of the models.

## Nonparametric estimation of the coefficient functions

To compare the accuracy of the penalized estimators with that of its unpenalized and oracle counterparts, we compute the relative estimation error (REE) as follows

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

where  $\vartheta_d(Z_{it})$  is either the oracle or unpenalized estimator, namely  $\hat{\beta}_{or,d}(Z_{it})$  or  $\hat{\beta}_d(Z_{it})$ .

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

## Variable selection

We now discuss the finite-sample performance of the SAREC-KLASSO procedure for selecting relevant regressors. It is useful to note that the vector of relevant regressors is  $X_{ita}^\top = \{X_{it,1}, X_{it,2}\}^\top$  for Model I, whereas it is  $X_{ita}^\top = \{X_{it,1}, \dots, X_{it,4}\}^\top$  for Model II, so that  $D_0 = 2$  and  $D_0 = 4$ , respectively. Table 4 presents the percentages of the simulation repetitions where the SAREC-KLASSO procedure is able to obtain the correct number of relevant regressors and is also capable of accurately selecting the regressors in question. Importantly, these results illustrate that the performance of our procedure is not affected by the fact that Model II contains constant coefficients. This finding paves the way for identifying constant coefficients in a semi-varying coefficient model by using the procedure introduced in Section 2.4. A higher number of nonzero coefficients leads to better finite-sample performance at smaller  $N$ . Nonetheless, the results for the two models converge when  $N$  increases to 300. The finite-sample performance of our selection procedure seems to worsen as  $P$  increases, but only marginally. This likely reflects the performance of the

spatial estimation, which was discussed in the previous section. Finally, it is imperative to note that the KLASSO procedure is incapable of operating under models associated with spatially correlated error components.

### Hypothesis testing of the coefficient constancy

We now examine the finite-sample performance of two versions of Fan and Zhang's (2000) hypothesis testing procedure of the coefficient constancy, namely (a) their original version, and (b) our extension incorporating the spatial interactions and random effects in order to obtain efficiency gain. To allow an investigation into the ability of the test to reject an untrue null hypothesis, we assume that  $y_{it}$  is generated in accordance with Model III below

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

where other details of the underlying data generating process are as previously specified.

Table 5 shows the percentage of correct rejections and non-rejections in the 200 replications of the two testing methods, namely *with* and *without* incorporating the spatial interactions and random effects. The null hypothesis of a constant coefficient is easily rejected for  $\beta_{0,1}(z)$ , where the percentages of rejections reach 100% even for  $N = 100$ . For  $\beta_{0,2}(z)$  and  $\beta_{0,3}(z)$ , addressing the SED and modeling the heterogeneity with random effects lead to efficiency gain which clearly helps to improve the percentages. These results suggest that the power values of the test depend on two factors. The first is the degree of nonlinearity of the coefficient functions. Note that  $\beta_{0,1}(z)$  in Model III demonstrates a much stronger nonlinearity than  $\beta_{0,2}(z)$  and  $\beta_{0,3}(z)$ . Secondly, the SED and heterogeneity must be appropriately considered for the inference methods to work effectively. The impacts of these are evident at all levels of observations and numbers of nonzero elements in the weighting matrices. On the contrary, the null hypothesis of a constant coefficient is true when tests are implemented for  $\beta_{0,4}(z)$  to  $\beta_{0,7}(z)$ . Hence, we can subtract the proportion shown in Table 5 from 100 results for the size of the test. From Table 5, it is obvious that the size converges to the 5% significance level in all cases. In addition, addressing the SED and modeling the heterogeneity with random effects lead to significant improvements.

## 4 Public expenditure on mental health in England

In this section, we analyze the determinants of the spending on MHS by local councils in England. In addition, we investigate whether different political preferences of residents within local authorities bring about disparities that can lead to unequal access to MHS.

### 4.1 Introduction and motivation

The issue of mental health is one of the emerging policy priorities in the UK. We begin by highlighting some notable facts, illustrating the prevalence of the mental health problems in the UK.

- The 2014 survey of Mental Health and Wellbeing in England found that 1 in 6 people aged 16+ had experienced symptoms of a common mental health problem, such as depression or anxiety, in the past week (see McManus et al. (2016) for more details). Women were more likely than men to experience common mental disorders.

- Monitoring from the Office for National Statistics (ONS) found that the prevalence of moderate or severe depressive symptoms among adults in Great Britain have risen sharply in recent years. In surveys taken between July 2019 and March 2020, the prevalence was 10%, but this rose to 19% by June 2020 (see Leach et al. (2022) for more details).

This section studies spending on MHS that is related to the primary sources of support for mental health, namely nursing, accommodation, direct payments, homecare, supported living, other long-term care, and other short-term support. Note that the most local funding for MHS is not ring-fenced, meaning that each local National Health Service (NHS) determines its own mental health budget from its overall funding allocation. This was the result of the public health reforms that took place in 2013, during which there was a significant transfer of responsibility for commissioning and providing of public health services from the NHS to the local authorities. In short, the government's view was that, *To ... avoid the problems of the past, we need to reform the public health system. Localism will be at the heart of this system, with responsibilities, freedoms, and funding devolved wherever possible* (Department of Health (2010)). This means that neither the government nor NHS England determines exactly how much funding goes to MHS in local areas.

Hence, an important question is what factors determine the MHS expenditure of the

clinical commissioning group within a local authority, how, and to what extent. The number of studies have investigated the determinants of the local authorities' MHS expenditure, particularly Aziz et al. (2003), Moscone et al. (2007), and Moscone and Tosetti (2010). The empirical models in these studies are based on a traditional reduced-form demand/supply analysis in which the MHS expenditure is explained by a set of risk factors of the need for MHS within the panel data regression framework. In consequence, these models only explain variations in the local governments' spending in accordance with changes in the risk factors of mental health, particularly the marginal effects of those risk factors.

On the contrary, the VCPD specification enables us to analyze the municipal disparities in the local governments' MHS expenditure specifically by allowing the marginal effects of each of the risk factors to vary and be driven by some idiosyncratic characteristics of the respective local authorities. In this section, we allow such variations to be driven by the political ideology (left-wing or right-wing) towards which voters in respective local authorities are leaning. An intuitive example behind this idea is that some councils decide to give more weight to the elderly whereas others give more to youth in terms of resources according to their political beliefs, particularly about which political party or ideology is in power. Studying these disparities in MHS expenditure is important because they can lead to unequal access to MHS given the current environment in which the austerity introduced by the UK government in 2010 squeezes local council budgets, leading to financial difficulties for the local councils.

Moreover, the spatial dependence brought about by measurement errors that spill across boundaries and/or spatially correlated unobservable latent variables can be captured in our model via the SED specification. In the context of MHS expenditure, one example of factors that may lead to these features is the closure of a large psychiatric hospital, which serves patients from various municipalities. High number of psychiatric hospital admission in two or more neighboring authorities may be caused by aviation, which impacts residential communities close to airports by affecting community annoyance, sleep deprivation, and other symptoms of a mental health problem. We shall elaborate on these points and provide an empirical illustration in the next sections.



## 4.2 Empirical model and data

The study in this section focuses on 151 councils in England out of 333 local authorities, who have responsibilities for social services. However, two local councils, namely City of London and Isles of Scilly, are excluded from our analysis because of their distinct socio economic and demographic characteristics. In the time dimension, we focus on the period between 2016/17 and 2019/20, which reflects our interest on the impact of the government’s public health reform of 2013 and the reduction in its spending on public health grants during the period. Therefore,  $N = 149$  and  $T = 4$ .

Our dependent variable is the MHS expenditure of a local authority standardized by the total population in each local authority. Figure 1 presents a per capita measure of the standardized MHS expenditure for persons aged between 18 and 64( *mhs* hereafter) for all the local authorities during 2016/17 and 2019/20. It is evident that *mhs* tends to be distributed in clusters, with the highest concentrations in metropolitan areas such as Greater London and Manchester. We assume that the data generating process behind *mhs* is as follows

$$mhs_{it} = X_{it}\beta_0(Z_{it}) + u_{it}, \quad (4.1)$$

where  $\beta_0(z)$  is a vector of smooth functions and  $u_{it}$  is the disturbance supporting the types of spatial interactions, which are defined in details in Section 2.1. We now discuss the regressors and covariate used in Model (4.1).

Let us begin with  $X_{it}^\top = (X_{1,it}, \dots, X_{D,it})^\top$ . The first proposition is to set  $X_{1,it} = 1$ , which implies that

$$mhs_{it} = \beta_{0,1}(Z_{it}) + X_{it}^*\beta_0^*(Z_{it}) + u_{it}, \quad (4.2)$$

where  $X_{it}^{*\top} = (X_{2,it}, \dots, X_{D,it})^\top$  and  $\beta_0^*(Z_{it}) = \{\beta_{0,2}(Z_{it}), \dots, \beta_{0,D}(Z_{it})\}^\top$ . In (4.2),  $\beta_{0,1}(Z)$  captures the direct effect of  $Z$ .

The remaining regressors are from two sources. Firstly, we select the explanatory variables suggested in previous studies as area-level characteristics potentially linked to the need for MHS (see e.g., McCrone and Jacobson (2004), Aziz et al. (2003), Moscone et al. (2007)). Our study explains the municipal disparity in *mhs* on the basis of a set of risk factors, namely (i) population density (*psq*), (ii) percentage of male population (*pmp*), (iii) percentage of population under 14 year of age (*pu14*), (iv) standardized mortality ratio

(*smr*), (v) number of jobs in a local area (*noj*), (vi) percentage of households headed by lone parent (*plp*), and (vii) number of unemployment benefit claimants (*nuc*). In addition, we include (ix) median house price (*mhp*), and (x) median weekly wage (*mww*) to control for the supply-side factors. Table 6 provides full descriptions and the sources of data, whereas Table 7 presents the descriptive statistics of all the variables and covariate used in the model.

The selection of these variables is in accordance with the evidence found in the literature, which suggests their association with mental health needs. In particular, *psq*, *nuc* and *noj* are indicators of general living condition in a particular local authority, whereas *smr* is an indicator of general health conditions. In addition, *pmp* is included because of the evidence that the gender is highly correlated with use of mental health care, whereas *pu14* and *plp* are included as it was shown that poorer mental health is associated with specific socio demographic characteristics, particularly a high proportion of old people and non married adults in the population. Although it is interesting to study the effects of other factors such as discrimination/deprivation measures and the percentage of immigrants at local level, this is found to cause the problem of multicollinearity that occurs when including variables such as deprivation indices, ethnicity, and the number of homeless people and refugees (see e.g., Moscone and Knapp (2005)). Hence, they are omitted in favour of a parsimonious model. Furthermore, the effects of immigration on MHS expenditures are not clear unless much more detailed information are available on the prevalence of health conditions in migrants such as pre-migration experiences, experience during and after migration to the UK, education, socio economic status and ethnicity (see Jass and Massey (2004), Antecol and Bedard (2006), and Giuntella et al. (2018) for more details).

We now discuss the covariate within the coefficient function. In the empirical work, we assume that British politics are essentially one-dimensional, so that political parties can be organized along a classical left-right axis. Accordingly, political endorsement is measured by the percentage of voters who voted for the parties that are traditionally considered to be right-wing. Such a percentage is denoted hereafter by *vote*. In the UK, this can be quite safely defined as the percentage of voters who have voted for the Conservative and UK Independence Parties in the local government elections (see Dunleavy et al. (2018) for more details). Table 6 provides the descriptions and data sources.

The final item required to complete the empirical model is the weight matrix. We argue that geographical-based weight matrices are the most relevant to our analysis, since they help us to capture how and to what extent the MHS expenditure of a local authority depends on that of its neighbors. This section considers a similar set of weight matrices to that used in Section 3 so that we can analyze if and how the results of the estimation change with weight matrices that differ in their degree of sparseness. In particular, we consider weight matrices based on (i) the  $k$ -nearest neighbors criterion, where  $k$  is 4, 10, or 16; and (ii) sphere of influence, referred to hereafter as KW4, KW10, KW16, and SW, respectively. Section 3 of the supplemental material discusses the spatial weight matrices in more detail.

### 4.3 Estimation results

The steps taken in our empirical analysis are congruent with those of the methodologies developed in Section 2. That is, we firstly estimate the empirical model in (4.2) and perform variable selection by using the proposed estimation procedure and the SAREC-KLASSO technique, respectively. Once relevant regressors are identified, we drop those considered to be irrelevant, then re-estimate the model and employ the testing procedure discussed in Section 2.4 to check whether the associated functional coefficients are constant. Below, we discuss some important findings. The implications of these on the determinants of and disparity in MHS expenditure in England are discussed in the next section.

Table 8 presents the estimates of the autoregressive parameter, which increase as higher numbers of nearest neighbors are taken into account. In addition, the outcome of the variable selection suggests that there are five regressors which are truly relevant, whereas the rest are not. This selection is robust across the spatial weight matrices. Similarly, the outcomes of the coefficient constancy test remain largely unchanged across different weight matrices. On the contrary, without incorporating the SAREC structure (a case that is denoted in the table by  $KW0$ ), the selection suggests that all of the variables except one are relevant. This finding is in significant contrast to that based on the SAREC-KLASSO method. The test statistics of the coefficient constancy test are also much larger compared with those associated with  $KW4$ ,  $KW10$ ,  $KW16$ , and  $SW$ . By applying the SAREC-

KLASSO method, we find that the intercept and four of the regressors are relevant in explaining the disparities in mental health spending by local councils in England, namely the percentage of male population ( $pmp$ ), standardized mortality rates ( $smr$ ), median weekly wage ( $mw$ ), and population density ( $psq$ ). Intriguingly, only the effect of  $pmp$  is found to depend nonlinearly on  $vote$ .

These findings are well supported by the estimates of the coefficient functions. Since the empirical results are robust to the different spatial weight matrices, here we focus only on those in Figure 2, which are based on  $KW4$  (see Section 4 of the supplemental material for the results based on other spatial weight matrices). In the figure, the red lines are confidence bands. The results of Corollary 2.6 and (2.24) suggest that we construct these bands as follows

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

where  $\hat{\beta}_d(z)$  is the unpenalized estimate obtained after excluding the irrelevant regressors,

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

Here,  $\alpha = 0.1$  (i.e., 90% confidence level), and  $d_N$  and  $\widehat{SD} \left\{ \hat{\beta}_d(z) \right\}$  are defined in (2.24) and (2.25), respectively. Moreover, the broken blue line is computed as follows

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

These confidence bands suggest that the estimates of the functional coefficients for  $smr$ ,  $mw$ , and  $psq$  are statistically significant, and their effects are independent from  $vote$ . These results are consistent with those in Table 8. The estimate of the functional coefficient for  $pmp$  is also statistically significant, but (unlike the previous cases) its effect depends nonlinearly on  $vote$ . These results are less conclusive. Finally, the estimates of the coefficient functions based on other spatial weight matrices are similar to those presented in Figures 3 and therefore are presented in the supplementary material.

## 4.4 Policy implications

This section discusses implications of our empirical findings on the determinants of and disparity in the MHS expenditure in England.

Independently from *vote*, the percentage of male population does not have statistically significant effect on the MHS expenditure across local councils. However, by conditioning its effect on political preference, it seems that *mpm* has a positive (negative) impact on *mhs* in councils that are dominated by central-left (central-right) politics. This suggests that the gender disparity exists in accessing MHS such that left-leaning and right-leaning councils react differently when facing a similar change in *mpm*. Since many previous studies suggested that females are more likely to be affected by mental health problems (see e.g., McManus et al. (2016)), a decrease in *mpm* should lead to an increase in *mhs* to maintain a similar level of accessibility. Our results suggest that only right-wing councils have made decisions accordingly.

Moreover, independently from *vote*, the standardized mortality ratio does not seem to have a statistically significant effect on the MHS expenditure across councils. However, by conditioning its effect on British local political preferences, it seems that *smr* contributes to positive effects on *mhs* by councils that are dominated by central-left politics. Hence, for these councils the allocation of mental health grants is linked more closely to the general indicators of deprivation and public health conditions. This appropriately reflects the social welfare, public provision, and collectivism are important characteristics of the left-wing political ideologies.

In addition, our results suggest that, irrespective of the political atmosphere within local authorities, the impact of population density on MHS expenditure is positive and significant. In other words, authorities generally anticipate higher demand for MHS in inner-city areas that are more densely populated. This anticipation is in agreement with empirical evidence found in many other studies (see e.g., Maconick et al. (2021)). When we consider the coefficient function more closely, it seems that political endorsement leads to some disparity in accessing MHS. For example, if there is an increase in the population density, right-leaning authorities tend to respond more generously. However, it should be noted that this nonlinearity is not statistically significant.

Moreover, median weekly earnings are found to have positive effects on MHS expenditure which is congruent with the findings of Moscone et al. (2007). It is highly likely that higher income often leads to a better living standard, which in turn helps to lessen other problems within a local authority that require funding from the council. This makes some

extra funding available, which can be spent by the local authority on MHS. We did not reject the null hypothesis of the coefficient constancy and therefore conclude that there is no disparity in spending on MHS because councils at both end of the political spectrum behave similarly.

We now focus on the selection result which suggests that percentage of lone parent,  $plp$ , is irrelevant when allocating MHS expenditure. This finding is quite alarming, particularly given the evidence that a person who takes care of their child or children without a husband, wife, or partner is more likely to require mental health support (see Moscone et al. (2007) for details). In England, the number of lone parent families has increased significantly in recent years (reaching 2.9 million families in 2020, accounting for 14.7% of families in the UK (Office of National Statistics (2020))).

## 5 Conclusions

We introduced a regression model that combines two important areas of methodological development in the analysis of panel data, namely the VC and SED specifications. Whereas the former allows a flexible nonlinear interaction, our SED enables the spatial interactions involving both the error and unit-specific error components. This allows the generalization that differs significantly from other random effects models in the literature, and broaden the applicability of our models to problems in health, urban and environmental economics.

Furthermore, we establish a novel estimation procedure. This can be viewed as a generalization of the QML method for spatial panel data models to the well-known conditional local kernel-weighted likelihood. Our main theoretical results are the uniform consistency of our proposed estimators, which are established based on a set of standard and primitive conditions. These results form the basis for establishing our inference methods, particularly the SAREC-KLASSO procedure for variable selection and hypothesis testing procedure for the parameter constancy. Our theory suggests that these methods can be implemented effectively even under the SED specification.

In addition, the empirical analysis has made some important contributions. Firstly, we illustrate the practicality and applicability of our model and statistical methods. To ensure their empirical validity and relevance, we conduct an extensive simulation exercise in order

731 to examine their finite-sample performance and robustness. Our proposed procedures are  
732 found to outperform existing methods, particularly when a true data generating process in-  
733 volves spatially correlated errors. Secondly, the empirical findings help to explain not only  
734 the determinants of the MHS expenditure of local councils, but also whether the political  
735 tendency of the local electorates bring about the municipal disparity and hence unequal  
736 access to MHS. Intriguingly, our results suggest that the gender disparity exists in access-  
737 ing MHS. This is in the sense that left-leaning and right-leaning councils react differently  
738 when facing a similar change in the percentage of male population. Furthermore, general  
739 indicators of deprivation and public health conditions seem to be given greater weight by  
740 left-leaning councils when allocating MHS funding. These could also bring about the mu-  
741 nicipal disparities and unequal access to MHS. Finally, the most concerning result suggests  
742 that (irrespective of the political ideologies), local councils are inattentive to percentage  
743 of lone parent when allocating funding for MHS; despite strong evidence that lone parents  
744 are more likely to require mental health support, as shown in previous studies.

# A Appendix A

This section consists of two subsections. Section A.1 presents a set of technical lemmas that are required for the mathematical proofs of the main theoretical results of the paper which are presented in Section A.2.

## A.1 Lemmas

We first derive the lower bound of the expected concentrated local likelihood, more specifically  $E(\tilde{\ell}_z^c(\delta))$ , which is denoted as  $\bar{\ell}_z^c(\delta)$  in (2.10).

**Lemma A.1.** *Under Assumptions of Theorem 2.1,*

$$\bar{\ell}_z^c(\delta) = -\frac{1}{2} \left[ \log\{2\pi\} + \log \left\{ \sigma_v^2(z) \frac{\text{tr}[\mathbb{Q}_{0N} \bar{\mathbb{Q}}_N^\top \bar{\mathbb{Q}}_N]}{NT} \right\} + \frac{\log |\mathbb{Q}_N|}{NT} + 1 \right] f(z) \leq E(\tilde{\ell}_z^c(\delta)) \quad (\text{A.1})$$

as  $N \rightarrow \infty$ .

*Derivation of Lemma A.1:* By Jensen's inequity, we have

$$\begin{aligned} E(\tilde{\ell}_z^c(\delta)) &= -\frac{1}{2} \left( \log\{2\pi\} + E(\log\{\tilde{\sigma}_v^2(z)\}) + \frac{\log\{|\mathbb{Q}_N|\}}{NT} + 1 \right) f(z) + o(1) \\ &\geq -\frac{1}{2} \left( \log\{2\pi\} + \log\{E(\tilde{\sigma}_v^2(z))\} + \frac{\log\{|\mathbb{Q}_N|\}}{NT} + 1 \right) f(z), \end{aligned}$$

where  $E(\tilde{\sigma}_v^2(z)) = \sigma_v^2(z) \frac{\text{tr}[\mathbb{Q}_{0N} \bar{\mathbb{Q}}_N^\top \bar{\mathbb{Q}}_N]}{NT}$ . By using the second order approximation of the expected ratio of two random processes,

$$E(\tilde{\sigma}_v^2(z)) \approx \frac{E(\tilde{\sigma}_{NU,v}^2(z))}{E(\tilde{\sigma}_{DE,v}^2(z))} + o(h^e),$$

where  $\tilde{\sigma}_{NU,v}^2(z)$  and  $\tilde{\sigma}_{DE,v}^2(z)$  denote the numerator and denominator of  $\tilde{\sigma}_v^2(z)$ , respectively. It is straightforward to show that  $E(\tilde{\sigma}_{DE,v}^2(z)) = f(z) + O(h^e)$ . Now let us consider  $E(\tilde{\sigma}_{NU,v}^2(z))$  as follows. By using the Taylor expansion argument,

$$\begin{aligned} E(\tilde{\sigma}_{NU,v}^2(z)) &= E \left( \frac{1}{NT} \sum_{j=1}^N \sum_{s=1}^T \{ \ddot{y}_{js} - \ddot{X}_{js} \tilde{\beta}(z) \}^2 K_h(Z_{js} - z) \right) \\ &= E \left( \frac{1}{NT} \sum_{j=1}^N \sum_{s=1}^T \left[ B_N(z, hv)^\top \ddot{X}_{js}^\top \ddot{X}_{js} B_N(z, hv) + R_N(z)^\top \ddot{X}_{js}^\top \ddot{X}_{js} R_N(z) + \ddot{u}_{js}^2 \right. \right. \\ &\quad \left. \left. - 2B_N(z, hv)^\top \ddot{X}_{js}^\top \ddot{X}_{js} R_N(z) + 2B_N(z, hv)^\top \ddot{X}_{js}^\top \ddot{u}_{js} - 2R_N(z)^\top \ddot{X}_{js}^\top \ddot{u}_{js} \right] K_h(Z_{js} - z) \right) \\ &\doteq E(F_{1,1}) + E(F_{1,2}) + E(F_{1,3}) + E(F_{1,4}) + E(F_{1,5}) + E(F_{1,6}), \end{aligned} \quad (\text{A.2})$$



760 where  $B_N(z, hv) = \beta(z)^{(1)}(hv) + \dots + \beta(z)^{(\varrho)}(hv)^{\varrho} + o(h^{\varrho})$ ,  $R_N(z) = (\ddot{X}_N^{\top} K_N \ddot{X}_N)^{-1} (\ddot{X}_N^{\top} K_N \ddot{u}_N)$ ,  
 761 under Assumption B5. Firstly, by using the iterated expectation argument, the three terms above  
 762 are as follows

$$E(F_{1,1}) = 2h^{\varrho} \kappa_{\varrho} f(z) \beta^{(\varrho-1)}(z)^{\top} \Omega(z) \frac{\text{tr}[\mathbb{Q}_{0N} \bar{\mathbb{Q}}_N^{\top} \bar{\mathbb{Q}}_N]}{NT} \beta^{(1)}(z) + o(h^{\varrho}) = o(1) \quad (\text{A.3})$$

763 under Assumptions of B1, B3, B4 and because  $\mathbb{Q}_{0N}$  and  $\bar{\mathbb{Q}}_N^{\top} \bar{\mathbb{Q}}_N$  are nonstochastic, where  $\kappa_{\varrho} =$   
 764  $\int u^{\varrho} K(u) du$ ,

$$E(F_{1,3}) = \sigma_v^2(z) \frac{\text{tr}[\mathbb{Q}_{0N} \bar{\mathbb{Q}}_N^{\top} \bar{\mathbb{Q}}_N]}{NT} f(z) + O(h^{\varrho}),$$

765 and  $E(F_{1,5}) = 0$  a.s. because  $E(u_{it}|X_{it}, Z_{it}) = 0$  a.s.. For the rest of the terms in (A.2), the second  
 766 order approximation of the expected ratio of two random processes and the iterated expectation  
 767 argument are used again as follows

$$E(F_{1,2}) = \frac{1}{NT h} \kappa \sigma_v^2(z) \Omega(z)^{-1} f(z)^{-1}$$

768 under Assumptions of B1, B3, and B4 because  $F_{1,2} = (\ddot{X}_N^{\top} K_N \ddot{u}_N)^{\top} (\ddot{X}_N^{\top} K_N \ddot{X}_N)^{-1} (\ddot{X}_N^{\top} K_N \ddot{u}_N)$ ,  
 769 and  $E(F_{1,4}) = E(F_{1,6}) = 0$  a.s. due to the similar argument to the case of  $E(F_{1,5})$ . Hence,  
 770  $E(\tilde{\sigma}_{NU,v}^2(z)) = \sigma_v^2(z) \frac{\text{tr}[\mathbb{Q}_{0N} \bar{\mathbb{Q}}_N^{\top} \bar{\mathbb{Q}}_N]}{NT} f(z) + o(1)$  and, subsequently, (A.1) follows.  $\square$

771 We now present the uniform consistency of the kernel estimations which will be subsequently  
 772 used in the proofs of the main theoretical results.

773 **Lemma A.2.** Suppose  $(\xi_{it}, Z_{it})$  are i.i.d. random vectors for all  $it = 1, \dots, NT$ , where  $\xi_{it}$ s are  
 774 scalar random variables. Under Assumptions of  $E|\xi|^m < \infty$ ,  $\sup_{z \in [0,1]} \int |\xi|^m f(\xi, z) d\xi < \infty$ , where  
 775  $f(\cdot, \cdot)$  denotes the joint density of  $(\xi, Z)$  and  $m > 2$ , and the kernel function  $K(u)$  is a symmetric  
 776 bounded positive function with a bounded support and Lipschitz continuous on its support, then

$$\sup_{z \in [0,1]} \left| \frac{1}{NT} \sum_{j=1}^N \sum_{s=1}^T [K_h(Z_{js} - z) \xi_{js} - E\{K_h(Z_{js} - z) \xi_{js}\}] \right| = O_p \left( \left( \frac{\log(1/h)}{(NT)h} \right)^{1/2} \right)$$

777 provided that  $N^{2\eta-1}h \rightarrow \infty$ , where  $\eta < 1 - m^{-1}$ .

778 *Proof:* The proof can be found in Mack and Silverman (1982).  $\square$

779 **Lemma A.3.** Under Assumptions of Lemma A.2, and the higher-order kernel function, particu-  
 780 larly Assumption B1, and if  $g(z) = E(\xi_{it}|z)$  has the higher-order bounded derivative, then

$$\sup_{z \in [0,1]} \left| \frac{1}{(NT)^2} \sum_{i=1}^N \sum_{t=1}^T \sum_{j=1}^N \sum_{s=1}^T [K_h(Z_{js} - Z_{it}) \xi_{js} - f(Z_{it}) g(Z_{it})] \right| = O_p \left( h^{\varrho} + \left( \left( \frac{\log(1/h)}{(NT)^2 h} \right)^{1/2} \right) \right),$$

781 where  $\varrho > 2$  and  $\frac{1}{2} + \frac{1}{4\varrho} < \eta < 1 - m^{-1}$ .

782 *Proof:* Under some additional regularity conditions, it should be clear that the Lemma A.3 is a  
 783 corollary of Lemma A.2.  $\square$

## 784 A.2 Proofs of the main theoretical results

785 We firstly present the proofs of Theorems 2.1 and 2.2, then subsequently those of Corollaries.

### 786 A.2.1 Proofs of Theorems 2.1 and 2.2

787 Note that the first derivatives of the matrices, namely  $\mathbb{Q}_N^{-1} = \bar{\mathbb{Q}}_N^\top \bar{\mathbb{Q}}_N$  and  $\mathbb{Q}_N$ , are frequently used  
 788 in the following proofs, hence they are presented for the sake of convenience below

$$\begin{aligned} \frac{\partial \mathbb{Q}_N^{-1}}{\partial \rho} &= -2[I_T \otimes (I_N - W_N^\top)]\{Q_{0,N} + (1 + T\phi)^{-1}Q_{1,N}\}[I_T \otimes (I_N - \rho W_N)] \\ \frac{\partial \mathbb{Q}_N^{-1}}{\partial \phi} &= -2[I_T \otimes (I_N - \rho W_N)^\top] \left\{ \frac{T/2}{(1 + T\phi)^2} \right\} [I_T \otimes (I_N - \rho W_N)] \\ \frac{\partial \mathbb{Q}_N}{\partial \rho} &= 2[I_T \otimes (I_N - \rho W_N^\top)^{-1}][I_T \otimes (I_N - W_N)][I_T \otimes (I_N - \rho W_N)^{-1}] \\ &\quad \times \{Q_{0,N} + (1 + \phi T)Q_{1,N}\}[I_T \otimes (I_N - \rho W_N^\top)^{-1}] \\ \frac{\partial \mathbb{Q}_N}{\partial \phi} &= [I_T \otimes (I_N - \rho W_N)^{-1}]\{TQ_{1,N}\}[I_T \otimes (I_N - \rho W_N)^{-1}]. \end{aligned}$$

789 **Proof of Theorem 2.1:** Theorem 2.1 can be proved in two main steps. The uniform consistency  
 790 of  $\hat{\delta}$  uniformly over  $\delta \in \Delta$  is established, then the unique identification of  $\delta_0$  is shown.

791 The uniform consistency over the compact parameter space can be shown by the following  
 792 two steps. It is established by firstly showing that

$$\sup_{z \in [0,1]} \left| \frac{1}{NT} \tilde{\mathcal{L}}^c(\delta) - \frac{1}{NT} \bar{\mathcal{L}}^c(\delta) \right| = O_P \left( (NT)^{-1/2} \right),$$

793 where  $\bar{\mathcal{L}}^c(\delta) = \sum_{i=1}^N \sum_{t=1}^T \bar{\ell}_{Z_{it}}^c(\delta)$ , then the uniform Lipschitz continuity of  $\left| \frac{1}{NT} \tilde{\mathcal{L}}^c(\delta) - \frac{1}{NT} \bar{\mathcal{L}}^c(\delta) \right|$   
 794 over  $\delta \in \Delta$  by establishing the stochastic equi-continuity. Let us present the first step as follows

$$\sup_{z \in [0,1]} \left| \frac{1}{NT} \tilde{\mathcal{L}}^c(\delta) - \frac{1}{NT} \bar{\mathcal{L}}^c(\delta) \right| \doteq \sup_{z \in [0,1]} \left| -\frac{1}{2} (L_{1,1} + L_{1,2} + L_{1,3} + L_{1,4}) \right|,$$

795 where

$$\begin{aligned} L_{1,1} &= \frac{1}{(NT)^2} \log\{2\pi\} \sum_{i=1}^N \sum_{t=1}^T \sum_{j=1}^N \sum_{s=1}^T K_h(Z_{js} - Z_{it}) - \log\{2\pi\} f(Z_{it}) \\ L_{1,2} &= \frac{1}{(NT)^2} \sum_{i=1}^N \sum_{t=1}^T \log\{\tilde{\sigma}_v^2(Z_{it})\} \sum_{j=1}^N \sum_{s=1}^T K_h(Z_{js} - Z_{it}) - \log \left\{ \sigma_v^2(Z_{it}) \frac{\text{tr}[\mathbb{Q}_{0N} \bar{\mathbb{Q}}_N^\top \bar{\mathbb{Q}}_N]}{NT} \right\} f(Z_{it}) \end{aligned}$$

$$\begin{aligned}
L_{1,3} &= \frac{1}{(NT)^2} \frac{\log |\mathbb{Q}_N|}{NT} \sum_{i=1}^N \sum_{t=1}^T \sum_{j=1}^N \sum_{s=1}^T K_h(Z_{js} - Z_{it}) - \frac{\log |\mathbb{Q}_N|}{NT} f(Z_{it}) \\
L_{1,4} &= \frac{1}{(NT)^2} \sum_{i=1}^T \sum_{t=1}^T \sum_{j=1}^N \sum_{s=1}^T K_h(Z_{js} - Z_{it}) - f(Z_{it}).
\end{aligned}$$

797 The uniform consistency of the three terms are similarly shown as follows  $\sup_{z \in [0,1]} |L_{1,k}| = O_p((NT)^{-1/2})$ ,  
 798 where  $k = 1, 3, 4$  by using the similar arguments in Lemma A.3. The last term,  $L_{1,2}$ , is as follows

$$\sup_{z \in [0,1]} \left| \tilde{\sigma}_v^2(Z_{it}) - \sigma_v^2(Z_{it}) \frac{\text{tr}[\mathbb{Q}_{0N} \bar{\mathbb{Q}}_N^\top \bar{\mathbb{Q}}_N]}{NT} \right| = O_p \left( h^\varrho + \left( \frac{\log(1/h)}{(NT)h} \right)^{1/2} \right)$$

799 and

$$\sup_{z \in [0,1]} \left| \frac{1}{NT} \sum_{j=1}^N \sum_{s=1}^T K_h(Z_{js} - Z_{it}) - f(Z_{it}) \right| = O_p \left( h^\varrho + \left( \frac{\log(1/h)}{(NT)h} \right)^{1/2} \right)$$

800 by using the same arguments as in Lemma A.2, under Assumptions of B1, B3 and B4, and

$$\sup_{z \in [0,1]} \left\| \tilde{\beta}(Z_{it}) - \beta(Z_{it}) \right\| = O_p \left( h^\varrho + \left( \frac{\log(1/h)}{(NT)h} \right)^{1/2} \right) \quad (\text{A.4})$$

801 under Assumptions of B1 to B5, where

$$\tilde{\beta}(Z_{it}) = \left[ \sum_{\iota=1}^N \sum_{\tau=1}^T \ddot{X}_{\iota\tau}^\top \ddot{X}_{\iota\tau} K_h(Z_{\iota\tau} - Z_{it}) \right]^{-1} \sum_{\iota=1}^N \sum_{\tau=1}^T \ddot{X}_{\iota\tau}^\top \ddot{y}_{\iota\tau} K_h(Z_{\iota\tau} - Z_{it}) = \beta(Z_{it}) + R_N(Z_{it})$$

802 with

$$\sup_{z \in [0,1]} \left\| \sum_{\iota=1}^N \sum_{\tau=1}^T \ddot{X}_{\iota\tau}^\top \ddot{X}_{\iota\tau} K_h(Z_{\iota\tau} - Z_{it}) - \Omega(Z_{it}) \frac{\text{tr}[\mathbb{Q}_{0N} \bar{\mathbb{Q}}_N^\top \bar{\mathbb{Q}}_N]}{NT} f(Z_{it}) \right\| = O_p \left( h^\varrho + \left( \frac{\log(1/h)}{(NT)h} \right)^{1/2} \right), \quad (\text{A.5})$$

803 and by using the facts that  $(A + hB)^{-1} = A^{-1} - hA^{-1}BA^{-1} + O(h^2)$  (see Fan and Zhang (2000)  
 804 for details),

$$\sup_{z \in [0,1]} \|R_N(Z_{it})\| = O_p \left( h^\varrho \left( \frac{\log(1/h)}{(NT)h} \right)^{1/2} + \left( \frac{\log(1/h)}{(NT)h} \right) \right) \quad (\text{A.6})$$

805 because

$$\sup_{z \in [0,1]} \left\| \ddot{X}_N^\top K_{N,it} \ddot{y}_N \right\| = O_p \left( \left( \frac{\log(1/h)}{(NT)h} \right)^{1/2} \right), \quad (\text{A.7})$$

806 where  $K_{N,it} = \text{diag}\{K_h(Z_{11} - Z_{it}), \dots, K_h(Z_{NT} - Z_{it})\}$ . Then, by using the similar arguments as in  
 807 Lemma A.3,  $\sup_{z \in [0,1]} |L_{1,2}| = o_P((NT)^{-1/2})$ . Hence,  $\sup_{z \in [0,1]} \left| \frac{1}{NT} \tilde{\mathcal{L}}^c(\delta) - \frac{1}{NT} \bar{\mathcal{L}}^c(\delta) \right| = O_p((NT)^{-1/2})$ .

808 The next step is establishing the uniform stochastic equip-continuity over  $\delta \in \Delta$  as follows

$$\begin{aligned}
& \sup_{\|\delta - \delta^*\| < \epsilon} \sup_{z \in [0,1]} \left\| \frac{1}{NT} \tilde{\mathcal{L}}^c(\delta) - \frac{1}{NT} \bar{\mathcal{L}}^c(\delta) - \left\{ \frac{1}{NT} \tilde{\mathcal{L}}^c(\delta^*) - \frac{1}{NT} \bar{\mathcal{L}}^c(\delta^*) \right\} \right\| \\
& \leq \sup_{z \in [0,1]} \left\| \frac{1}{NT} \{\tilde{\mathcal{L}}^c(\bar{\delta})\}^{(1)} - \frac{1}{NT} \{\bar{\mathcal{L}}^c(\bar{\delta})\}^{(1)} \right\| \cdot \sup_{\|\delta - \delta^*\| < \epsilon} \|\delta - \delta^*\| = o_P(1),
\end{aligned}$$

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

$$\sup_{z \in [0,1]} \left\| \frac{1}{NT} \{\tilde{\mathcal{L}}^c(\bar{\delta})\}^{(1)} - \frac{1}{NT} \{\bar{\mathcal{L}}^c(\bar{\delta})\}^{(1)} \right\| = O_P(1).$$

Now, let us consider

$$\sup_{z \in [0,1]} \left\| \frac{1}{NT} \{\tilde{\mathcal{L}}^c(\bar{\delta})\}^{(1)} - \frac{1}{NT} \{\bar{\mathcal{L}}^c(\bar{\delta})\}^{(1)} \right\| \doteq \sup_{z \in [0,1]} \left\| -\frac{1}{2} (L_{2,1} + L_{2,2}) \right\|,$$

where

$$\begin{aligned} L_{2,1} &= \frac{1}{(NT)^2} \sum_{i=1}^N \sum_{t=1}^T \sum_{j=1}^N \sum_{s=1}^T \frac{\{\tilde{\sigma}_v^2(Z_{it})\}^{(1)}}{\tilde{\sigma}_v^2(Z_{it})} K_h(Z_{js} - Z_{it}) - \frac{\text{tr} \left[ \mathbb{Q}_{0N} \frac{\partial \mathbb{Q}_N^{-1}}{\partial \delta} \right]}{\text{tr} [\mathbb{Q}_{0N} \mathbb{Q}_N^{-1}]} f(Z_{it}) \\ L_{2,2} &= \frac{1}{(NT)^2} \sum_{i=1}^N \sum_{t=1}^T \sum_{j=1}^N \sum_{s=1}^T \frac{\text{tr} \left[ \mathbb{Q}_N^{-1} \frac{\partial \mathbb{Q}_N}{\partial \delta} \right]}{NT} K_h(Z_{js} - Z_{it}) - \frac{\text{tr} \left[ \mathbb{Q}_N^{-1} \frac{\partial \mathbb{Q}_N}{\partial \delta} \right]}{NT} f(Z_{it}) \end{aligned}$$

with  $\{\tilde{\sigma}_v^2(Z_{it})\}^{(1)}$  denoting the gradients of  $\tilde{\sigma}_v^2(Z_{it})$ . Let us consider  $\{\tilde{\sigma}_v^2(Z_{it})\}^{(1)}$  below

$$\begin{aligned} \{\tilde{\sigma}_v^2(Z_{it})\}^{(1)} &= \left[ \frac{1}{NT} \sum_{\iota=1}^N \sum_{\tau=1}^T K_h(Z_{\iota\tau} - Z_{it}) \right]^{-1} \\ &\times 2 \left[ \frac{1}{NT} \sum_{\iota=1}^N \sum_{\tau=1}^T \{\dot{y}_{\iota\tau} - \ddot{X}_{\iota\tau} \tilde{\beta}(Z_{it})\} \left\{ \dot{y}_{\iota\tau} - \dot{X}_{\iota\tau} \tilde{\beta}(Z_{it}) - \ddot{X}_{\iota\tau} \frac{\partial \tilde{\beta}(Z_{it})}{\partial \delta} \right\} K_h(Z_{\iota\tau} - Z_{it}) \right], \end{aligned}$$

where  $\dot{y}_{\iota\tau}$  and  $\dot{X}_{\iota\tau}$  are defined in the same manner as those of  $\ddot{y}_{\iota\tau}$  and  $\ddot{X}_{\iota\tau}$ , respectively, with  $\frac{\partial \bar{\mathbb{Q}}_N}{\partial \delta}$  instead of  $\bar{\mathbb{Q}}_N$ , and

$$\begin{aligned} \frac{\partial \tilde{\beta}(Z_{it})}{\partial \delta} &= (\ddot{X}_N^\top K_{N,it} \ddot{X}_N)^{-1} \{ (\dot{X}_N^\top K_{N,it} \ddot{u}_N) + (\ddot{X}_N^\top K_{N,it} \dot{u}_N) \} \\ &\quad - 2(\ddot{X}_N^\top K_{N,it} \ddot{X}_N)^{-1} \{ (\ddot{X}_N^\top K_{N,it} \ddot{u}_N)^\top (\dot{X}_N^\top K_{N,it} \ddot{X}_N) \} (\ddot{X}_N^\top K_{N,it} \ddot{X}_N)^{-1} \\ &\doteq \dot{R}_{N,X}(Z_{it}) + \dot{R}_{N,U}(Z_{it}) - 2\dot{\mathcal{R}}_N(Z_{it}) \end{aligned}$$

by using the argument of  $\tilde{\beta}(Z_{it}) = \beta(Z_{it}) + R_N(Z_{it})$  with  $\dot{u}_{\iota\tau}$  being similarly defined as  $\dot{y}_{\iota\tau}$  and  $\dot{X}_{\iota\tau}$  above. Then, by using the similar arguments to (A.6),

$$\sup_{z \in [0,1]} \left\| \dot{R}_{N,X}(Z_{it}) + \dot{R}_{N,U}(Z_{it}) \right\| = o_P \left( h^e + \left( \frac{\log(1/h)}{(NT)h} \right)^{1/2} \right), \quad (\text{A.8})$$

and

$$\sup_{z \in [0,1]} \left\| \dot{\mathcal{R}}_N(Z_{it}) \right\| = o_P \left( h^e + \left( \frac{\log(1/h)}{(NT)h} \right)^{1/2} \right) \quad (\text{A.9})$$

821 because

$$\dot{\mathcal{R}}_N(Z_{it}) \leq \frac{1}{(\hat{\lambda}_{\min})^2} \left( \left\| \ddot{X}_N^\top K_{N,it} \ddot{u}_N \right\|^2 \right)^{1/2} \left( \left\| \dot{X}_N^\top K_{N,it} \ddot{X}_N \right\|^2 \right)^{1/2},$$

822 where  $\hat{\lambda}_{\min} = \min\{\hat{\lambda}_{it}^{\min}\}$  with  $\hat{\lambda}_{it}^{\min}$  being the smallest eigenvalue of  $\ddot{X}_N^\top K_{N,it} \ddot{X}_N$  and  $P(\hat{\lambda}_{\min} \rightarrow$   
823  $\lambda_0^{\min}) \rightarrow 1$  from (A.5) where  $\lambda_0^{\min} = \inf_{z \in [0,1]} \lambda_{\min}(f(z)\Omega(z))$  with  $\lambda_{\min}(\cdot)$  denoting for the minimal  
824 eigenvalues of an arbitrarily positive definite matrix. By using the results of (A.4), (A.8) and  
825 (A.9),  $\frac{\{\tilde{\sigma}_v^2(Z_{it})\}^{(1)}}{\tilde{\sigma}_v^2(Z_{it})}$  is given below

$$\begin{aligned} \frac{\{\tilde{\sigma}_v^2(Z_{it})\}^{(1)}}{\tilde{\sigma}_v^2(Z_{it})} &= \left[ \sum_{\iota=1}^N \sum_{\tau=1}^T \{ \ddot{X}_{\iota\tau} B_N(Z_{it}, hv) + \ddot{u}_{\iota\tau} + o_P(1) \}^2 K_h(Z_{\iota\tau} - Z_{it}) \right]^{-1} \\ &\times \left[ \sum_{\iota=1}^N \sum_{\tau=1}^T \{ \ddot{X}_{\iota\tau} B_N(Z_{it}, hv) + \ddot{u}_{\iota\tau} + o_P(1) \}^\top \{ \dot{X}_{\iota\tau} B_N(Z_{it}, hv) + \dot{u}_{\iota\tau} + o_P(1) \} K_h(Z_{\iota\tau} - Z_{it}) \right] \\ &\doteq L_{2,1,1} + L_{2,1,2} + L_{2,1,3} + L_{2,1,4}. \end{aligned}$$

826 Each terms above are considered as follows. By using the similar arguments to those of (A.3) and  
827 (A.6),

$$\begin{aligned} \sup_{z \in [0,1]} \|L_{2,1,1}\| &= \sup_{z \in [0,1]} \left\| \left[ \frac{1}{NT} \sum_{\iota=1}^N \sum_{\tau=1}^T \{ \ddot{X}_{\iota\tau} B_N(Z_{it}, hv) + \ddot{u}_{\iota\tau} \}^2 K_h(Z_{\iota\tau} - Z_{it}) \right]^{-1} \right. \\ &\quad \times \left. \left[ \frac{1}{NT} \sum_{\iota=1}^N \sum_{\tau=1}^T \{ B_N(Z_{it}, hv)^\top \ddot{X}_{\iota\tau}^\top \dot{X}_{\iota\tau} B_N(Z_{it}, hv) \} K_h(Z_{\iota\tau} - Z_{it}) \right] \right\| \\ &= o_P \left( h^\varrho + \left( \frac{\log(1/h)}{(NT)h} \right)^{1/2} \right), \end{aligned}$$

828 and using the facts that  $E(u_{it}|X_{it}, Z_{it}) = 0$  a.s.,

$$\begin{aligned} \sup_{z \in [0,1]} \|L_{2,1,2} + L_{2,1,3}\| &= \sup_{z \in [0,1]} \left\| \left[ \frac{1}{NT} \sum_{\iota=1}^N \sum_{\tau=1}^T \{ \ddot{X}_{\iota\tau} B_N(Z_{it}, hv) + \ddot{u}_{\iota\tau} \}^2 K_h(Z_{\iota\tau} - Z_{it}) \right]^{-1} \right. \\ &\quad \times \left. \left[ \frac{1}{NT} \sum_{\iota=1}^N \sum_{\tau=1}^T \{ B_N(Z_{it}, hv)^\top \ddot{X}_{\iota\tau}^\top \dot{u}_{\iota\tau} + \ddot{u}_{\iota\tau}^\top \dot{X}_{\iota\tau} B_N(Z_{it}, hv) \} K_h(Z_{\iota\tau} - Z_{it}) \right] \right\| \\ &= o_P \left( h^\varrho + \left( \frac{\log(1/h)}{(NT)h} \right)^{1/2} \right). \end{aligned}$$

829 The last term is as follows

$$\sup_{z \in [0,1]} \left\| L_{2,1,4} - \frac{\text{tr} \left[ \mathbb{Q}_{0N} \frac{\partial \mathbb{Q}_N^{-1}}{\partial \delta} \right]}{\text{tr}[\mathbb{Q}_{0N} \mathbb{Q}_N^{-1}]} \right\| = o_P \left( h^\varrho + \left( \frac{\log(1/h)}{(NT)h} \right)^{1/2} \right), \quad (\text{A.10})$$

830 where

$$L_{2,1,4} = \left[ \sum_{\iota=1}^N \sum_{\tau=1}^T \{ \ddot{X}_{\iota\tau} B_N(Z_{\iota\tau}, hv) + \ddot{u}_{\iota\tau} \}^2 K_h(Z_{\iota\tau} - Z_{\iota\tau}) \right]^{-1} \left[ \sum_{\iota=1}^N \sum_{\tau=1}^T \ddot{u}_{\iota\tau} \dot{u}_{\iota\tau} K_h(Z_{\iota\tau} - Z_{\iota\tau}) \right].$$

831 Hence,  $\sup_{z \in [0,1]} \|L_{2,1} + L_{2,2}\| = o_P(1)$ . Consequently, the uniform stochastic equip-continuity is  
 832 established.

833 Lastly, the identification condition of  $\delta_0$  is presented. The identification of  $\delta_0$  is established  
 834 by showing the counter argument as follows. Consider the Jensen's inequality below

$$\begin{aligned} \bar{\ell}_z^c(\delta) - \bar{\ell}_z^c(\delta_0) &= -\frac{1}{2} \left( \frac{1}{NT} \log \left\{ |\mathbb{Q}_N \mathbb{Q}_{0N}^{-1}| \right\} + \log \left\{ \frac{\text{tr}[\mathbb{Q}_{0N} \mathbb{Q}_N^{-1}]}{NT} \right\} \right) f(z) \\ &\leq 0. \end{aligned}$$

835 The above inequality 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  
 836 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  $\bar{D}_\epsilon(\cdot)$   
 837 represent an open  $\epsilon$ -neighborhood and its complement, respectively, and  $\lim_{N \rightarrow \infty} \mathbb{Q}_N(\delta^*) \rightarrow \lim_{N \rightarrow \infty} \mathbb{Q}_{0N}$ .  
 838 Hence, the unique identification condition requires (2.11), namely  $\lim_{N \rightarrow \infty} \left( \max_{\delta \in \bar{D}_\epsilon(\delta_0) \cap \Delta} \bar{\ell}_z^c(\delta) \right) \neq$   
 839  $\bar{\ell}_z^c(\delta_0)$  for any  $\delta$ .  $\square$

840 **Proof of Theorem 2.2:** Let us first introduce the Taylor expansion of  $\hat{\mathbb{Q}}_N$  as follows

$$\hat{\mathbb{Q}}_N = \bar{\mathbb{Q}}_N^0 + \frac{\partial \bar{\mathbb{Q}}_N^0}{\partial \rho_0} (\hat{\rho} - \rho_0) + \frac{\partial \bar{\mathbb{Q}}_N^0}{\partial \phi_0} (\hat{\phi} - \phi_0) + o_p((NT)^{-1/2}) \quad (\text{A.11})$$

841 uniformly over  $z \in [0, 1]$  using the result of Theorem 2.1. Then, by using (A.11),  $\hat{\beta}(z)$  can be  
 842 rewritten as follows

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

843 where  $\ddot{y}_{0,j,s}$  and  $\ddot{X}_{0,j,s}$  are defined similarly as those of  $\ddot{y}_{js}$  and  $\ddot{X}_{js}$ , respectively, with  $\bar{\mathbb{Q}}_N^0$  instead  
 844 of  $\bar{\mathbb{Q}}_N$ , and  $\dot{D}_{js}(z) = \dot{D}_{jz,\rho}(z) + \dot{D}_{jz,\phi}(z)$  and  $\dot{N}_{js}(z) = \dot{N}_{js,\rho}(z) + \dot{N}_{js,\phi}(z)$  with

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

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

846 where  $\dot{y}_{0,js}$  and  $\dot{X}_{0,js}$  are defined similarly as those of  $\ddot{y}_{js}$  and  $\ddot{X}_{js}$ , respectively, with  $\frac{\partial \bar{\mathbb{Q}}_N^0}{\partial \delta_0}$  instead  
 847 of  $\bar{\mathbb{Q}}_N$ . Notice that  $\sup_{z \in [0,1]} \|\dot{D}_{js}(z)\| = \sup_{z \in [0,1]} \|\dot{N}_{js}(z)\| = O_P\left(h^e + \left(\frac{\log(1/h)}{(NT)h}\right)^{1/2}\right)$  by using the  
 848 same arguments as in Lemma A.2. Hence, it is the case that

$$\sup_{z \in [0,1]} \|\hat{\beta}(z) - \beta_0(z)\| = O_P\left(h^e + \left(\frac{\log(1/h)}{(NT)h}\right)^{1/2}\right)$$

849 by using the similar arguments to those of (A.5) and (A.7),

$$\sup_{z \in [0,1]} \left\| \sum_{j=1}^N \sum_{s=1}^T \ddot{X}_{0,js}^\top \ddot{X}_{0,js} K_h(Z_{js} - z) - \Omega(z)f(z) \right\| = O_P\left(h^e + \left(\frac{\log(1/h)}{(NT)h}\right)^{1/2}\right)$$

850 and

$$\sup_{z \in [0,1]} \left\| \sum_{j=1}^N \sum_{s=1}^T \ddot{u}_{0,js}^\top \ddot{X}_{0,js} K_h(Z_{js} - z) \right\| = O_P\left(\left(\frac{\log(1/h)}{(NT)h}\right)^{1/2}\right).$$

851 Now, let us consider the convergence of  $\hat{\sigma}_v^2(z)$ . By using (A.11),  $\hat{\sigma}_v^2(z)$  can be rewritten as

$$\begin{aligned} \hat{\sigma}_v^2(z) &= \left[ \sum_{j=1}^N \sum_{s=1}^T K_h(Z_{js} - z) \right]^{-1} \left( \sum_{j=1}^N \sum_{s=1}^T \left\{ \left[ \ddot{X}_{0,js} + \dot{X}_{0,js} O_P((NT)^{-1/2}) \right] \beta_0(z) \right. \right. \\ &\quad \left. \left. - \left[ \ddot{X}_{0,js} + \dot{X}_{0,js} O_P((NT)^{-1/2}) \right] \hat{\beta}(z) + \ddot{u}_{0,js} + \dot{u}_{0,js} O_P((NT)^{-1/2}) \right\}^2 K_h(Z_{js} - z) \right) \\ &\doteq S_1 + S_2 + S_3 + S_4 + 2S_5 + 2S_6 + 2S_7 + 2S_8 + 2S_9 + 2S_{10}, \end{aligned}$$

852 where

$$\begin{aligned} &\sup_{z \in [0,1]} |S_1 + S_2 + S_5| \\ &= \sup_{z \in [0,1]} \left\| \left[ \sum_{j=1}^N \sum_{s=1}^T K_h(Z_{js} - z) \right]^{-1} \left[ \sum_{j=1}^N \sum_{s=1}^T \{ \beta_0(z) - \hat{\beta}(z) \}^\top \left( \ddot{X}_{0,js}^\top \ddot{X}_{0,js} \right. \right. \right. \\ &\quad \left. \left. + \dot{X}_{0,js}^\top \dot{X}_{0,js} O_P((NT)^{-1}) + 2\ddot{X}_{0,js}^\top \dot{X}_{0,js} O_P((NT)^{-1/2}) \right) \{ \beta_0(z) - \hat{\beta}(z) \} K_h(Z_{js} - z) \right] \right\| \\ &= o_P\left(h^e + \left(\frac{\log(1/h)}{(NT)h}\right)^{1/2}\right) \end{aligned}$$

853 by the uniform convergence result of  $\hat{\beta}(z)$  above and using the similar arguments in (A.5),

$$\begin{aligned}
& \sup_{z \in [0,1]} |S_6 + S_7 + S_8 + S_9| \\
&= \sup_{z \in [0,1]} \left\| \left[ \sum_{j=1}^N \sum_{s=1}^T K_h(Z_{js} - z) \right]^{-1} \left[ \sum_{j=1}^N \sum_{s=1}^T 2\{\beta_0(z) - \hat{\beta}(z)\}^\top \left( \ddot{X}_{0,js}^\top \ddot{u}_{0,js} + \ddot{X}_{0,js}^\top \dot{u}_{0,js} O_P((NT)^{-1/2}) \right. \right. \right. \\
&\quad \left. \left. \left. + \dot{X}_{0,js}^\top \ddot{u}_{0,js} O_P((NT)^{-1/2}) + \dot{X}_{0,js}^\top \dot{u}_{0,js} O_P((NT)^{-1}) \right) K_h(Z_{js} - z) \right] \right\| \\
&= o_P \left( h^\varrho + \left( \frac{\log(1/h)}{(NT)h} \right)^{1/2} \right) \tag{A.12}
\end{aligned}$$

854 by the uniform convergence result of  $\hat{\beta}(z)$  above and using the similar arguments in (A.7), and

$$\begin{aligned}
& \sup_{z \in [0,1]} |S_4 + S_{10}| = \sup_{z \in [0,1]} \left\| \left[ \sum_{j=1}^N \sum_{s=1}^T K_h(Z_{js} - z) \right]^{-1} \right. \\
&\quad \times \left. \left[ \sum_{j=1}^N \sum_{s=1}^T \left( \dot{u}_{0,js}^2 O_P((NT)^{-1}) + 2\ddot{u}_{0,js} \dot{u}_{0,js} O_P((NT)^{-1/2}) \right) K_h(Z_{js} - z) \right] \right\| \\
&= o_P \left( h^\varrho + \left( \frac{\log(1/h)}{(NT)h} \right)^{1/2} \right)
\end{aligned}$$

855 by using the similar arguments in (A.10). Finally, it is

$$\begin{aligned}
\sup_{z \in [0,1]} |S_3| &= \sup_{z \in [0,1]} \left\| \left[ \sum_{j=1}^N \sum_{s=1}^T K_h(Z_{js} - z) \right]^{-1} \left[ \sum_{j=1}^N \sum_{s=1}^T \ddot{u}_{0,js}^2 K_h(Z_{js} - z) \right] \right\| \\
&= \sigma_{v,0}^2(z) + O_P \left( h^\varrho + \left( \frac{\log(1/h)}{(NT)h} \right)^{1/2} \right). \quad \square
\end{aligned}$$

## 856 A.2.2 Proofs of Corollaries 2.1 to 2.6

857 For the proofs of Corollaries 2.1 to 2.6, the uniform negligibility of the difference of using  $\hat{\delta}$  and  
858  $\delta$  is the main focus. Once the uniform negligibility is established, the rest of the proofs can be  
859 found in Wang and Xia (2009).

860 **Proof of Corollary 2.1:** In this proof, the uniform negligibility between  $\hat{\beta}_\lambda(z)$  and  $\tilde{\beta}_\lambda(z)$  is  
861 shown as follows

$$\sup_{z \in [0,1]} \left\| \hat{\beta}_\lambda(z) - \tilde{\beta}_\lambda(z) \right\| = o_P(1),$$

862 where

$$\tilde{\beta}_\lambda(z) = \left( \sum_{j=1}^N \sum_{s=1}^T \ddot{X}_{js}^\top \ddot{X}_{js} K_h(Z_{js} - z) + \hat{\mathcal{D}} \right)^{-1} \left( \sum_{j=1}^N \sum_{s=1}^T \ddot{X}_{js}^\top \ddot{y}_{js} K_h(Z_{js} - z) \right).$$



863 We rewrite  $\hat{\beta}_\lambda(z)$  by using (A.11) as follows

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

864 This leads to  $\sup_{z \in [0,1]} \|\hat{\beta}_\lambda(z) - \tilde{\beta}_\lambda(z)\| = o_P\left(h^\varrho + \left(\frac{\log(1/h)}{(NT)h}\right)^{1/2}\right)$  by using the similar arguments in  
865 the proof of Theorem 2.2. The rest of the proof can be found in Wang and Xia (2009).  $\square$

866 **Proof of Corollary 2.2:** The proof of this Corollary can be established by showing that

$$\sup_{z \in [0,1]} \|\hat{\beta}_{\lambda,a}(z) - \tilde{\beta}_{\lambda,a}(z)\| + \sup_{z \in [0,1]} \|\hat{\beta}_a(z) - \tilde{\beta}_a(z)\| = o_P\left(h^\varrho + \left(\frac{\log(1/h)}{(NT)h}\right)^{1/2}\right),$$

867 where  $\tilde{\beta}_{\lambda,a}(z)$  and  $\tilde{\beta}_a(z)$  are defined in the same manner to those of  $\hat{\beta}_{\lambda,a}(z)$  and  $\hat{\beta}_a(z)$  with  $\delta$   
868 instead of  $\hat{\delta}$ , respectively. Since it is

$$\begin{aligned}\sup_{z \in [0,1]} \|\hat{\beta}_{\lambda,a}(z) - \hat{\beta}_a(z)\| &\leq \sup_{z \in [0,1]} \|\hat{\beta}_{\lambda,a}(z) - \tilde{\beta}_{\lambda,a}(z)\| + \sup_{z \in [0,1]} \|\hat{\beta}_a(z) - \tilde{\beta}_a(z)\| \\ &\quad + \sup_{z \in [0,1]} \|\tilde{\beta}_{\lambda,a}(z) - \tilde{\beta}_a(z)\|,\end{aligned}$$

869 where  $\sup_{z \in [0,1]} \|\hat{\beta}_{\lambda,a}(z) - \tilde{\beta}_{\lambda,a}(z)\| = \sup_{z \in [0,1]} \|\hat{\beta}_a(z) - \tilde{\beta}_a(z)\| = o_P\left(h^\varrho + \left(\frac{\log(1/h)}{(NT)h}\right)^{1/2}\right)$ , the proof is  
870 completed by using the similar arguments to those in the proof of Corollary 6.1.  $\square$

871 **Proof of Corollary 2.3:** In this proof, we establish the uniform negligibility of the following

$$\sup_{z \in [0,1]} |RSS - \widetilde{RSS}| = o_P\left(h^\varrho + \left(\frac{\log(1/h)}{(NT)h}\right)^{1/2}\right),$$

872 where  $RSS$  is defines in (2.23) and

$$\widetilde{RSS} = \frac{1}{(NT)^2} \sum_{i=1}^N \sum_{t=1}^T \sum_{j=1}^N \sum_{s=1}^T \left\{ \ddot{y}_{js} - \ddot{X}_{js} \tilde{\beta}_\lambda(Z_{it}) \right\}^2 K_h(Z_{js} - Z_{it}).$$

873 By using (A.11) and the result of Corollary 2.1,  $RSS$  can be rewritten below

$$RSS \doteq \widetilde{RSS} + \mathcal{R}_{js},$$

874 where  $\mathcal{R}_{js} \doteq \mathcal{R}_{1,js} + 2\mathcal{R}_{2,js}$  with

$$\begin{aligned}\sup_{z \in [0,1]} |\mathcal{R}_{1,js}| &= \sup_{z \in [0,1]} \left| \frac{1}{(NT)^2} \sum_{i=1}^N \sum_{t=1}^T \sum_{j=1}^N \sum_{s=1}^T \left\{ \dot{y}_{js} - \dot{X}_{js} \tilde{\beta}_\lambda(Z_{it}) O_P((NT)^{-1/2}) \right\}^2 K_h(Z_{js} - Z_{it}) \right| \\ &= o_P\left(h^\varrho + \left(\frac{\log(1/h)}{(NT)h}\right)^{1/2}\right),\end{aligned}$$

875 and

$$\begin{aligned} \sup_{z \in [0,1]} |\mathcal{R}_{2,js}| &= \sup_{z \in [0,1]} \left| \frac{1}{(NT)^2} \sum_{i=1}^N \sum_{t=1}^T \sum_{j=1}^N \sum_{s=1}^T \left\{ \ddot{y}_{js} - \ddot{X}_{js} \tilde{\beta}_\lambda(Z_{it}) \right\} \right. \\ &\quad \times \left. \left\{ \dot{y}_{js} - \dot{X}_{js} \tilde{\beta}_\lambda(Z_{it}) O_P((NT)^{-1/2}) \right\} K_h(Z_{js} - Z_{it}) \right| \\ &= o_P \left( h^\varrho + \left( \frac{\log(1/h)}{(NT)h} \right)^{1/2} \right) \end{aligned}$$

876 by using the similar arguments to those of (A.12). The proof is completed.  $\square$

877 **Proof of Corollary 2.4:** Corollary 2.4 is established by showing that

$$\sup_{z \in [0,1]} \|\hat{\beta}_\lambda^{(m+1)}(z) - \tilde{\beta}_\lambda^{(m+1)}(z)\| = o_P \left( h^\varrho + \left( \frac{\log(1/h)}{(NT)h} \right)^{1/2} \right),$$

878 where

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

879 The above uniform negligibility is established by using the similar arguments to those of Corollary  
880 2.1 and the rest of the proof can be found in (Wang and Xia (2009)).  $\square$

881 **Proof of Corollary 2.5:** The proof of Corollary 2.5 can be shown similarly to that of Corollary  
882 2.2 as follows

$$\sup_{z \in [0,1]} \|\hat{\beta}_{\lambda,a}^{(m+1)}(z) - \tilde{\beta}_{\lambda,a}^{(m+1)}(z)\| + \sup_{z \in [0,1]} \|\hat{\beta}_a(z) - \tilde{\beta}_a(z)\| = o_P \left( h^\varrho + \left( \frac{\log(1/h)}{(NT)h} \right)^{1/2} \right),$$

883 where

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

884 **Proof of Corollary 2.6:** The uniform negligibility,  $\sup_{z \in [0,1]} \|\hat{\beta}(z) - \tilde{\beta}(z)\| = o_P \left( h^\varrho + \left( \frac{\log(1/h)}{(NT)h} \right)^{1/2} \right)$ ,  
885 is shown by using the similar arguments to those of Corollary 2.1. Recall also that these are the  
886 estimators defined in (2.7) and (2.12), respectively.  $\square$

## 887 B Additional discussions

### 888 B.1 Discussion and justification of conditions in Assumption B

889 Assumptions B1 to B5 are standard regularity conditions in the semi/nonparametric literature.  
890 Assumption B1 (i) is required to achieve the parametric convergence rate for estimating the finite

coefficients (point parameters), and is commonly used in the semiparametric literature (see e.g., Robinson (1988), and Fan and Li (1999)). Additionally, the order of the derivatives in Assumptions B3 to B5 is  $\varrho > 2$  corresponding to the order of the kernel function in Assumption B1. These higher-order derivatives conditions are necessary in order to obtain the parametric convergence rate for our proposed QML estimator. Therefore, it is that  $\lim_{N \rightarrow \infty} Nh^{2\varrho} \rightarrow 0$  and this higher-order condition also provides the lower bound for  $\eta$ , namely  $\frac{1}{2} + \frac{1}{4\varrho} < \eta$ , satisfying  $\lim_{N \rightarrow \infty} N^{2\eta-1}h \rightarrow \infty$  which is the necessary condition for the uniform consistency of the nonparametric estimation (see e.g., Mack and Silverman (1982)). Furthermore, Assumptions B2 to B5 are necessary conditions in the varying coefficient literature (see e.g., Cai et al. (2000), Fan and Zhang (2000), and Xia et al. (2004)). In particular, Assumption B4 ensures that the observed index values are sufficiently dense implying that the maximal distance between two consecutive index variables is only of the order  $O_P\left(\frac{\log NT}{NT}\right)$ . For an arbitrarily index value  $z \in [0, 1]$ , let  $z^*$  be the nearest neighbor among the observed index values,  $z^* = \arg \min_{\tilde{z} \in \{Z_{it}: 11 \leq it \leq NT\}} |z - \tilde{z}|$ . Assumption B5 imposes a smoothness condition on the functional coefficients which implies that  $\|\beta_0(z) - \beta_0(z^*)\| = O_P\left(\frac{\log NT}{NT}\right)$  (see Xia et al. (2004) for full discussion).

## B.2 Asymptotic behaviors of our test statistics

To formally state the asymptotic behaviors of our test statistic, we write  $C = (C_d e_{NT})^\top$ , where  $e_{NT}$  is  $NT \times 1$  vector of ones, and  $\bar{\beta} = \{\beta_{d,0}(X_{11}), \dots, \beta_{d,0}(X_{NT})\}^\top \in \mathbb{R}^{NT}$ . Also, let  $A_{C_d} = \left[\frac{1}{NT} \|\bar{\beta} - C\|^2\right]^{1/2}$  denote the normalized  $l_2$  distance, such that  $A_{C_d} \geq C_A$  if the  $H_0$  is false, for some constant  $C_A > 0$ . Under Assumptions A to C and the procedure introduced in Section 2.2, Corollary 2.6 ensures that

$$\lim_{N \rightarrow \infty} P(T_d > c_\alpha) = 1.$$

In other words, the power of our test is nontrivial and approaching one as the number of observations increases. Moreover, it is the case that

$$\lim_{N \rightarrow \infty} P(T_d > c_\alpha) = 1$$

under  $H_0$ , i.e., the size value approaches the significance level.

# C Tables and Figures

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,  $\hat{\rho}_{or}$ .

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.

Table 3: Nonparametric estimation of the coefficient functions

| Model I | $N = 100$  |            | $N = 200$  |            | $N = 300$  |            |
|---------|------------|------------|------------|------------|------------|------------|
|         | $REE_{or}$ | $REE_{un}$ | $REE_{or}$ | $REE_{un}$ | $REE_{or}$ | $REE_{un}$ |
| $P = 2$ | 1.278      | 0.391      | 1.079      | 0.357      | 1.032      | 0.350      |
| $P = 5$ | 1.289      | 0.393      | 1.075      | 0.354      | 1.032      | 0.350      |
| $P = 8$ | 1.291      | 0.394      | 1.074      | 0.354      | 1.030      | 0.349      |

| Model II | $N = 100$  |            | $N = 200$  |            | $N = 300$  |            |
|----------|------------|------------|------------|------------|------------|------------|
|          | $REE_{or}$ | $REE_{un}$ | $REE_{or}$ | $REE_{un}$ | $REE_{or}$ | $REE_{un}$ |
| $P = 2$  | 1.048      | 0.608      | 1.011      | 0.604      | 0.998      | 0.599      |
| $P = 5$  | 1.057      | 0.611      | 1.014      | 0.605      | 0.998      | 0.601      |
| $P = 8$  | 1.057      | 0.609      | 1.013      | 0.604      | 0.997      | 0.600      |

Table 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     |

Table 5: Hypothesis test of the 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            |
| $H_0 : \beta_2(z) = C_2$ | 67          | 63             | 81          | 79             | 94          | 85             |
| $H_0 : \beta_3(z) = C_3$ | 69          | 65             | 77          | 76             | 80          | 76             |
| $H_0 : \beta_4(z) = 0$   | 74          | 67             | 85          | 79             | 90          | 86             |
| $H_0 : \beta_5(z) = 0$   | 73          | 66             | 85          | 77             | 83          | 79             |
| $H_0 : \beta_6(z) = 0$   | 73          | 72             | 82          | 74             | 83          | 79             |
| $H_0 : \beta_7(z) = 0$   | 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            |
| $H_0 : \beta_2(z) = C_2$ | 65          | 63             | 81          | 81             | 94          | 85             |
| $H_0 : \beta_3(z) = C_3$ | 66          | 65             | 77          | 75             | 80          | 75             |
| $H_0 : \beta_4(z) = 0$   | 75          | 70             | 83          | 79             | 90          | 85             |
| $H_0 : \beta_5(z) = 0$   | 73          | 66             | 85          | 77             | 82          | 77             |
| $H_0 : \beta_6(z) = 0$   | 74          | 72             | 83          | 71             | 83          | 77             |
| $H_0 : \beta_7(z) = 0$   | 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 utilized in order to obtain efficiency gain.

Figure 1: Per capita measure of MHS spending for persons age between 18 and 64 (*mhs*)

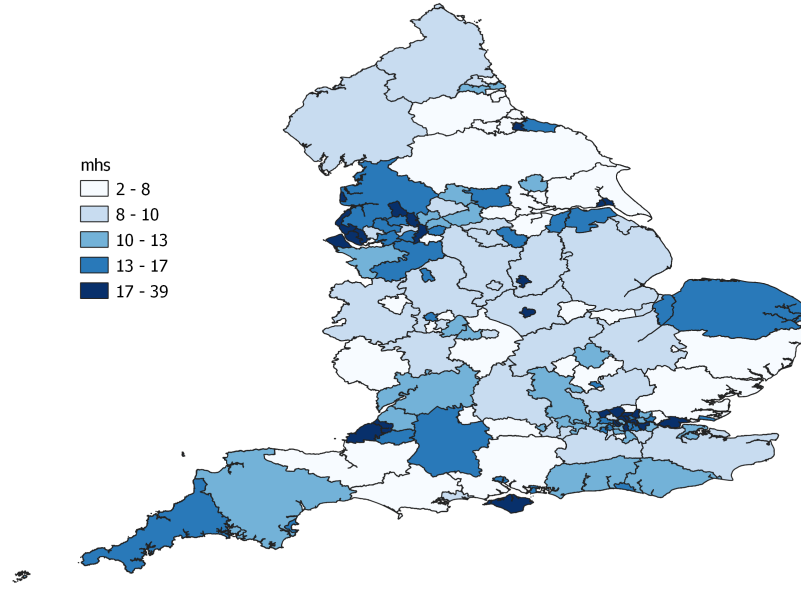
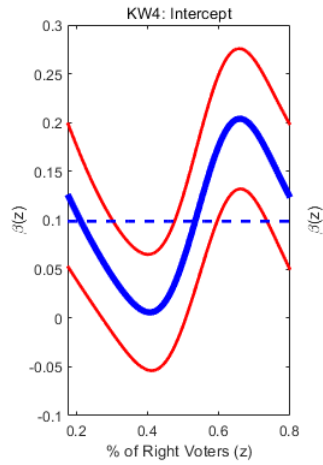


Figure 2: Estimates coefficient function of the intercept:  $Z$  represents *vote*



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



Table 6: Our data and its sources

| Symbols     | Descriptions                                                                                                                                                                                                                                                        |
|-------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>vote</i> | Percentage of voters with right-wing ideology<br>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-standardized total public health by local authority<br>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 standardized MHS for persons age between 18 and 64<br>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<br>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<br>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<br>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-standardized mortality rates for 2016 to 2019 standardized to the 2013 European Standard Population expressed per 100,000 population<br>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<br>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<br>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<br>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<br>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<br>Source: Estimates of the population for the UK available at <a href="http://www.ons.gov.uk">www.ons.gov.uk</a>                                                                                     |

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 |

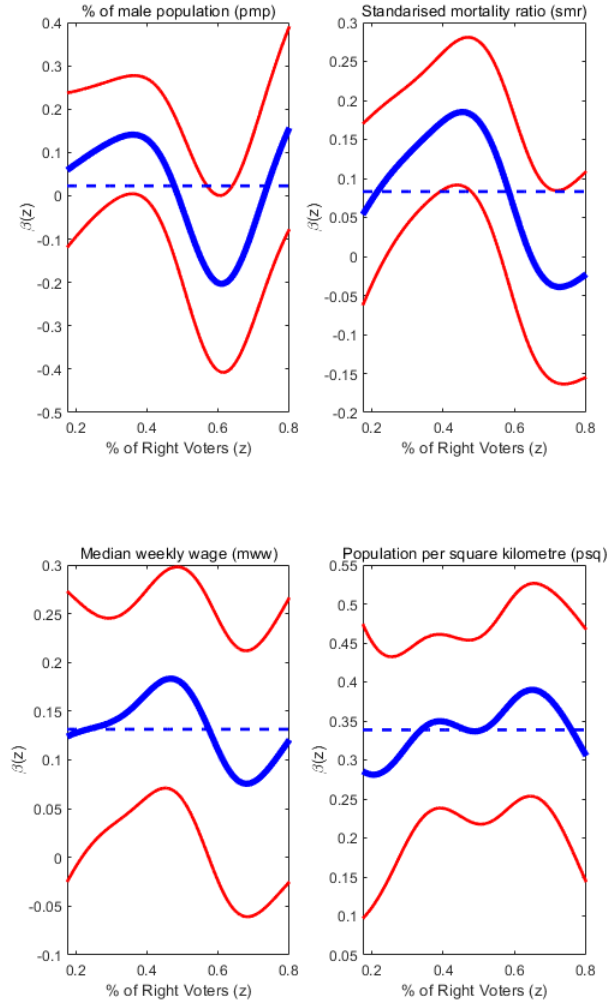
Table 8: Estimation results

$Z$  is defined as percentage of right-wing voters (*vote*)

| $W$  | $\rho$ | $\phi$ | $\hat{K}$ | <i>inct</i> | <i>nuc</i> | <i>pmp</i> | <i>pu14</i> | <i>smr</i> | <i>noj</i> | <i>plp</i> | <i>mhp</i> | <i>mww</i> | <i>psq</i> |
|------|--------|--------|-----------|-------------|------------|------------|-------------|------------|------------|------------|------------|------------|------------|
| KW4  | 0.16   | 1.77   | 5         | ✖           | ·          | ×          | ·           | ×          | ·          | ·          | ·          | ×          | ×          |
|      |        |        |           | 3.83        | ·          | 2.56       | ·           | 2.28       | ·          | ·          | ·          | 0.15       | 0.72       |
| KW10 | 0.21   | 1.78   | 5         | ✖           | ·          | ✖          | ·           | ×          | ·          | ·          | ·          | ×          | ×          |
|      |        |        |           | 3.74        | ·          | 3.42       | ·           | 2.14       | ·          | ·          | ·          | 0.45       | 1.18       |
| KW16 | 0.27   | 1.81   | 5         | ✖           | ·          | ✖          | ·           | ×          | ·          | ·          | ·          | ×          | ×          |
|      |        |        |           | 3.27        | ·          | 3.45       | ·           | 1.87       | ·          | ·          | ·          | 0.13       | 1.24       |
| SW   | 0.21   | 1.75   | 5         | ✖           | ·          | ✖          | ·           | ×          | ·          | ·          | ·          | ×          | ×          |
|      |        |        |           | 3.63        | ·          | 3.75       | ·           | 2.34       | ·          | ·          | ·          | 0.37       | 0.75       |
| KW0  | ·      | ·      | ·         | ✖           | ·          | ✖          | ✖           | ✖          | ✖          | ✖          | ✖          | ✖          | ✖          |
|      |        |        |           | 13.07       | ·          | 8.76       | 4.96        | 10.85      | 10.743     | 6.547      | 4.292      | 9.829      | 8.58       |

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.

Figure 3: 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  $\hat{C}_d$  in (4.4).

## References

- Amemiya, T. (1985). *Advanced Econometrics*. Harvard university press.
- Anselin, L. (2009). Spatial regression. *The SAGE Handbook of Spatial Analysis 1*, 255–276.
- Antecol, H. and K. Bedard (2006). Unhealthy assimilation: why do immigrants converge to american health status levels? *Demography* 43(2), 337–360.
- Aziz, F., P. McCrone, S. Boyle, and M. Knapp (2003). Financing mental health services in london: central funding and local expenditure.
- Baltagi, B. H., G. Bresson, and A. Pirotte (2012). Forecasting with spatial panel data. *Computational Statistics & Data Analysis* 56(11), 3381–3397.
- Baltagi, B. H., S. H. Song, and W. Koh (2003). Testing panel data regression models with spatial error correlation. *Journal of Econometrics* 117(1), 123–150.
- Cai, Z., J. Fan, and R. Li (2000). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association* 95(451), 888–902.
- Chen, J., J. Gao, and D. Li (2012). Semiparametric trending panel data models with cross-sectional dependence. *Journal of Econometrics* 171(1), 71–85.
- Department of Health (2010). *Healthy lives, healthy people: Our strategy for public health in England*. London, UK: Department of Health.
- Dong, C., J. Gao, and B. Peng (2021). Varying-coefficient panel data models with nonstationarity and partially observed factor structure. *Journal of Business & Economic Statistics* 39(3), 700–711.
- Dunleavy, P., A. Park, and R. Taylor (2018). *The UK’s changing democracy: the 2018 Democratic Audit*. LSE Press.
- Fan, J., M. Farnen, and I. Gijbels (1998). Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60(3), 591–608.
- Fan, J. and W. Zhang (2000). Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scandinavian Journal of Statistics* 27(4), 715–731.
- Fan, J. and W. Zhang (2008). Statistical methods with varying coefficient models. *Statistics and Its Interface* 1(1), 179.
- Fan, Y. and Q. Li (1999). Root-n-consistent estimation of partially linear time series models. *Journal of Nonparametric Statistics* 11(1-3), 251–269.

946 Feng, G., J. Gao, B. Peng, and X. Zhang (2017). A varying-coefficient model with fixed effects:  
947 theory and an application to us commercial banks. *Journal of Econometrics* 196(1), 68–82.

948 Fingleton, B. (2009). Prediction using panel data regression with spatial random effects. *Inter-*  
949 *national Regional Science Review* 32(2), 195–220.

950 Giuntella, O., Z. Kone, I. Ruiz, and C. Vargas-Silva (2018). Reason for immigration and immi-  
951 grants’ health. *Public Health* 158, 102–109.

952 Hu, T. and Y. Xia (2012). Adaptive varying coefficient model selection. *Statistica Sinica*, 575–599.

953 Hunter, D. and R. Li (2005). Variable selection using mm. *Annals of Statistics* (33), 1617.

954 Ihlanfeldt, K. and T. Mayock (2010). Panel data estimates of the effects of different types of  
955 crime on housing prices. *Regional Science and Urban Economics* 40(2-3), 161–172.

956 Jass, G. and D. S. Massey (2004). Immigrant health: Selectivity and acculturation. Technical  
957 report, IFS Working Papers.

958 Kapoor, M., H. H. Kelejian, and I. R. Prucha (2007). Panel data models with spatially correlated  
959 error components. *Journal of Econometrics* 140(1), 97–130.

960 Kelejian, H. H. and I. R. Prucha (1999). A generalized moments estimator for the autoregressive  
961 parameter in a spatial model. *International Economic Review* 40(2), 509–533.

962 Kim, N., P. Wongsart, and I. J. Bateman (2023). A new model for agricultural land-use modeling  
963 and prediction in england using spatially high-resolution data. In *Essays in Honor of Joon Y.*  
964 *Park: Econometric Methodology in Empirical Applications*, Volume 45, pp. 291–317. Emerald  
965 Publishing Limited.

966 Leach, C., K. Finning, G. Kerai, and T. Vizard (2022). Coronavirus and depression in adults,  
967 great britain: July to august 2021. office for national statistics. 2021.

968 Lee, L. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial  
969 autoregressive models. *Econometrica* 72(6), 1899–1925.

970 Lee, L.-f. and J. Yu (2010). Estimation of spatial autoregressive panel data models with fixed  
971 effects. *Journal of Econometrics* 154(2), 165–185.

972 Li, D., J. Chen, and J. Gao (2011). Non-parametric time-varying coefficient panel data models  
973 with fixed effects. *The Econometrics Journal* 14(3), 387–408.

974 Li, D.-K., C.-L. Mei, and N. Wang (2019). Tests for spatial dependence and heterogeneity in spa-  
975 tially autoregressive varying coefficient models with application to boston house price analysis.  
976 *Regional Science and Urban Economics* 79, 103470.

- 977 Liang, X., J. Gao, and X. Gong (2022). Semiparametric spatial autoregressive panel data  
978 model with fixed effects and time-varying coefficients. *Journal of Business & Economic Statis-*  
979 *tics* 40(4), 1784–1802.
- 980 Liu, S. F. and Z. Yang (2015). Asymptotic distribution and finite sample bias correction of qml  
981 estimators for spatial error dependence model. *Econometrics* 3(2), 376–411.
- 982 Mack, Y.-p. and B. W. Silverman (1982). Weak and strong uniform consistency of kernel regression  
983 estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 61(3), 405–415.
- 984 Maconick, L., L. Sheridan Rains, R. Jones, B. Lloyd-Evans, and S. Johnson (2021). Investigating  
985 geographical variation in the use of mental health services by area of england: a cross-sectional  
986 ecological study. *BMC Health Services Research* 21(1), 1–10.
- 987 Magnus, J. R. and C. Muris (2010). Specification of variance matrices for panel data models.  
988 *Econometric Theory* 26(1), 301–310.
- 989 Manski, C. F. (1993). Identification problems in the social sciences. *Sociological Methodology*,  
990 1–56.
- 991 McCrone, P. and B. Jacobson (2004). Indicators of mental health activity in london: Adjusting  
992 for sociodemographic need. *London: London Development Centre for Mental Health*.
- 993 McManus, S., P. E. Bebbington, R. Jenkins, and T. Brugha (2016). *Mental health and wellbeing*  
994 *in England: the adult psychiatric morbidity survey 2014*. NHS digital.
- 995 Meng, Y., J. Gao, X. Zhang, and X. Zhao (2021). A panel data model of length of stay in hospitals  
996 for hip replacements. *Econometric Reviews* 40(7), 688–707.
- 997 Moscone, F. and M. Knapp (2005). Exploring the spatial pattern of mental health expenditure.  
998 *Journal of Mental Health Policy and Economics* 8(4), 205.
- 999 Moscone, F., M. Knapp, and E. Tosetti (2007). Mental health expenditure in england: a spatial  
1000 panel approach. *Journal of Health Economics* 26(4), 842–864.
- 1001 Moscone, F. and E. Tosetti (2010). Testing for error cross section independence with an application  
1002 to us health expenditure. *Regional Science and Urban Economics* 40(5), 283–291.
- 1003 Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Hand-*  
1004 *book of Econometrics* 4, 2111–2245.
- 1005 Office of National Statistics (2020). *Families and households in the UK: 2020*. London, UK: Office  
1006 of National Statistics.

- 1007 Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, 931–954.
- 1008 Robinson, P. M. (2011). Asymptotic theory for nonparametric regression with spatial data.  
1009 *Journal of Econometrics* 165(1), 5–19.
- 1010 Sarafidis, V. and T. Wansbeek (2021). Celebrating 40 years of panel data analysis: Past, present  
1011 and future. *Journal of Econometrics* 220(2), 215–226.
- 1012 Sun, Y. and E. Malikov (2018). Estimation and inference in functional-coefficient spatial autore-  
1013 gressive panel data models with fixed effects. *Journal of Econometrics* 203(2), 359–378.
- 1014 Tomal, M. (2020). Modelling housing rents using spatial autoregressive geographically weighted  
1015 regression: A case study in cracow, poland. *ISPRS International Journal of Geo-*  
1016 *Information* 9(6), 346.
- 1017 Tomal, M. and M. Helbich (2023). A spatial autoregressive geographically weighted quantile  
1018 regression to explore housing rent determinants in amsterdam and warsaw. *Environment and*  
1019 *Planning B: Urban Analytics and City Science* 50(3), 579–599.
- 1020 Walker, I. F., J. Stansfield, L. Makurah, H. Garnham, C. Robson, C. Lugton, N. Hey, and G. Hen-  
1021 derson (2019). Delivering national public mental health–experience from england. *Journal of*  
1022 *Public Mental Health* 18(2), 112–123.
- 1023 Wang, H. and C. Leng (2007). Unified lasso estimation by least squares approximation. *Journal*  
1024 *of the American Statistical Association* 102(479), 1039–1048.
- 1025 Wang, H. and Y. Xia (2009). Shrinkage estimation of the varying coefficient model. *Journal of*  
1026 *the American Statistical Association* 104(486), 747–757.
- 1027 White, H. (1996). *Estimation, inference and specification analysis*. Number 22. Cambridge uni-  
1028 versity press.
- 1029 Xia, Y., W. Zhang, and H. Tong (2004). Efficient estimation for semivarying-coefficient models.  
1030 *Biometrika* 91(3), 661–681.
- 1031 Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables.  
1032 *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68(1), 49–67.
- 1033 Zhang, Y. and D. Shen (2015). Estimation of semi-parametric varying-coefficient spatial panel  
1034 data models with random-effects. *Journal of Statistical Planning and Inference* 159, 64–80.
- 1035 Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical*  
1036 *Association* (101), 1418–1429.