TEST (2015) 24:417–440
DOI 10.1007/s11749-014-0419-x

ORIGINAL PAPER

A varying-coefficient approach to estimating multi-level clustered data models

Jinhong You · Alan T. K. Wan · Shu Liu · Yong Zhou

Received: 31 August 2013 / Accepted: 10 November 2014 / Published online: 27 November 2014

Abstract Most of the literature on clustered data models emphasizes two-level clustering, and within-cluster correlation. While multi-level clustered data models can arise in practice, analysis of multi-level clustered data models poses additional difficulties owing to the existence of error correlations both within and across the clusters. It is perhaps for this reason that existing approaches to multi-level clustered data models have been mostly parametric. The purpose of this paper is to develop a varying-coefficient nonparametric approach to the analysis of three-level clustered data models. Because the nonparametric functions are restricted only to some of the variables, this approach has the appeal of avoiding many of the curse of dimensionality problems commonly associated with other nonparametric methods. By applying an undersmoothing technique, taking into account the correlations within and across clusters, we develop an efficient two-stage local polynomial estimation procedure for the unknown coefficient functions. The large and finite sample properties of the resultant estimators are examined; in particular, we show that the resultant estimators are asymptotically

Electronic supplementary material The online version of this article (doi:10.1007/s11749-014-0419-x) contains supplementary material, which is available to authorized users.

J. You · S. Liu
School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China
e-mail: johnyou07@163.com

A. T. K. Wan (✉)
Department of Management Sciences, City University of Hong Kong, Kowloon, Hong Kong
e-mail: msawan@cityu.edu.hk; Alan.Wan@cityu.edu.hk

Y. Zhou
Institute of Applied Mathematics, Chinese Academy of Sciences, Beijing, China
e-mail: yzhou@amss.ac.cn

normal, and exhibit considerably smaller asymptotic variability than the traditional local polynomial estimators that neglect the correlations within and among clusters. An application example is presented based on a data set extracted from the World Bank's STARS database.

Keywords Asymptotic normality · Correlation · Nonparametric · Clustered data · Two-stage estimation

Mathematics Subject Classification 62G08 · 62G20

1 Introduction

The analysis of clustered or longitudinal data has been the subject of one of the most active bodies of research in statistics. Examples of this type of data abound in many fields; see Diggle et al. (1994) and Baltagi (2008) for relevant examples. The widespread attention enjoyed by clustered data models is amply justified by a number of attractions, most notably, observation heterogeneity and more efficient estimation of parameters owing to the richer source of information. The jointness of the equations introduces additional information over and above what is available when the equations are treated separately. Indeed, a great deal of the literature on clustered data models has focused on ways of taking into account the interactions of the equations to improve the sharpness of inference.

The traditional literature on clustered data models has focused on two-level clustering. In practice, three or higher level clustering also has sound practical applications. Using data from 1960–1969 to 1980–1987, taken from the World Bank's STARS database, the example in Sect. 7 considers the estimation of production functions for eighty-one different countries. For each country during both time periods, gross domestic product is determined by real capital, labour supply, and the average level of schooling of the work force. Given the likelihood that similar factors may be responsible for the random effects, it seems reasonable to suppose that the error terms associated with the production functions of different countries are contemporaneously correlated. We also suspect an inherent jointness of the equations within the same country across the different years of the same period, and across the two periods under consideration. Our suspicion is confirmed by the test results of Sect. 7, which show that the error correlations both within and across clusters are significant. This example clearly illustrates that three-level data clustering arises commonly in practice. Indeed, there is a strong body of empirical literature emphasizing three-level clustering. For example, Beierlein et al. (1981) considered demand models of electricity and natural gas for commercial, industrial and residential sectors over a period of ten years in several northeastern US states; Chapman et al. (2003) discussed a clinical trial where various biomarkers for kidney malfunctioning were collected from different patients over a specified time period. Other studies involving multi-level clustered data models abound, but these examples illustrate the range of applications for which this model is appropriate.

The analysis of three-level clustered data models is complicated by the existence of error dependency both within and across clusters, whereas with conventional two-level

models the latter type of dependency is nonexistent. In fact, the majority of nonparametric estimation methods developed for two-level clustered models ignore within cluster dependency; e.g., Hoover et al. (1998). This can be justified theoretically when the cluster size is finite and a kernel approach is used in estimating the nonparametric function, but does not apply under other circumstances (Wang 2003). In addition, Welsh et al. (2002) showed that for time varying covariates, spline methods that take into account the within-cluster correlations work better than kernel methods that ignore this dependency. There has been some recent interest in the development of alternative nonparametric and semiparametric estimation methods that properly account for the within-cluster error dependency for two-level clustered data models. See the work of Wang (2003), Fan et al. (2007), and Zhang (2009).

In the case of three-level clustered data models, empirical studies utilizing these models are all based on parametric approaches. This is not surprising given the absence of a rigorous theoretical nonparametric literature on this subject. For example, the aforementioned study of Beierlein et al. (1981) employed an estimation method that is a generalization of Zellner (1962) seemingly unrelated parametric regression approach. To our knowledge, the recent contributions of Chen and Zhong (2011) and Zhou et al. (2011) are the only theoretical nonparametric studies with an explicit focus on the multi-level clustered models. Chen and Zhong (2011) developed ANOVA tests for multi-treatment comparisons in clustered data models, allowing for the possibility of missing data; however, their model assumes independence of errors both within and across clusters. Assuming three-level clustering and an error components specification of the disturbances, Zhou et al. (2011) developed a nonparametric seemingly unrelated regression approach to the estimation of the unknown coefficient functions. The approach of Zhou et al. is based on undersmoothing and an estimation of the covariance matrix of the errors, taking into account both within and across errors dependency, which improves the performance of the nonparametric function estimators. One limitation of this approach, however, is that the disturbances are restricted to an error components' specification.

In this paper, we develop a varying-coefficient approach to the estimation of a three-level clustered data model. The varying-coefficient model, introduced by Cleveland et al. (1991) and Hastie and Tibshiran (1993), has the important appeal that it allows the coefficients that describe the effect of the regressor to vary nonparametrically as a function of other variables, called the effect modifiers. Because smoothing is applied only to the effect modifiers, the curse of dimensionality problem commonly associated with other nonparametric approaches can be avoided. The varying-coefficient model has been an important development in the nonparametric statistics literature, and has received much attention for its ability to avoid the difficulties caused by high-dimension covariates. Throughout this paper, we work with the following three-level clustered varying-coefficient model:

$$\begin{aligned} Y_{\text{sit}} &= X_{sit1}\alpha_{s1}(U_{\text{sit}}) + \cdots + X_{\text{sit}p_s}\alpha_{sp_s}(U_{\text{sit}}) + \varepsilon_{\text{sit}}, \\ &\quad \text{for } s = 1, \ldots, m, i = 1, \ldots, n,\ t = 1, \ldots, T_s, \end{aligned} \tag{1.1}$$

where Y_{sit} is the observed measurement of the sth response variable of the ith individual at the tth observation point; $(\mathbf{X}_{\text{sit}}^{\tau}, U_{\text{sit}})^{\tau} = (X_{\text{sit}1}, \ldots, X_{\text{sit}p_s}, U_{\text{sit}})^{\tau}$ is the observ-

able covariate; $\boldsymbol{\alpha}(\cdot) = (\boldsymbol{\alpha}_1^{\tau}(\cdot), \ldots, \boldsymbol{\alpha}_m^{\tau}(\cdot))^{\tau} = (\alpha_{11}(\cdot), \ldots, \alpha_{1p_1}(\cdot), \ldots, \alpha_{mp_m}(\cdot))^{\tau}$ are unknown functions with p_s being the number of nonparametric varying coefficient functions in the sth group. In addition, ε_{sit} is the random error with $E(\boldsymbol{\varepsilon}_{\cdot i\cdot}) = 0$, $Cov(\boldsymbol{\varepsilon}_{\cdot i\cdot}) = \Sigma = (\Sigma_{s_1 s_2})_{s_1,s_2=1}^{m} > 0$ and $E(\boldsymbol{\varepsilon}_{\cdot i_1\cdot}\boldsymbol{\varepsilon}_{\cdot i_2\cdot}^{\tau}) = \mathbf{0}_{\sum_{s=1}^{m} T_s \times \sum_{s=1}^{m} T_s}$ for $i_1 \neq i_2$, where ε_{sit} is independent of the covariate $(\mathbf{X}_{\text{sit}}^{\tau}, U_{\text{sit}})^{\tau}$ and $\boldsymbol{\varepsilon}_{\cdot i\cdot} = (\varepsilon_{1i1}, \ldots, \varepsilon_{1iT_1}, \ldots, \varepsilon_{mi1}, \ldots, \varepsilon_{miT_m})^{\tau}$. Model (1.1) says that the coefficients of X change with U, the effect modifier. We assume that U is a single variable, but in general U can be a low-dimensional vector of variables. Because only low-dimensional functions are estimated, the model avoids the curse of dimensionality difficulties even if p_s is large. Also, note that (1.1) allows for unbalanced panels as T_s need not be the same for each s. In addition, the form of Σ is not restrictive, and this makes (1.1) flexible. The popular one-way error component random effect covariance structure is a special case of Σ by setting $\varepsilon_{\text{sit}} = \mu_{si} + \nu_{\text{sit}}$, where μ_{si}s' and ν_{sit}s' are i.i.d. over i and (i, t) respectively, with both μ_{si} and ν_{sit} having a zero mean, $\text{var}(\mu_{si}) = \sigma_{s\mu}^2$ and $\text{var}(\nu_{\text{sit}}) = \sigma_{s\nu}^2$. Then, $\Sigma_{ss} = \sigma_{s\mu}^2 \mathbf{1}_{T_s}\mathbf{1}_{T_s}^{\tau} + \sigma_{s\nu}^2 \mathbf{I}_{T_s \times T_s}$, where $\mathbf{1}_{T_s}$ is a T_s-dimensional unit vector and $\mathbf{I}_{T_s \times T_s}$ is a $T_s \times T_s$ identity matrix.

Model (1.1) is fairly general and includes a variety of data models as special cases. Our emphasis here is on the inference aspect of model (1.1). Specifically, we aim to develop methods of estimation that take into account the correlations within and across clusters. While there have been some adaptations of the varying-coefficient approach to clustered data models focusing on two-level clustering, few of the existing studies make full use of the information implied by the correlation matrix of the disturbances. For example, the estimation methods based on varying-coefficient models developed by Hoover et al. (1998) and Fan and Zhang (2000) both assume independence of errors within clusters. In addition, the estimation procedure used by Zhang (2009) only partially utilizes the within-cluster correlation by estimating the random effect functions but ignores the within-cluster correlation in the random errors. A related study by Wang (2003) developed a nonparametric estimation method that takes full account of the within-cluster correlations; however, her procedure pertains only to the two-level clustered model and is not based on the varying-coefficient approach.

The method we develop here is a two-stage procedure. In the first stage, an undersmoothing technique is applied to construct estimators that neglect the correlations within and across clusters. These "pilot estimators" form the basis of the estimation of covariance in the second stage; the final estimators of the unknown coefficient functions are then constructed by local polynomial fitting based on information contained in the estimated covariance matrix. We prove that the estimator of the unknown function resulting from this two-stage procedure is asymptotically normal and has the same magnitude of bias as (but a substantially smaller variance than) the estimator that treats each equation separately. The pilot estimators obtained from the first stage also allow the development of test statistics useful for detecting correlations within and across responses.

The layout of the remainder of this paper is as follows. In Sect. 2, we present the pilot estimators of the unknown coefficient functions and error covariance matrix, while in Sect. 3, we develop test statistics to detect correlations within and among responses. In Sect. 4, an efficient two-stage local polynomial estimator is constructed.

The asymptotic and finite sample properties of this estimator are investigated in Sects. 5 and 6 respectively. Section 7 contains a real data analysis. In Sect. 8, we offer some concluding remarks. Proofs of results are contained in an on-line supplementary file.

2 Pilot estimators of unknown coefficient functions and error covariance matrix

For a fixed s, if correlations within the response are ignored, a local linear regression technique can be employed to estimate the coefficient functions $\{\alpha_{sj}(\cdot), j = 1, \ldots, p_s\}$ (e.g., Fan and Gijbels 1996). The procedure works as follows: consider that U_{sit} in the close neighbourhood of the local value u, $\alpha_{sj}(U_{\text{sit}})$ can be approximated locally by

$$\alpha_{sj}(U_{\text{sit}}) \approx \alpha_{sj}(u) + \alpha'_{sj}(u)(U_{\text{sit}} - u) \equiv a_{sj} + b_{sj}(U_{\text{sit}} - u), \quad j = 1, \ldots, p_s,$$

where $\alpha'_{sj}(u) = \partial\alpha_{sj}(u)/\partial u$. This leads to the following weighted local least-squares problem. Find $\{(a_{sj}, b_{sj}), j = 1, \ldots, p_s\}$ that minimizes

$$\sum_{i=1}^{n}\sum_{t=1}^{T_s}\left[Y_{\text{sit}} - \sum_{j=1}^{p_s}\{a_{sj} + b_{sj}(U_{\text{sit}} - u)\}X_{\text{sit}j}\right]^2 K_{h_s}(U_{\text{sit}} - u), \tag{2.1}$$

where $K_{h_s}(\cdot) = K(\cdot/h_s)/h_s$, $K(\cdot)$ is a kernel function and h_s a bandwidth. The solution to problem (2.1) is given by

$$\left(\tilde{a}_{s1}, \ldots, \tilde{a}_{sp_s}, \tilde{b}_{s1}, \ldots, \tilde{b}_{sp_s}\right)^\tau = (\mathbf{D}_{su}^\tau \mathbf{W}_{su}\mathbf{D}_{su})^{-1}\mathbf{D}_{su}^\tau \mathbf{W}_{su}\mathbf{Y}_s, \tag{2.2}$$

where

$$\mathbf{D}_{su} = \begin{pmatrix} \mathbf{X}_{s11}, & \cdots, & \mathbf{X}_{s1T_s}, & \cdots, & \mathbf{X}_{snT_s} \\ (U_{s11} - u)\mathbf{X}_{s11}, & \cdots, & (U_{s1T_s} - u)\mathbf{X}_{s1T_s}, & \cdots, & (U_{snT_s} - u)\mathbf{X}_{snT_s}^\tau \end{pmatrix}^\tau,$$

$$\mathbf{W}_{su} = \text{diag}(K_{h_s}(U_{s11} - u), \ldots, K_{h_s}(U_{s1T_s} - u), \ldots, K_{h_s}(U_{snT_s} - u))$$

and $\mathbf{Y}_s = (Y_{s11}, \ldots, Y_{s1T_s}, \ldots, Y_{snT_s})^\tau$. Hence, the estimator of $(\alpha_{s1}(u), \ldots, \alpha_{sp_s}(u), \alpha'_{s1}(u), \ldots, \alpha'_{sp_s}(u))^\tau$ has the same form as (2.2). In particular, the estimator of $\alpha_{sj}(u)$ is $\widehat{\alpha}_{sj}(u) = \mathbf{e}_{j,2p_s}^\tau(\mathbf{D}_{su}^\tau \mathbf{W}_{su}\mathbf{D}_{su})^{-1}\mathbf{D}_{su}^\tau \mathbf{W}_{su}\mathbf{Y}_s$, where $\mathbf{e}_{j,2p_s}$ is a $2p_s$-vector containing zeros everywhere except for the jth element which equals 1.

The following assumptions are required for the derivation of asymptotic properties of $\widehat{\alpha}_{sj}(u)$:

Assumption 1 For a given s, U_{sit}'s are generated from a distribution with bounded support on $[0, 1]$, and a Lipschitz continuous density function $p_{st}(\cdot)$ that satisfies $0 < \inf_{[0,1]} p_{st}(\cdot) \leq \sup_{[0,1]} p_{st}(\cdot) < \infty$.

Assumption 2 For a given s, $(\mathbf{X}_{si1}^{\tau}, \ldots, \mathbf{X}_{siT_s}^{\tau}, U_{si1}, \ldots, U_{siT_s})$'s and $\boldsymbol{\varepsilon}_{si\cdot} = (\varepsilon_{si1}, \ldots, \varepsilon_{siT_s})^{\tau}$'s are independently and identically distributed for $i = 1, \ldots, n$. The mean and variance of $\boldsymbol{\varepsilon}_{si\cdot}$ are 0 and Σ_{ss} respectively. In addition, $E(\boldsymbol{\varepsilon}_{s_1 i\cdot}\boldsymbol{\varepsilon}_{s_2 i\cdot}^{\tau}) = \Sigma_{s_1 s_2}$ for $1 \le s_1 \ne s_2 \le m$, $E(\varepsilon_{\mathrm{sit}}^4) \le \infty$ and $\sum_{s=1}^{m}\sum_{t=1}^{T_s} ||\mathbf{X}_{\mathrm{sit}}||^4 \le c < \infty$ for $s = 1, \ldots, m$ and $t = 1, \ldots, T_s$.

Assumption 3 $\alpha_{sj}(\cdot)$ has continuous second derivatives on $[0, 1]$ for $s = 1, \ldots, m$ and $j = 1, \ldots, p_s$.

Assumption 4 $K(\cdot)$ is a density function with a bounded support on $[-1, 1]$.

Assumption 5 The bandwidth h_s satisfies $nh_s^8/(\log\log n)^{1/2} \to 0$ and $nh_s^2/(\log n)^2 \to \infty$ as $n \to \infty$.

To facilitate notations, we denote

$$\varsigma_j = \int_{-\infty}^{\infty} u^j K(u)du, \quad \varrho_j = \int_{-\infty}^{\infty} u^j K^2(u)du, \quad \Gamma_{st}(u) = E(\mathbf{X}_{\mathrm{sit}}\mathbf{X}_{\mathrm{sit}}^{\tau}|U_{\mathrm{sit}} = u),$$
$$\boldsymbol{\alpha}_s(\cdot) = (\alpha_{s1}(\cdot), \ldots, \alpha_{sp_s}(\cdot))^{\tau}, \quad \boldsymbol{\alpha}_s'(\cdot) = (\alpha_{s1}'(\cdot), \ldots, \alpha_{sp_s}'(\cdot))^{\tau},$$
$$\widehat{\boldsymbol{\alpha}}_s(\cdot) = (\widehat{\alpha}_{s1}(\cdot), \ldots, \widehat{\alpha}_{sp_s}(\cdot))^{\tau}, \quad \widehat{\boldsymbol{\alpha}}_s'(\cdot) = (\widehat{\alpha}_{s1}'(\cdot), \ldots, \widehat{\alpha}_{sp_s}'(\cdot))^{\tau},$$

and $\mathbf{H}_s = \mathrm{diag}(1, h_s) \otimes \mathbf{I}_{p_s}$, where $\mathbf{I}_{p_s}$ is a $p_s \times p_s$ identity matrix.

The following theorem demonstrates the asymptotic normality of $(\widehat{\boldsymbol{\alpha}}_s^{\tau}(\cdot), \widehat{\boldsymbol{\alpha}}_s'^{\tau}(\cdot))^{\tau}$.

Theorem 1 *Suppose that Assumptions 1 through 5 hold. Then*

$$\sqrt{nh_s}\left[\mathbf{H_s}^{-1}\left\{\begin{pmatrix}\widehat{\boldsymbol{\alpha}}_s(u)\\ \widehat{\boldsymbol{\alpha}}_s'(u)\end{pmatrix} - \begin{pmatrix}\boldsymbol{\alpha}_s(u)\\ \boldsymbol{\alpha}_s'(u)\end{pmatrix}\right\} - \frac{h_s^2}{2}\begin{pmatrix}\Im_1\boldsymbol{\alpha}_s''(u)\\ \Im_2\boldsymbol{\alpha}_s''(u)\end{pmatrix} + o(h_s^2)\right]$$
$$\xrightarrow{D} N(0, \Sigma_{(\boldsymbol{\alpha}_s, \boldsymbol{\alpha}_s')})$$

as $n \to \infty$, *where* $\boldsymbol{\alpha}_s''(u) = (\alpha_{s1}''(u), \ldots, \alpha_{sp_s}''(u))^{\tau}$ *with* $\alpha_{sj}''(u) = \partial^2\alpha_{sj}(u)/\partial u^2$,

$$\Sigma_{(\boldsymbol{\alpha}_s, \boldsymbol{\alpha}_s')} = \left(\sum_{t=1}^{T}\Gamma_{st}(u)p_{st}(u)\right)^{-1}\sum_{t=1}^{T}\sigma_{stst}^2\Gamma_{st}(u)p_{st}(u)$$
$$\times\left(\sum_{t=1}^{T}\Gamma_{st}(u)p_{st}(u)\right)^{-1} \otimes \begin{pmatrix}\Im_{11} & \Im_{12}\\ \Im_{21} & \Im_{22}\end{pmatrix},$$
$$\Sigma_{ss} = (\sigma_{st_1st_2}^2)_{t_1=1,t_2=1}^{T_s}, \quad \Im_1 = \frac{\varsigma_2 - \varsigma_1\varsigma_3}{\varsigma_2 - \varsigma_1^2}, \quad \Im_2 = \frac{\varsigma_3 - \varsigma_1\varsigma_2}{\varsigma_2 - \varsigma_1^2},$$
$$\Im_{11} = \varsigma_2^2\varrho_0 - 2\varsigma_1\varsigma_2\varrho_1 + \varsigma_1^2\varrho_2, \Im_{12} = (\varsigma_1^2 + \varsigma_2)\varrho_1 - \varsigma_1\varsigma_2\varrho_0 - \varsigma_1\varrho_2,$$
$$\Im_{21} = (\varsigma_1^2 + \varsigma_2)\varrho_1 - \varsigma_1\varsigma_2\varrho_0 - \varsigma_1\varrho_2, \text{ and } \Im_{22} = \varrho_2 - \varsigma_1(2\varrho_1 + \varsigma_1\varrho_0).$$

Proof See the online supplementary file.

If no contemporaneous correlations exist across the responses, then by arguments used in Fan and Gijbels (1996), we can show that $\widehat{\alpha}_{sj}(\cdot)$ is asymptotically efficient. In the present case, however, $\widehat{\alpha}_{sj}(\cdot)$ cannot be efficient because it neglects the contemporaneous correlations across responses. Nevertheless, Theorem 1 shows that it is a consistent estimator even though it is inefficient. We will use the regression residuals obtained through the application of $\widehat{\alpha}_{sj}(\cdot)$ to estimate the error covariance matrix and contemporaneous correlations across responses. The residuals are

$$\begin{aligned}\widehat{\varepsilon}_{\text{sit}} &= Y_{\text{sit}} - X_{\text{sit}1}\widehat{\alpha}_{s1}(U_{\text{sit}}) - \cdots - X_{\text{sit}p_s}\widehat{\alpha}_{sp_s}(U_{\text{sit}}),\\ &s = 1, \ldots, m,\ i = 1, \ldots, n,\ t = 1, \ldots, T_s.\end{aligned}$$

Hence, the error covariance matrix may be estimated by $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{\varepsilon}}_{\cdot i\cdot}\widehat{\boldsymbol{\varepsilon}}_{\cdot i\cdot}^{\tau}$ with $\widehat{\boldsymbol{\varepsilon}}_{\cdot i\cdot} = (\widehat{\varepsilon}_{1i1}, \ldots, \widehat{\varepsilon}_{1iT_1}, \ldots, \widehat{\varepsilon}_{miT_m})^{\tau}$. The following theorem gives the asymptotic property of $\widehat{\Sigma}$.

Theorem 2 *Suppose that Assumptions 1 through 5 hold. Then*

$$\sqrt{n}\left(Vech(\widehat{\Sigma}) - Vech(\Sigma)\right) \xrightarrow{D} N\left(0, \mathbf{L}_{\sum_{s=1}^{m} T_s} Cov(\boldsymbol{\varepsilon}_{\cdot 1\cdot} \otimes \boldsymbol{\varepsilon}_{\cdot 1\cdot})\mathbf{L}_{\sum_{s=1}^{m} T_s}^{\tau}\right), \quad as\ n \to \infty,$$

where Vech is a column stacking operator that stacks only the elements on or below the main diagonal of the matrix, $\mathbf{L}_{\sum_{s=1}^{m} T_s}$ *is the* $\frac{1}{2}\sum_{s=1}^{m} T_s(\sum_{s=1}^{m} T_s + 1) \times (\sum_{s=1}^{m} T_s)^2$ *elimination matrix, and* $\boldsymbol{\varepsilon}_{\cdot i\cdot} = (\varepsilon_{1i1}, \ldots, \varepsilon_{1iT_1}, \ldots, \varepsilon_{miT_m})^{\tau}$.

Proof See the online supplementary file.

This estimated error covariance matrix allows the detection of correlations within and across responses, as well as the construction of improved estimators of $\alpha_{sj}(\cdot)$ that properly account for these correlations. The subsequent sections explore these properties in detail.

3 Detection of correlations within and across responses

3.1 Detection of correlations within response

The null hypothesis of interest is that the elements in $\boldsymbol{\varepsilon}_{si\cdot} = (\varepsilon_{si1}, \ldots, \varepsilon_{siT_s})^{\tau}$ are independent, or equivalently, $\Sigma_{ss} = E(\boldsymbol{\varepsilon}_{si\cdot}\boldsymbol{\varepsilon}_{si\cdot}^{\tau})$ is a diagonal matrix. That is

$$H_{0s}: \quad \Sigma_{ss} = \text{diag}\left(\sigma_{s1s1}^2, \sigma_{s2s2}^2, \ldots, \sigma_{sT_ssT_s}^2\right).$$

We use a testing method that adopts the idea of Tsay (2004). It is based on the following observations. First, if the elements in $\boldsymbol{\varepsilon}_{si\cdot}$ are mutually independent, then $Z_{\text{sit}_1t_2} = \varepsilon_{\text{sit}_1}\varepsilon_{\text{sit}_2}$ are uncorrelated with $Z_{\text{sit}_3t_4} = \varepsilon_{\text{sit}_3}\varepsilon_{\text{sit}_4}$, as long as $t_1 \neq t_3$ or $t_2 \neq t_4$. Thus, every $T_s(T_s - 1)/2$ combination of $\varepsilon_{\text{sit}_1}\varepsilon_{\text{sit}_2}$ $(t_1 \neq t_2)$ in $\boldsymbol{\varepsilon}_{si\cdot} = (\varepsilon_{si1}, \ldots, \varepsilon_{siT_s})^{\tau}$ will also be uncorrelated with each other. Second, if the elements in $\boldsymbol{\varepsilon}_{si\cdot}$ are mutually

independent, then as $n \to \infty$, $n^{-\frac{1}{2}} \sum_{i=1}^{n} \varepsilon_{\text{sit}_1} \varepsilon_{\text{sit}_2} \xrightarrow{\mathcal{D}} N(0, \sigma^2_{st_1st_1}\sigma^2_{st_2st_2})$, where $\sigma^2_{st_1st_2}$ is the (t_1, t_2)th element of Σ_{ss}. Then, it holds that

$$\Lambda_s = n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathbf{Z}_{si} \xrightarrow{D} N(0, \Gamma_s) \quad \text{as } n \to \infty,$$

where $\mathbf{Z}_{si} = (Z_{si12}, \ldots, Z_{si1T_s}, Z_{si23}, \ldots, Z_{si2T_s}, \ldots, Z_{si(T_s-1)T_s})^\tau$, $Z_{\text{sit}_1t_2} = \varepsilon_{\text{sit}_1} \varepsilon_{\text{sit}_2}$, and

$$\Gamma_s = \text{diag}\left(\Gamma_{s12}, \Gamma_{s13}, \cdots, \Gamma_{s(T_s-1)T_s}\right) \quad \text{with } \Gamma_{st_1t_2} = \sigma^2_{st_1st_1}\sigma^2_{st_2st_2}.$$

In practice, $\varepsilon_{\text{sit}_1}\varepsilon_{\text{sit}_2}$ is unobservable and may be replaced by $\widehat{\varepsilon}_{\text{sit}_1}\widehat{\varepsilon}_{\text{sit}_2}$, where $\widehat{\varepsilon}_{\text{sit}} = Y_{\text{sit}} - X_{\text{sit}1}\widehat{\alpha}_{s1}(U_{\text{sit}}) - \cdots - X_{\text{sit}p_s}\widehat{\alpha}_{sp_s}(U_{\text{sit}})$. The test of H_{0s} is based on the test statistic

$$\widehat{M}_s = \widehat{\Lambda}_s^\tau \widehat{\Gamma}_s^{-1} \widehat{\Lambda}_s,$$

where $\widehat{\mathbf{Z}}_{si} = (\widehat{Z}_{si12}, \widehat{Z}_{si13}, \ldots, \widehat{Z}_{si(T_s-1)T_s})^\tau$, $\widehat{Z}_{\text{sit}_1t_2} = \widehat{\varepsilon}_{\text{sit}_1}\widehat{\varepsilon}_{\text{sit}_2}$, and $\widehat{\Gamma}_s$ has the same definition as Γ_s except that $\Gamma_{st_1t_2}$ is replaced by $\widehat{\Gamma}_{st_1t_2} = \widehat{\sigma}^2_{st_1st_1}\widehat{\sigma}^2_{st_2st_2}$. In addition, $\widehat{\Lambda}_s = n^{-\frac{1}{2}} \sum_{i=1}^{n} \widehat{\mathbf{Z}}_{si}$.

The following theorem gives the asymptotic distribution of $\widehat{M}_s$ under H_{0s}.

Theorem 3 *Suppose that Assumptions 1 through 5 hold. Then under H_{0s}, as $n \to \infty$, $\widehat{M}_s \xrightarrow{D} \chi^2_{T_s(T_s-1)/2}$.*

Proof See the online supplementary file.

3.2 Detection of correlations among responses

To detect correlations among Y_{s_1it} and Y_{s_2it} for $1 \le s_1 \ne s_2 \le m$, the null hypothesis of interest is

$$H_{0s_1s_2}: \quad \Sigma_{s_1s_2} = E(\boldsymbol{\varepsilon}_{s_1i\cdot}\boldsymbol{\varepsilon}^\tau_{s_2i\cdot}) = \mathbf{0}_{T_{s_1}\times T_{s_2}}.$$

We define the following test statistic based on arguments along the lines of those presented in Sect. 3.1:

$$\widehat{M}_{s_1s_2} = \widehat{\Lambda}^\tau_{s_1s_2} \widehat{\Gamma}^{-1}_{s_1s_2} \widehat{\Lambda}_{s_1s_2},$$

where $\widehat{\Lambda}_{s_1s_2} = n^{-\frac{1}{2}} \sum_{i=1}^{n} \widehat{\mathbf{Z}}_{s_1s_2i}$ with $\widehat{\mathbf{Z}}_{s_1s_2i} = (\widehat{Z}_{s_1s_2i11}, \ldots, \widehat{Z}_{s_1s_2i1T_{s_2}}, \widehat{Z}_{s_1s_2i21}, \ldots, \widehat{Z}_{s_1s_2i2T_{s_2}}, \ldots, \widehat{Z}_{s_1s_2iT_{s_1}T_{s_2}})^\tau$, $\widehat{Z}_{s_1s_2it_1t_2} = \widehat{\varepsilon}_{s_1it_1}\widehat{\varepsilon}_{s_2it_2}$, and

$$\widehat{\Gamma}_{s_1s_2} = \text{diag}\left(\widehat{\Gamma}_{s_1s_211}, \widehat{\Gamma}_{s_1s_212}, \cdots, \widehat{\Gamma}_{s_1s_2T_{s_1}T_{s_2}}\right), \quad \text{with } \widehat{\Gamma}_{s_1s_2t_1t_2} = \widehat{\sigma}^2_{s_1t_1s_1t_1}\widehat{\sigma}^2_{s_2t_2s_2t_2}.$$

The following theorem gives the asymptotic distribution of $\widehat{M}_{s_1s_2}$.

Theorem 4 *Suppose that Assumptions 1 through 5 hold. Then under* $H_{0s_1s_2}$, *as* $n \to \infty$, $\widehat{M}_{s_1s_2} \xrightarrow{D} \chi^2_{T_{s_1}T_{s_2}}$.

Proof See the online supplementary file.

3.3 Detection of correlations within and among responses

Under the null hypothesis of no correlation within or among responses, $\Sigma = E(\boldsymbol{\varepsilon}_{\cdot i}.\boldsymbol{\varepsilon}_{\cdot i}^{\tau}.)$ is a diagonal matrix. That is

$$H_0: \quad \Sigma = \left(\sigma^2_{1111}, \sigma^2_{1212}, \ldots, \sigma^2_{1T_11T_1}, \ldots, \sigma^2_{mT_mmT_m}\right).$$

Along the lines of arguments used in Sect. 3.1, H_0 may be tested using the following statistic:

$$\widehat{M} = \widehat{\Lambda}^{\tau}\widehat{\Gamma}^{-1}\widehat{\Lambda},$$

where $\widehat{\Lambda} = n^{-\frac{1}{2}}\sum_{i=1}^{n}\widehat{\mathbf{Z}}_i$ with $\widehat{\mathbf{Z}}_i = (\widehat{Z}_{11i12}, \ldots, \widehat{Z}_{11i1T_1}, \widehat{Z}_{11i23}, \ldots, \widehat{Z}_{11i2T_1}, \ldots, \widehat{Z}_{11i(T_1-1)T_1}, \widehat{Z}_{12i11}, \ldots, \widehat{Z}_{12i1T_2}, \ldots, \widehat{Z}_{12iT_1T_2}, \ldots, \widehat{Z}_{\text{mmi}12}, \ldots, \widehat{Z}_{\text{mmi}1T_1}, \widehat{Z}_{\text{mmi}23}, \ldots, \widehat{Z}_{\text{mmi}2T_1}, \ldots, \widehat{Z}_{\text{mmi}(T_m-1)T_m})^{\tau}$, $\widehat{Z}_{s_1s_2it_1t_2} = \widehat{\varepsilon}_{s_1it_1}\widehat{\varepsilon}_{s_2it_2}$, and $\widehat{\Gamma} = (\widehat{\Gamma}_{s_1s_2})_{s_1,s_2=1}^{m}$, with

$$\widehat{\Gamma}_{ss} = \text{diag}\left(\widehat{\Gamma}_{ss12}, \widehat{\Gamma}_{ss13}, \cdots, \widehat{\Gamma}_{ss(T_s-1)T_s}\right),$$
$$\widehat{\Gamma}_{s_1s_2} = \left(\widehat{\Gamma}_{s_1s_211}, \widehat{\Gamma}_{s_1s_212}, \cdots, \widehat{\Gamma}_{s_1s_2T_{s_1}T_{s_2}}\right)$$

and $\widehat{\Gamma}_{s_1s_2t_1t_2} = \widehat{\sigma}^2_{s_1t_1s_1t_1}\widehat{\sigma}^2_{s_2t_2s_2t_2}$.

The following theorem provides the asymptotic distribution of $\widehat{M}$.

Theorem 5 *Suppose that Assumptions 1 through 5 hold. Then under* H_0, *as* $n \to \infty$, $\widehat{M} \xrightarrow{D} \chi^2_{\sum_{s=1}^{m} T_s(\sum_{s=1}^{m} T_s-1)/2}$.

Proof See the online supplementary file.

The statistics $\widehat{M}_s$, $\widehat{M}_{s_1s_2}$ and $\widehat{M}$ for testing H_{0s}, $H_{0s_1s_2}$ and H_0 have some distinct appeal. The fact that they are all χ^2 distributed under their respective null with degrees of freedom independent of sample size means that they are easy to apply. Also, the burden associated with the computation of $\widehat{M}_s$, $\widehat{M}_{s_1s_2}$ and $\widehat{M}$ increases only linearly with the magnitude of T_s. Moreover, these tests are flexible in the sense that they permit testing of block-diagonality in addition to diagonality in the error covariance matrix.

4 A two-stage procedure for estimating unknown coefficient functions

The difficulties associated with applying nonparametric regression to a three-level clustered data model stem largely from the fact that kernel smoothing is a 'local' technique, whereas correlations within and across responses are a 'global' issue. For the panel with a single response variable, Wang (2003) proposed a two-stage seemingly unrelated kernel regression method that accounts for correlations within the response. When correlations both within and across responses are present and represented by a one-way error component structure, Zhou et al. (2011) proposed a two-stage local polynomial estimation procedure that accounts for these correlations. However, their method relies heavily on the assumption of a one-way error component structure, which fails to represent common situations (e.g., in economics) where an unobserved shock affects the behavioral relationship for a prolonged period.

To describe our method, denote $\Sigma_{s(-s)} = \left(\Sigma_{s1}, \cdots, \Sigma_{s,s-1}, \Sigma_{s,s+1}, \cdots, \Sigma_{s,m}\right)$, $\Sigma_{(-s)(-s)} = (\Sigma_{s_1 s_2})_{s_1,s_2\neq s}^{m}$ and $\boldsymbol{\varepsilon}_{si\cdot} = (\varepsilon_{si1}, \ldots, \varepsilon_{siT_s})^{\tau}$. Note that

$$\begin{aligned}
&\Sigma_{s(-s)}\Sigma_{(-s)(-s)}^{-1} \\
&= \operatorname{argmin}_{\Delta\in\Re^{T_s\times\sum_{s_1\neq s}^{m}T_{s_1}}} E\left\{\boldsymbol{\varepsilon}_{si\cdot} - \Delta\begin{pmatrix}\boldsymbol{\varepsilon}_{1i\cdot}\\ \vdots\\ \boldsymbol{\varepsilon}_{s-1,i\cdot}\\ \boldsymbol{\varepsilon}_{s+1,i\cdot}\\ \vdots\\ \boldsymbol{\varepsilon}_{mi\cdot}\end{pmatrix}\right\}\left\{\boldsymbol{\varepsilon}_{si\cdot} - \Delta\begin{pmatrix}\boldsymbol{\varepsilon}_{1i\cdot}\\ \vdots\\ \boldsymbol{\varepsilon}_{s-1,i\cdot}\\ \boldsymbol{\varepsilon}_{s+1,i\cdot}\\ \vdots\\ \boldsymbol{\varepsilon}_{mi\cdot}\end{pmatrix}\right\}^{\tau}.
\end{aligned}$$

We first describe the method of estimating $\boldsymbol{\alpha}_s(\cdot) = (\alpha_{s1}(\cdot), \ldots, \alpha_{sp_s})^{\tau}$, $\boldsymbol{\varepsilon}_{s_1 i\cdot}$ with $s_1 \neq s$ under the assumption of known $\Sigma_{s(-s)}$ and $\Sigma_{(-s)(-s)}$. This assumption will be relaxed at a later stage.

Consider the pseudo responses

$$\begin{aligned}
\mathbf{Y}_{si\cdot}^{*} &= (\mathbf{Y}_{si1}^{*}, \ldots, \mathbf{Y}_{siT_s}^{*})^{\tau} \\
&= \mathbf{Y}_{si\cdot} - \Sigma_{s(-s)}\Sigma_{(-s)(-s)}^{-1}\left(\boldsymbol{\varepsilon}_{1i\cdot}^{\tau}, \ldots, \boldsymbol{\varepsilon}_{s-1,i\cdot}^{\tau}, \boldsymbol{\varepsilon}_{s+1,i\cdot}^{\tau}, \ldots, \boldsymbol{\varepsilon}_{mi\cdot}^{\tau}\right)^{\tau}.
\end{aligned}$$

Conditional on $\mathbf{X}_{\mathrm{sit}}$ and U_{sit},

$$E(\mathbf{Y}_{\mathrm{sit}}^{*}) = X_{\mathrm{sit}1}\alpha_{s1}(U_{\mathrm{sit}}) + \cdots + X_{\mathrm{sit}p_s}\alpha_{sp_s}(U_{\mathrm{sit}})$$

and

$$\mathrm{Cov}(\mathbf{Y}_{si\cdot}^{*}) = \Sigma_{ss} - \Sigma_{s(-s)}\Sigma_{(-s)(-s)}^{-1}\Sigma_{s(-s)}^{\tau} \leq \Sigma_{ss}.$$

The latter implies that the error variance of $\mathbf{Y}_{si\cdot}$ can be reduced by taking into account the correlations among the responses. Additionally, let $\Delta_{ss} = (\delta_{st_1st_2})_{t_1,t_2=1}^{T_s} = \Sigma_{ss} - \Sigma_{s(-s)}\Sigma_{(-s)(-s)}^{-1}\Sigma_{s(-s)}^{\tau}$ and $\Delta_{ss}^{-1} = (\delta_{ss}^{t_1t_2})_{t_1,t_2=1}^{T_s}$. The tth element of $\Delta_{ss}^{-1}\mathbf{Y}_{si\cdot}^{*}$ is

$$Y_{\mathrm{sit}}^{**} = \sum_{t_1=1}^{T_s} \delta_{ss}^{tt_1} Y_{\mathrm{sit}_1}^{*} = \delta_{ss}^{tt} Y_{\mathrm{sit}}^{*} + \sum_{t_1 \neq t}^{T_s} \delta_{ss}^{tt_1} Y_{\mathrm{sit}_1}^{*}.$$

Denote $a_{\mathrm{sit}_1} = \alpha_{s1}(U_{\mathrm{sit}_1}) X_{\mathrm{sit}_1 1} + \cdots + \alpha_{sp_s}(U_{\mathrm{sit}_1}) X_{\mathrm{sit}_1 p_s}$ and assume that a_{sit_1}'s are known. We then have

$$\begin{aligned} Y_{\mathrm{sit}}^{***} &= (\delta_{ss}^{tt})^{-1} \left(Y_{\mathrm{sit}}^{**} - \sum_{t_1 \neq t}^{T_s} \delta_{ss}^{tt_1} a_{\mathrm{sit}_1} \right) \\ &= \alpha_{s1}(U_{\mathrm{sit}}) X_{\mathrm{sit}1} + \cdots + \alpha_{sp_s}(U_{\mathrm{sit}}) X_{\mathrm{sit}p_s} + (\delta_{ss}^{tt})^{-1} \sum_{t_1=1}^{T_s} \delta_{ss}^{tt_1} \varepsilon_{\mathrm{sit}_1}. \end{aligned}$$

Conditional on $\mathbf{X}_{\mathrm{sit}}$ and U_{sit}, Y_{sit}^{***} has a mean of $\alpha_{s1}(U_{\mathrm{sit}}) X_{\mathrm{sit}1} + \cdots + \alpha_{sp_s}(U_{\mathrm{sit}}) X_{\mathrm{sitp}_s}$ and variance $(\delta_{ss}^{tt})^{-1}$, which is smaller than σ_{stst}^2. So, applying a local polynomial estimation to Y_{sit}^{***} can result in a more efficient estimator of the unknown coefficient functions in model (1.1). Now, an estimator of $(\alpha_{s1}(u), \ldots, \alpha_{sp_s}(u), \alpha'_{s1}(u), \ldots, \alpha'_{sp_s}(u))^{\tau}$ has the form

$$(\tilde{\alpha}_{s1}^{TS}(u), \ldots, \tilde{\alpha}_{sp_s}^{TS}(u), \tilde{\alpha}_{s1}^{'TS}(u), \ldots, \tilde{\alpha}_{sp_s}^{'TS}(u))^{\tau} = (\mathbf{D}_{su}^{*\tau} \mathbf{W}_{su}^{*} \mathbf{D}_{su}^{*})^{-1} \mathbf{D}_{su}^{*\tau} \mathbf{W}_{su}^{*} \mathbf{Y}_{s}^{***},$$

where $\mathbf{D}_{su}^{*}$ and $\mathbf{W}_{su}^{*}$ are the same as $\mathbf{D}_{su}$ and $\mathbf{W}_{su}$, respectively, except that h_s is replaced by h_s^{*}, and $\mathbf{Y}_{s}^{***} = (Y_{s11}^{***}, \ldots, Y_{s1T_s}^{***}, \ldots, Y_{snT_s}^{***})$.

Clearly, in practice, $\boldsymbol{\varepsilon}_{si\cdot}$, $\delta_{ss}^{t_1 t_2}$ and a_{sit_1} are unknown. To make the estimator operational, we replace these unknown quantities by $\widehat{\boldsymbol{\varepsilon}}_{si\cdot} = (\widehat{\varepsilon}_{si1}, \ldots, \widehat{\varepsilon}_{siT_s})^{\tau}$, $\widehat{\delta}_{ss}^{t_1 t_2}$ and $\widehat{a}_{\mathrm{sit}}$ respectively. Thus, a feasible two-stage local polynomial estimator of $(\alpha_{s1}(u), \ldots, \alpha_{sp_s}(u), \alpha'_{s1}(u), \ldots, \alpha'_{sp_s}(u))^{\tau}$ is

$$(\widehat{\alpha}_{s1}^{TS}(u), \ldots, \widehat{\alpha}_{sp_s}^{TS}(u), \widehat{\alpha}_{s1}^{'TS}(u), \ldots, \widehat{\alpha}_{sp_s}^{'TS}(u))^{\tau} = (\mathbf{D}_{su}^{*\tau} \mathbf{W}_{su}^{*} \mathbf{D}_{su}^{*})^{-1} \mathbf{D}_{su}^{*\tau} \mathbf{W}_{su}^{*} \widehat{\mathbf{Y}}_{s}^{***}$$

where $\widehat{\mathbf{Y}}_{s}^{***} = (\widehat{Y}_{s11}^{***}, \ldots, \widehat{Y}_{s1T_s}^{***}, \ldots, \widehat{Y}_{snT_s}^{***})^{\tau}$,

$$\widehat{Y}_{\mathrm{sit}}^{***} = (\widehat{\delta}_{ss}^{tt})^{-1} \left(\widehat{Y}_{\mathrm{sit}}^{**} - \sum_{t_1 \neq t}^{T_s} \widehat{\delta}_{ss}^{tt_1} \widehat{a}_{\mathrm{sit}_1} \right), \quad i = 1, \ldots, n, \ t = 1, \ldots, T_s$$

with $\widehat{\Delta}_{ss}^{-1} = (\widehat{\Sigma}_{ss} - \widehat{\Sigma}_{s(-s)} \widehat{\Sigma}_{(-s)(-s)}^{-1} \widehat{\Sigma}_{s(-s)}^{\tau})^{-1} = (\widehat{\delta}_{ss}^{t_1 t_2})_{t_1, t_2 = 1}^{T}$, $\widehat{\Sigma}_{s(-s)} = (\widehat{\Sigma}_{s1}, \ldots, \widehat{\Sigma}_{s,s-1}, \widehat{\Sigma}_{s,s+1}, \ldots, \widehat{\Sigma}_{sm})$, $\widehat{\Sigma}_{(-s)(-s)} = (\widehat{\Sigma}_{s_1 s_2})_{s_1, s_2 \neq s}^{m}$, $\widehat{Y}_{\mathrm{sit}}^{**} = \widehat{\delta}_{ss}^{tt} \widehat{Y}_{\mathrm{sit}}^{*} + \sum_{t_1 \neq t}^{T_s} \widehat{\delta}_{ss}^{tt_1} \widehat{Y}_{\mathrm{sit}_1}^{*}$,

$$\begin{aligned} \widehat{\mathbf{Y}}_{si\cdot}^{*} &= (\widehat{\mathbf{Y}}_{si1}^{*}, \ldots, \widehat{\mathbf{Y}}_{siT_s}^{*})^{\tau} \\ &= \mathbf{Y}_{si\cdot} - \widehat{\Sigma}_{s(-s)} \widehat{\Sigma}_{(-s)(-s)}^{-1} \left(\widehat{\boldsymbol{\varepsilon}}_{1i\cdot}^{\tau}, \ldots, \widehat{\boldsymbol{\varepsilon}}_{s-1,i\cdot}^{\tau}, \widehat{\boldsymbol{\varepsilon}}_{s+1,i\cdot}^{\tau}, \ldots, \widehat{\boldsymbol{\varepsilon}}_{mi\cdot}^{\tau} \right)^{\tau}, \\ \widehat{\boldsymbol{\varepsilon}}_{si\cdot} &= (\widehat{\varepsilon}_{si1}, \ldots, \widehat{\varepsilon}_{siT_s})^{\tau}, \quad \widehat{\varepsilon}_{\mathrm{sit}} = Y_{\mathrm{sit}} - \widehat{\alpha}_{s1}(U_{\mathrm{sit}}) X_{\mathrm{sit}1} - \cdots - \widehat{\alpha}_{sp_s}(U_{\mathrm{sit}}) X_{\mathrm{sitp}_s} \end{aligned}$$

and

$$\widehat{a}_{\mathrm{sit}_1} = \widehat{\alpha}_{s1}(U_{\mathrm{sit}_1}) X_{\mathrm{sit}_1 1} + \cdots + \widehat{\alpha}_{sp_s}(U_{\mathrm{sit}_1}) X_{\mathrm{sit}_1 p_s}.$$

In particular, a two-stage local polynomial estimator of $\alpha_{sj}(u)$ is given by

$$\widehat{\alpha}_{sj}^{TS}(u) = \mathbf{e}_{j,2p_s}^{\tau}(\mathbf{D}_{su}^{*\tau}\mathbf{W}_{su}^{*}\mathbf{D}_{su}^{*})^{-1}\mathbf{D}_{su}^{*\tau}\mathbf{W}_{su}^{*}\widehat{\mathbf{Y}}_{s}^{***},$$

where $\mathbf{e}_{j,2p_s}$ is a $2p_s$-vector with zero everywhere except for the jth element which equals 1.

5 Asymptotic properties of proposed estimators

The following assumption, in addition to Assumptions 1 through 5, is required for the investigation of the asymptotic properties of $(\widehat{\boldsymbol{\alpha}}_{s}^{TS\tau}(\cdot), \widehat{\boldsymbol{\alpha}}_{s}^{'TS\tau}(\cdot))^{\tau}$.

Assumption 6 The bandwidth h_s^* satisfies $nh_s^{*8}/(\log\log n)^{1/2} \to 0$ and $nh_s^{*2}/(\log n)^2 \to \infty$ as $n \to \infty$. In addition, $\max_{1\le s\le m} h_s = o(h_s^*)$.

The core asymptotic properties of $(\widehat{\boldsymbol{\alpha}}_{s}^{TS\tau}(\cdot), \widehat{\boldsymbol{\alpha}}_{s}^{'TS\tau}(\cdot))^{\tau}$ are summarized in the following theorem.

Theorem 6 *Suppose that Assumptions 1 through 6 hold. Then*

$$\sqrt{nh_s^*}\left[\mathbf{H}_s^{*-1}\left\{\begin{pmatrix}\widehat{\boldsymbol{\alpha}}_{s}^{TS}(u)\\ \widehat{\boldsymbol{\alpha}}_{s}^{'TS}(u)\end{pmatrix}-\begin{pmatrix}\boldsymbol{\alpha}_s(u)\\ \boldsymbol{\alpha}_s'(u)\end{pmatrix}\right\}-\frac{h_s^{*2}}{2}\begin{pmatrix}\Im_1\boldsymbol{\alpha}_s''(u)\\ \Im_2\boldsymbol{\alpha}_s''(u)\end{pmatrix}+o(h_s^{*2})\right]$$
$$\xrightarrow{D} N(0, \Sigma_{(\boldsymbol{\alpha}_s,\boldsymbol{\alpha}_s')}^{TS})$$

as $n \to \infty$, *where* $\mathbf{H}_s^* = diag(1, h_s^*) \otimes \mathbf{I}_{p_s}$, $\boldsymbol{\alpha}_s''(u) = (\alpha_{s1}''(u), \ldots, \alpha_{sp_s}''(u))^{\tau}$ *with* $\alpha_{sj}''(u) = \partial^2\alpha_{sj}(u)/\partial u^2$,

$$\Sigma_{(\boldsymbol{\alpha}_s,\boldsymbol{\alpha}_s')}^{TS} = \left(\sum_{t=1}^{T}\Gamma_{st}(u)p_{st}(u)\right)^{-1}\sum_{t=1}^{T}(\delta_{ss}^{tt})^{-1}\Gamma_{st}(u)p_{st}(u)\left(\sum_{t=1}^{T}\Gamma_{st}(u)p_{st}(u)\right)^{-1}$$
$$\otimes\begin{pmatrix}\Im_{11} & \Im_{12}\\ \Im_{21} & \Im_{22}\end{pmatrix},$$

and other symbols are defined in Theorem 1.

Proof See the online supplementary file.

Remark 1 Recall that $\Delta_{ss} = (\delta_{st_1st_2}^2)_{t_1,t_2=1}^{T_s} = \Sigma_{ss} - \Sigma_{s(-s)}\Sigma_{(-s)(-s)}^{-1}\Sigma_{s(-s)}^{\tau} \le \Sigma_{ss}$ and $\Delta_{ss}^{-1} \ge \Sigma_{ss}^{-1}$. This implies $\delta_{ss}^{tt} \ge \sigma_{ss}^{tt} = (\sigma_{stst}^2 - \Sigma_{sst(-t)}\Sigma_{ss(-t)(-t)}^{-1}\Sigma_{sst(-t)}^{\tau})^{-1} \ge (\sigma_{stst}^2)^{-1}$, where $\Sigma_{sst(-t)} = E(\varepsilon_{si}\text{t}(\varepsilon_{si1}, \ldots, \varepsilon_{si,t-1}, \varepsilon_{si,t+1}, \ldots, \varepsilon_{siT_s})^{\tau})$ and $\Sigma_{ss(-t)(-t)} = E((\varepsilon_{si1}, \ldots, \varepsilon_{si,t-1}, \varepsilon_{si,t+1}, \ldots, \varepsilon_{siT_s})(\varepsilon_{si1}, \ldots, \varepsilon_{si,t-1}, \varepsilon_{si,t+1}, \ldots, \varepsilon_{siT_s})^{\tau})$. So, $\Sigma_{(\boldsymbol{\alpha}_s,\boldsymbol{\alpha}_s')}^{TS} - \Sigma_{(\boldsymbol{\alpha}_s,\boldsymbol{\alpha}_s')}$ is a negative semi-definite matrix, or $(\widehat{\boldsymbol{\alpha}}_{s}^{TS\tau}(\cdot),$

$\widehat{\boldsymbol{\alpha}}_s^{TS^{'}\tau}(\cdot))^\tau$ is variance superior to $(\widehat{\boldsymbol{\alpha}}_s^\tau(\cdot), \widehat{\boldsymbol{\alpha}}_s^{'\tau}(\cdot))^\tau$. Additionally, from Theorems 1 and 6, we can see that $(\widehat{\boldsymbol{\alpha}}_s^{TS\tau}(\cdot), \widehat{\boldsymbol{\alpha}}_s^{TS^{'}\tau}(\cdot))^\tau$ and $(\widehat{\boldsymbol{\alpha}}_s^\tau(\cdot), \widehat{\boldsymbol{\alpha}}_s^{'\tau}(\cdot))^\tau$ have the same asymptotic bias.

Suppose correlations exist only within and not across responses. The two-stage estimator of $(\boldsymbol{\alpha}_s^\tau(\cdot), \boldsymbol{\alpha}_s^{'\tau}(\cdot))^\tau$ then has the form

$$(\check{\alpha}_{s1}^{TS}(u), \ldots, \check{\alpha}_{sp_s}^{TS}(u), \check{\alpha}_{s1}^{'TS}(u), \ldots, \check{\alpha}_{sp_s}^{'TS}(u))^\tau = (\mathbf{D}_{su}^{*\tau}\mathbf{W}_{su}^*\mathbf{D}_{su}^*)^{-1}\mathbf{D}_{su}^{*\tau}\mathbf{W}_{su}^*\check{\mathbf{Y}}_s^{**},$$

where $\check{\mathbf{Y}}_s^{**} = (\check{Y}_{s11}^{**}, \ldots, \check{Y}_{s1T_s}^{**}, \ldots, \check{Y}_{snT_s}^{**})^\tau$,

$$\check{Y}_{\mathrm{sit}}^{**} = (\widehat{\sigma}_{ss}^{tt})^{-1}\left(\check{Y}_{\mathrm{sit}}^{*} - \sum_{t_1\neq t}^{T_s}\widehat{\sigma}_{ss}^{tt_1}\widehat{a}_{\mathrm{sit}_1}\right), \quad i = 1, \ldots, n,\ t = 1, \ldots, T_s$$

with $\widehat{\Sigma}_{ss}^{-1} = (\widehat{\sigma}_{ss}^{t_1t_2})_{t_1,t_2=1}^{T_s}$, $\check{Y}_{\mathrm{sit}}^{*} = \widehat{\sigma}_{ss}^{tt}Y_{\mathrm{sit}} + \sum_{t_1\neq t}^{T}\widehat{\sigma}_{ss}^{tt_1}Y_{\mathrm{sit}_1}$ and

$$\widehat{a}_{\mathrm{sit}_1} = \widehat{\alpha}_{s1}(U_{\mathrm{sit}_1})X_{\mathrm{sit}_1 1} + \cdots + \widehat{\alpha}_{sp_s}(U_{\mathrm{sit}_1})X_{\mathrm{sit}_1 p_s}.$$

Similar to Theorem 6, we have the following asymptotic result for $(\widehat{\boldsymbol{\alpha}}_s^{TS\tau}(\cdot), \widehat{\boldsymbol{\alpha}}_s^{TS^{'}\tau}(\cdot))^\tau$:

$$\begin{aligned}&\sqrt{nh_s^*}\left[\mathbf{H}_s^{*-1}\left\{\begin{pmatrix}\check{\boldsymbol{\alpha}}_s^{TS}(u)\\ \check{\boldsymbol{\alpha}}_s^{'TS}(u)\end{pmatrix} - \begin{pmatrix}\boldsymbol{\alpha}_s(u)\\ \boldsymbol{\alpha}_s^{'}(u)\end{pmatrix}\right\} - \frac{h_s^{*2}}{2}\begin{pmatrix}\Im_1\boldsymbol{\alpha}_s^{''}(u)\\ \Im_2\boldsymbol{\alpha}_s^{''}(u)\end{pmatrix} + o(h_s^{*2})\right]\\ &\quad\xrightarrow{D} N(0, \Sigma_{(\boldsymbol{\alpha}_s,\boldsymbol{\alpha}_s')}^{*TS})\end{aligned}$$

as $n \to \infty$, where

$$\begin{aligned}\overset{*}{\Sigma}_{(\boldsymbol{\alpha}_s,\boldsymbol{\alpha}_s')}^{TS} = &\left(\sum_{t=1}^{T}\Gamma_{st}(u)p_{st}(u)\right)^{-1}\sum_{t=1}^{T}(\sigma_{ss}^{tt})^{-1}\Gamma_{st}(u)p_{st}(u)\left(\sum_{t=1}^{T}\Gamma_{st}(u)p_{st}(u)\right)^{-1}\\ &\otimes\begin{pmatrix}\Im_{11} & \Im_{12}\\ \Im_{21} & \Im_{22}\end{pmatrix}.\end{aligned}$$

As $\delta_{ss}^{tt} \geq \sigma_{ss}^{tt}$, we have $\overset{*}{\Sigma}_{(\boldsymbol{\alpha}_s,\boldsymbol{\alpha}_s')}^{TS} \geq \Sigma_{(\boldsymbol{\alpha}_s,\boldsymbol{\alpha}_s')}^{TS}$, and accordingly $(\widehat{\boldsymbol{\alpha}}_s^{TS\tau}(\cdot), \widehat{\boldsymbol{\alpha}}_s^{TS^{'}\tau}(\cdot))^\tau$ is asymptotically more efficient than $(\check{\boldsymbol{\alpha}}_s^{TS\tau}(\cdot), \check{\boldsymbol{\alpha}}_s^{TS^{'}\tau}(\cdot))^\tau$.

On the other hand, if correlations exist only across and not within responses, then the two-stage estimator of $(\boldsymbol{\alpha}_s^\tau(\cdot), \boldsymbol{\alpha}_s^{'\tau}(\cdot))^\tau$ has the form

$$(\bar{\alpha}_{s1}^{TS}(u), \ldots, \bar{\alpha}_{sp_s}^{TS}(u), \bar{\alpha}_{s1}^{'TS}(u), \ldots, \bar{\alpha}_{sp_s}^{'TS}(u))^\tau = (\mathbf{D}_{su}^{*\tau}\mathbf{W}_{su}^*\mathbf{D}_{su}^*)^{-1}\mathbf{D}_{su}^{*\tau}\mathbf{W}_{su}^*\bar{\mathbf{Y}}_s^*,$$

where $\bar{\mathbf{Y}}_s^* = (\bar{Y}_{s11}^*, \ldots, \bar{Y}_{s1T_s}^*, \ldots, \bar{Y}_{snT_s}^*)^\tau$,

$$\begin{aligned}\bar{\mathbf{Y}}_{si\cdot}^* &= (\bar{\mathbf{Y}}_{si1}^*, \ldots, \bar{\mathbf{Y}}_{siT_s}^*)^\tau \\ &= \mathbf{Y}_{si\cdot} - \widehat{\Sigma}_{s(-s)}\widehat{\Sigma}_{(-s)(-s)}^{-1}\left(\widehat{\boldsymbol{\varepsilon}}_{1i\cdot}^\tau, \ldots, \widehat{\boldsymbol{\varepsilon}}_{s-1,i\cdot}^\tau, \ldots, \widehat{\boldsymbol{\varepsilon}}_{s+1,i\cdot}^\tau, \ldots, \widehat{\boldsymbol{\varepsilon}}_{mi\cdot}^\tau\right)^\tau.\end{aligned}$$

Similar to Theorem 6, we have the following asymptotic result for $(\bar{\boldsymbol{\alpha}}_s^{TS\tau}(\cdot), \bar{\boldsymbol{\alpha}}_s^{'TS\tau}(\cdot))^\tau$:

$$\begin{aligned}&\sqrt{nh_s^*}\left[\mathbf{H}_s^{*-1}\left\{\begin{pmatrix}\bar{\boldsymbol{\alpha}}_s^{TS}(u)\\ \bar{\boldsymbol{\alpha}}_s^{'TS}(u)\end{pmatrix} - \begin{pmatrix}\boldsymbol{\alpha}_s(u)\\ \boldsymbol{\alpha}_s'(u)\end{pmatrix}\right\} - \frac{h_s^{*2}}{2}\begin{pmatrix}\Im_1\boldsymbol{\alpha}_s''(u)\\ \Im_2\boldsymbol{\alpha}_s''(u)\end{pmatrix} + o(h_s^{*2})\right] \\ &\xrightarrow{D} N(0, \Sigma_{(\boldsymbol{\alpha}_s, \boldsymbol{\alpha}_s')}^{**TS})\end{aligned}$$

as $n \to \infty$, where

$$\begin{aligned}\Sigma_{(\boldsymbol{\alpha}_s, \boldsymbol{\alpha}_s')}^{**TS} = &\left(\sum_{t=1}^T \Gamma_{st}(u)p_{st}(u)\right)^{-1}\sum_{t=1}^T \delta_{ststst}^2\Gamma_{st}(u)p_{st}(u)\left(\sum_{t=1}^T \Gamma_{st}(u)p_{st}(u)\right)^{-1} \\ &\otimes\begin{pmatrix}\Im_{11} & \Im_{12}\\ \Im_{21} & \Im_{22}\end{pmatrix}.\end{aligned}$$

Because $\delta_{ss}^{tt} \geq \delta_{ststst}^{-2}$, $(\widehat{\boldsymbol{\alpha}}_s^{TS\tau}(\cdot), \widehat{\boldsymbol{\alpha}}_s^{'TS\tau}(\cdot))^\tau$ is also asymptotically more efficient than $(\bar{\boldsymbol{\alpha}}_s^{TS\tau}(\cdot), \bar{\boldsymbol{\alpha}}_s^{'TS\tau}(\cdot))^\tau$ under the case where correlations exist only across and not within responses.

The following consistent estimator of $\Sigma_{(\boldsymbol{\alpha}_s, \boldsymbol{\alpha}_s')}^{TS}$ is required for the purpose of statistical inference such as the construction of piecewise confidence band of $(\boldsymbol{\alpha}_s^\tau(\cdot), \boldsymbol{\alpha}_s^{'\tau}(\cdot))^\tau$:

$$\begin{aligned}\widehat{\Sigma}_{(\boldsymbol{\alpha}_s, \boldsymbol{\alpha}_s')}^{TS} = &\left(\frac{1}{n}\sum_{i=1}^n\sum_{t=1}^T \mathbf{X}_{\mathrm{sit}}\mathbf{X}_{\mathrm{sit}}^\tau K_{h_s^*}(U_{\mathrm{sit}} - u)\right)^{-1} \\ &\times\frac{1}{n}\sum_{i=1}^n\sum_{t=1}^T (\widehat{\delta}_{ss}^{tt})^{-1}\mathbf{X}_{\mathrm{sit}}\mathbf{X}_{\mathrm{sit}}^\tau K_{h_s^*}(U_{\mathrm{sit}} - u) \\ &\cdot\left(\frac{1}{n}\sum_{i=1}^n\sum_{t=1}^T \mathbf{X}_{\mathrm{sit}}\mathbf{X}_{\mathrm{sit}}^\tau K_{h_s^*}(U_{\mathrm{sit}} - u)\right)^{-1} \otimes\begin{pmatrix}\Im_{11} & \Im_{12}\\ \Im_{21} & \Im_{22}\end{pmatrix}\end{aligned}$$

The next theorem shows the consistency of $\widehat{\Sigma}_{(\boldsymbol{\alpha}_s, \boldsymbol{\alpha}_s')}^{TS}$.

Theorem 7 *Suppose that Assumptions 1 through 6 hold. Then,* $\widehat{\Sigma}_{(\boldsymbol{\alpha}_s, \boldsymbol{\alpha}_s')}^{TS} \to_p \Sigma_{(\boldsymbol{\alpha}_s, \boldsymbol{\alpha}_s')}^{TS}$ *as* $n \to \infty$ *for any fixed* $u \in \mathcal{U}_s$.

Proof See the online supplementary file.

6 Simulation studies

In this section, we conduct simulation experiments to investigate the finite sample performance of the proposed estimation and testing methods.

Experiment 1 The data are generated from the following three-level clustered data varying-coefficient regression model:

$$Y_{\mathrm{sit}} = \alpha_{s1}(U_{\mathrm{sit}})X_{\mathrm{sit}1} + \cdots + \alpha_{sp_s}(U_{\mathrm{sit}})X_{\mathrm{sitp}_s} + \varepsilon_{\mathrm{sit}},$$
$$s = 1, \ldots, m, \;\; i = 1, \ldots, n, \;\; \text{and} \;\; t = 1, \ldots, T_s,$$

where $m = 3$, $T_1 = 2$, $T_2 = 3$ and $T_3 = 2$, $X_{1it1} = 2\xi_{1it1} + \xi_{1i1}$, with $\xi_{1it1} \sim$ i.i.d.$N(0, 1)$ and $\xi_{1i1} \sim$ i.i.d. $N(0, 1)$, $X_{1it2} = 1.5 + \xi_{1it2} + \xi_{1it}$, with $\xi_{1it2} \sim$ i.i.d.$N(0, 1)$ and $\xi_{1it} \sim$ i.i.d. $N(0, 1)$, $X_{2it1} = \xi_{2it1}^2 + \xi_{2i1}$, with $\xi_{2it1} \sim$ i.i.d.$N(0, 1)$ and $\xi_{2i1} \sim$ i.i.d. $N(0, 1)$, $X_{2it2} = 1 - \xi_{2it2} + \xi_{2it}$, with $\xi_{2it2} \sim$ i.i.d.$N(0, 1)$ and $\xi_{2it} \sim$ i.i.d. $N(0, 1)$, $X_{3it1} = \xi_{3it1} + \xi_{3i1}$, with $\xi_{3it1} \sim$ i.i.d.$N(0, 1)$ and $\xi_{3i1} \sim$ i.i.d. $N(0, 1)$, $X_{3it2} = 3 - \xi_{3it2} + \xi_{3it}$, with $\xi_{3it2} \sim$ i.i.d.$N(0, 1)$ and $\xi_{3it} \sim$ i.i.d. $N(0, 1)$, $U_{1it} \sim$ i.i.d.$U(0, 1)$, $U_{2it} \sim$ i.i.d.$U(0, 1)$, $U_{3it} \sim$ i.i.d.$U(0, 1)$, $\alpha_{11}(U_{1it}) = (U_{1it}/1.6)^2$, $\alpha_{12}(U_{1it}) = 2\cos(\pi U_{1it})$, $\alpha_{21}(U_{2it}) = U_{2it}^2$, $\alpha_{22}(U_{2it}) = \sin(2\pi U_{2it})$, $\alpha_{31}(U_{3it}) = 3U_{3it}^3$, and $\alpha_{22}(U_{3it}) = \sin(\pi U_{3it}) + U_{3it}$.

In addition,

$$\mathrm{Cov}((\varepsilon_{1i1}, \varepsilon_{1i2}, \varepsilon_{2i1}, \varepsilon_{2i2}, \varepsilon_{2i3}, \varepsilon_{3i1}, \varepsilon_{3i2})^\tau) = \begin{pmatrix} \mathbf{A}_{11}(\eta_1) & \mathbf{A}_{12}(\eta_4) & \mathbf{A}_{13}(\eta_5) \\ \mathbf{A}_{12}^\tau(\eta_4) & \mathbf{A}_{22}(\eta_2) & \mathbf{A}_{23}(\eta_6) \\ \mathbf{A}_{13}^\tau(\eta_5) & \mathbf{A}_{23}^\tau(\eta_6) & \mathbf{A}_{33}(\eta_3) \end{pmatrix},$$

with

$$\mathbf{A}_{11}(\eta) = \mathbf{A}_{33}(\eta) = \begin{pmatrix} 1+\eta & \eta \\ \eta & 1+\eta \end{pmatrix}, \; \mathbf{A}_{12}(\eta) = \begin{pmatrix} \eta & 0 & 0 \\ 0 & \eta & 0 \end{pmatrix}, \; \mathbf{A}_{13}(\eta) = \begin{pmatrix} \eta & 0 \\ 0 & \eta \end{pmatrix}$$
$$\mathbf{A}_{22}(\eta) = \begin{pmatrix} 1+\eta & \eta & \eta \\ \eta & 1+\eta & \eta \\ \eta & \eta & 1+\eta \end{pmatrix} \quad \text{and} \quad \mathbf{A}_{23}(\eta) = \begin{pmatrix} \eta & 0 \\ 0 & \eta \\ 0 & 0 \end{pmatrix}.$$

Further, we let $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6) = (0.5, 0.5, 0.5, 0.5, 0.5, 0.5)$, $(0.75, 0.75, 0.75, 0.75, 0.75, 0.75)$ and $(1.0, 1.0, 1.0, 1.0, 1.0, 1.0)$.

Sample sizes of $n = 100, 200, 300$ are considered. In each case, the number of replications is 1,000. We adopt the truncated Gaussian kernel with support on $[-1, 1]$. The bandwidth h_s^* is selected by the plug-in method, and $0.5h_s^*$ is used in the first stage of the estimation. Three estimators are considered: the usual local linear estimator which neglects all correlations ($\widehat{\alpha}(\cdot)$); the two-stage estimator which only takes the correlations within the response into account ($\breve{\alpha}^{TS}(\cdot)$), and the proposed two-stage estimator which takes both correlations within and across responses into account

($\widehat{\alpha}^{TS}(\cdot)$). The efficiency of these estimators is evaluated via the root averaged squared errors (RASE):

$$\mathrm{RASE}(\alpha_{sj}(\cdot)) = \left[n^{-1} \sum_{i=1}^{n} T_s^{-1} \sum_{t=1}^{T_s} \{\widehat{\alpha}_{sj}(U_{\mathrm{sit}}) - \alpha_{sj}(U_{\mathrm{sit}})\}^2 \right]^{\frac{1}{2}},$$

where $\widehat{\alpha}_{sj}(U_{\mathrm{sit}})$ is any estimator of $\alpha_{sj}(U_{\mathrm{sit}})$. The results are summarized in Table 1, where *sm* and *std* represent the sample mean and the standard deviation of the RASE based on the replicated samples. We also report the RASE of the residuals defined analogously.

We see from Table 1 that taking the correlations both within and across the responses into account leads to an improvement in the performance of the unknown coefficient function estimators of model (1.1). For example, when $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6) = (1.0, 1.0, 1.0, 1.0, 1.0, 1.0)$ and $n = 300$, the RASE of $\widehat{\alpha}_{11}^{TS}(\cdot)$ is only around 80% of that of $\widehat{\alpha}_{11}(\cdot)$, and 90% of that of $\check{\alpha}_{11}^{TS}(\cdot)$. The improvement obtained from using $\widehat{\alpha}^{TS}(\cdot)$ appears to be more significant when the correlations within and across the responses are stronger (i.e., the η's are large) than when they are small (i.e., the η's are small). Additionally, other things being equal, an increase in n, the number of observations, has the effect of improving the performance of all estimators of the unknown coefficient function. On the other hand, the two-stage estimator does not result in an improvement of the RASE of the residuals. At first glance, this may be a somewhat curious finding. However, it is not uncommon for estimators designed to reduce estimator variance to result in larger prediction residuals. For example, in a linear regression with autoregressive errors, the feasible generalised least squares estimator usually results in smaller estimator variance than the ordinary least squares estimator, but this reduction in estimator variance is not always matched by a simultaneous reduction in residual variability.

Experiment 2 The purpose of this experiment is to compare our methods with Zhou et al. (2011). As mentioned earlier, Zhou et al.'s method is based on a nonparametric seemingly unrelated regression approach to the estimation of the unknown coefficient functions, with the disturbances being restricted to an error component specification. The basic model framework of this experiment design is the same as that in the last experiment, but we only consider the scenario of $m = 2, T_1 = 2$ and $T_2 = 2$ for the first two groups, and assume that $(\eta_1, \eta_2, \eta_3) = (0.5, 0.5, 0.5), (0.75, 0.75, 0.75), (1.0, 1.0, 1.0)$, and

$$\mathrm{Cov}((\varepsilon_{1i1}, \varepsilon_{1i2}, \varepsilon_{2i1}, \varepsilon_{2i2})^{\tau}) = \begin{pmatrix} \mathbf{A}_{11}(\eta_1) & \mathbf{A}_{12}(\eta_3) \\ \mathbf{A}_{12}^{\tau}(\eta_3) & \mathbf{A}_{22}(\eta_2) \end{pmatrix},$$

with

$$\mathbf{A}_{11}(\eta) = \mathbf{A}_{22}(\eta) = \begin{pmatrix} 1+\eta & \eta \\ \eta & 1+\eta \end{pmatrix}, \quad \text{and} \quad \mathbf{A}_{12}(\eta) = \begin{pmatrix} \eta & 0 \\ 0 & \eta \end{pmatrix}.$$

Table 1 The finite sample performance of the estimators for the unknown coefficient functions for experiment 1

		$(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6)$ (0.5, 0.5, 0.5, 0.5, 0.5, 0.5)			$(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6)$ (0.75, 0.75, 0.75, 0.75, 0.75, 0.75)			$(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6)$ (1.0, 1.0, 1.0, 1.0, 1.0, 1.0)		
		$n = 100$	$n = 200$	$n = 300$	$n = 100$	$n = 200$	$n = 300$	$n = 100$	$n = 200$	$n = 300$
RASE($\widehat{\alpha}_{11}(\cdot)$)	sm	0.0768	0.0574	0.0474	0.0829	0.0620	0.0513	0.0898	0.0658	0.0555
	std	0.0266	0.0183	0.0147	0.0287	0.0198	0.0159	0.0306	0.0200	0.0163
RASE($\widehat{\alpha}_{12}(\cdot)$)	sm	0.1014	0.0731	0.0625	0.1075	0.0776	0.0662	0.1124	0.0840	0.0712
	std	0.0309	0.0214	0.0171	0.0333	0.0230	0.0184	0.0351	0.0249	0.0199
RASE($\widehat{\alpha}_{21}(\cdot)$)	sm	0.0809	0.0573	0.0487	0.0879	0.0622	0.0526	0.0971	0.0674	0.0562
	std	0.0276	0.0180	0.0145	0.0302	0.0194	0.0157	0.0339	0.0211	0.0171
RASE($\widehat{\alpha}_{22}(\cdot)$)	sm	0.1384	0.1118	0.0938	0.1434	0.1148	0.0962	0.1501	0.1178	0.0981
	std	0.0338	0.0239	0.0195	0.0366	0.0254	0.0209	0.0376	0.0266	0.0220
RASE($\widehat{\alpha}_{31}(\cdot)$)	sm	0.1311	0.0979	0.0812	0.1406	0.1055	0.0869	0.1489	0.1124	0.0938
	std	0.0435	0.0295	0.0245	0.0475	0.0315	0.0264	0.0497	0.0350	0.0278
RASE($\widehat{\alpha}_{32}(\cdot)$)	sm	0.0646	0.0477	0.0397	0.0682	0.0504	0.0420	0.0721	0.0524	0.0434
	std	0.0210	0.0149	0.0120	0.0224	0.0160	0.0130	0.0236	0.0168	0.0138
RASE(residuals)	sm	1.4396	1.4693	1.4771	1.6738	1.7121	1.7209	1.9190	1.9552	1.9710
RASE($\check{\alpha}_{11}^{TS}(\cdot)$)	sm	0.0734	0.0547	0.0449	0.0764	0.0569	0.0468	0.0797	0.0576	0.0490
	std	0.0245	0.0172	0.0140	0.0254	0.0179	0.0146	0.0274	0.0181	0.0145
RASE($\check{\alpha}_{12}^{TS}(\cdot)$)	sm	0.0987	0.0707	0.0606	0.1019	0.0729	0.0625	0.1047	0.0763	0.0649
	std	0.0297	0.0202	0.0163	0.0311	0.0210	0.0169	0.0320	0.0223	0.0181
RASE($\check{\alpha}_{21}^{TS}(\cdot)$)	sm	0.0759	0.0532	0.0450	0.0784	0.0548	0.0462	0.0829	0.0561	0.0472
	std	0.0253	0.0167	0.0138	0.0266	0.0172	0.0144	0.0284	0.0182	0.0145
RASE($\check{\alpha}_{22}^{TS}(\cdot)$)	sm	0.1355	0.1092	0.0915	0.1373	0.1099	0.0918	0.1384	0.1108	0.0922
	std	0.0309	0.0223	0.0181	0.0317	0.0227	0.0186	0.0329	0.0224	0.0190

Table 1 continued

		$(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6)$ (0.5, 0.5, 0.5, 0.5, 0.5, 0.5)			$(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6)$ (0.75, 0.75, 0.75, 0.75, 0.75, 0.75)			$(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6)$ (1.0, 1.0, 1.0, 1.0, 1.0, 1.0)		
		$n = 100$	$n = 200$	$n = 300$	$n = 100$	$n = 200$	$n = 300$	$n = 100$	$n = 200$	$n = 300$
RASE($\check{\alpha}_{31}^{TS}(\cdot)$)	sm	0.1270	0.0936	0.0784	0.1320	0.0976	0.0814	0.1348	0.1010	0.0835
	std	0.0415	0.0289	0.0233	0.0439	0.0296	0.0242	0.0454	0.0311	0.0259
RASE($\check{\alpha}_{32}^{TS}(\cdot)$)	sm	0.0642	0.0470	0.0392	0.0666	0.0486	0.0407	0.0685	0.0493	0.0409
	std	0.0208	0.0145	0.0120	0.0220	0.0151	0.0129	0.0227	0.0159	0.0135
RASE(residuals)	sm	1.4491	1.4748	1.4813	1.6905	1.7220	1.7284	1.9424	1.9700	1.9821
RASE($\widehat{\alpha}_{11}^{TS}(\cdot)$)	sm	0.0652	0.0477	0.0393	0.0560	0.0399	0.0327	0.0356	0.0196	0.0146
	std	0.0228	0.0151	0.0124	0.0199	0.0126	0.0102	0.0138	0.0067	0.0051
RASE($\widehat{\alpha}_{12}^{TS}(\cdot)$)	sm	0.0925	0.0654	0.0559	0.0859	0.0601	0.0515	0.0730	0.0496	0.0427
	std	0.0273	0.0179	0.0149	0.0246	0.0157	0.0130	0.0182	0.0101	0.0073
RASE($\widehat{\alpha}_{21}^{TS}(\cdot)$)	sm	0.0717	0.0494	0.0412	0.0674	0.0461	0.0379	0.0600	0.0393	0.0314
	std	0.0244	0.0158	0.0126	0.0235	0.0153	0.0119	0.0235	0.0141	0.0108
RASE($\widehat{\alpha}_{22}^{TS}(\cdot)$)	sm	0.1328	0.1075	0.0898	0.1300	0.1052	0.0876	0.1246	0.1011	0.0839
	std	0.0279	0.0205	0.0166	0.0259	0.0186	0.0152	0.0225	0.0148	0.0120
RASE($\widehat{\alpha}_{31}^{TS}(\cdot)$)	sm	0.1154	0.0834	0.0701	0.1045	0.0738	0.0617	0.0811	0.0514	0.0422
	std	0.0373	0.0260	0.0210	0.0346	0.0236	0.0191	0.0308	0.0202	0.0173
RASE($\widehat{\alpha}_{32}^{TS}(\cdot)$)	sm	0.0608	0.0442	0.0365	0.0588	0.0418	0.0347	0.0540	0.0351	0.0294
	std	0.0202	0.0140	0.0111	0.0220	0.0144	0.0117	0.0206	0.0152	0.0133
RASE(residuals)	sm	1.4616	1.4831	1.4873	1.7154	1.7392	1.7413	1.9857	1.9996	2.0050

Zhou et al.'s estimator is denoted as $(\bar{\alpha}^{TS}(\cdot))$ in Table 2. From the Table, we can see that the two estimators behave very closely to each other, meaning that our method works well.

Experiment 3 (Testing for correlations within and across responses) The purpose of this experiment is to investigate the properties of the tests proposed in Sect. 3. The data are generated from the following three-level clustered data varying-coefficient regression model:

$$Y_{\text{sit}} = \alpha_{s1}(U_{\text{sit}})X_{\text{sit1}} + \cdots + \alpha_{sp_s}(U_{\text{sit}})X_{\text{sitp}_s} + \varepsilon_{\text{sit}},$$
$$s = 1, \ldots, m, \quad i = 1, \ldots, n, \quad \text{and} \quad t = 1, \ldots, T_s,$$

where $m = 2$, T_1, T_2, X_{1it1}, X_{1it2}, X_{2it1}, X_{2it2}, U_{1it}, U_{2it} and $\alpha_{11}(U_{1it})$ have the same definitions as in Experiment 1, $\alpha_{12}(U_{1it}) = 2\cos(2\pi U_{1it}) + U_{1it}^2$, $\alpha_{21}(U_{2it}) = 3\sqrt{U_{2it}(2 - U_{2it})}\sin((2.1\pi)/(U_{2it}+0.65))$, and $\alpha_{22}(U_{2it}) = -0.5+\exp(U_{2it})/(1+\exp(U_{2it}))$. In addition, η = 0.0,0.1,0.2,0.3,0.4,0.5,0.6,0.7 and 0.8, and

$$\text{Cov}((\varepsilon_{1i1}, \varepsilon_{1i2}, \varepsilon_{2i1}, \varepsilon_{2i2}, \varepsilon_{2i3})^{\tau}) = \begin{pmatrix} \mathbf{A}_{11}(\eta) & \mathbf{A}_{12}(\eta) \\ \mathbf{A}_{12}^{\tau}(\eta) & \mathbf{A}_{22}(\eta) \end{pmatrix},$$

with

$$\mathbf{A}_{11}(\eta) = \begin{pmatrix} 1+\eta & \eta \\ \eta & 1+\eta \end{pmatrix}, \quad \mathbf{A}_{12}(\eta) = \begin{pmatrix} \eta & 0 & 0 \\ 0 & \eta & 0 \end{pmatrix} \quad \text{and}$$
$$\mathbf{A}_{22}(\eta) = \begin{pmatrix} 1+\eta & \eta & \eta \\ \eta & 1+\eta & \eta \\ \eta & \eta & 1+\eta \end{pmatrix}.$$

Four tests are considered: $\widehat{M}_1$ and $\widehat{M}_2$ test for error correlations within the response levels $s = 1$ and $s = 2$ respectively, $\widehat{M}_{12}$ tests for correlations between these two response levels, while $\widehat{M}$ tests for correlations both within and between the response levels. In each case, we set the nominal significance level to 0.05. The empirical powers of these tests are summarized in Table 3. The power value at η=0 is the true significance level of the test in each case. We observe that the true test size in each case is very close to the nominal 0.05 level, and all tests exhibit very reasonable power. As η increases, the powers of all tests invariably approach 1, and the speed in which the test power approaches 1 increases as n increases, *ceteris paribus*.

7 An application

We now consider an application of the proposed estimation and testing procedures using a dataset extracted from the STARS database of the World Bank. We obtained values of Gross Domestic Product (GDP) and aggregate physical capital stock (in 1987 prices converted into US dollars at the prevailing exchange rate) for 81 countries. We

Table 2 The finite sample performance of the estimators for the unknown coefficient functions for experiment 2

		(η_1, η_2, η_3) (0.5, 0.5, 0.5)			(η_1, η_2, η_3) (0.75, 0.75, 0.75)			(η_1, η_2, η_3) (1.0, 1.0, 1.0)		
		$n = 100$	$n = 200$	$n = 300$	$n = 100$	$n = 200$	$n = 300$	$n = 100$	$n = 200$	$n = 300$
RASE($\widehat{\alpha}_{11}^{TS}(\cdot)$)	sm	0.0680	0.0496	0.0415	0.0612	0.0432	0.0357	0.0415	0.0231	0.0171
	std	0.0244	0.0153	0.0127	0.0219	0.0137	0.0112	0.0155	0.0082	0.0060
RASE($\widehat{\alpha}_{12}^{TS}(\cdot)$)	sm	0.0924	0.0667	0.0584	0.0879	0.0624	0.0545	0.0744	0.0503	0.0443
	std	0.0272	0.0183	0.0156	0.0252	0.0165	0.0142	0.0197	0.0109	0.0086
RASE($\widehat{\alpha}_{21}^{TS}(\cdot)$)	sm	0.0904	0.0596	0.0499	0.0828	0.0536	0.0439	0.0614	0.0349	0.0256
	std	0.0328	0.0200	0.0156	0.0302	0.0184	0.0140	0.0248	0.0139	0.0101
RASE($\widehat{\alpha}_{22}^{TS}(\cdot)$)	sm	0.1510	0.1255	0.1048	0.1452	0.1217	0.1009	0.1316	0.1122	0.0923
	std	0.0352	0.0249	0.0205	0.0314	0.0222	0.0180	0.0222	0.0132	0.0091
RASE($\bar{\alpha}_{11}^{TS}(\cdot)$)	sm	0.0674	0.0495	0.0414	0.0598	0.0428	0.0353	0.0369	0.0197	0.0143
	std	0.0242	0.0153	0.0126	0.0215	0.0133	0.0109	0.0139	0.0068	0.0048
RASE($\bar{\alpha}_{12}^{TS}(\cdot)$)	sm	0.0926	0.0665	0.0581	0.0879	0.0616	0.0537	0.0723	0.0470	0.0413
	std	0.0270	0.0182	0.0156	0.0245	0.0163	0.0139	0.0170	0.0090	0.0067
RASE($\bar{\alpha}_{21}^{TS}(\cdot)$)	sm	0.0900	0.0596	0.0498	0.0813	0.0532	0.0434	0.0554	0.0310	0.0225
	std	0.0324	0.0199	0.0156	0.0292	0.0180	0.0137	0.0218	0.0115	0.0081
RASE($\bar{\alpha}_{22}^{TS}(\cdot)$)	sm	0.1499	0.1256	0.1048	0.1433	0.1218	0.1009	0.1277	0.1117	0.0920
	std	0.0352	0.0248	0.0203	0.0312	0.0220	0.0178	0.0213	0.0124	0.0088

Table 3 Empirical test powers at the nominal 0.05 significance level

	n	$\eta = 0.0$	$\eta = 0.1$	$\eta = 0.2$	$\eta = 0.3$	$\eta = 0.4$	$\eta = 0.5$	$\eta = 0.6$	$\eta = 0.7$	$\eta = 0.8$
$\widehat{M}_1$	100	0.0630	0.1940	0.5400	0.7520	0.9140	0.9790	0.9956	0.9990	1.0000
	200	0.0580	0.3140	0.7670	0.9950	0.9980	1.0000	1.0000	1.0000	1.0000
	300	0.0520	0.4220	0.8780	0.9940	1.0000	1.0000	1.0000	1.0000	1.0000
$\widehat{M}_2$	100	0.0540	0.3262	0.7650	0.9550	0.9890	1.0000	1.0000	1.0000	1.0000
	200	0.0560	0.5490	0.9600	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	300	0.0530	0.6900	0.8780	0.9970	1.0000	1.0000	1.0000	1.0000	1.0000
$\widehat{M}$	100	0.0540	0.5920	0.9680	0.9970	1.0000	1.0000	1.0000	1.0000	1.0000
	200	0.0510	0.8640	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	300	0.0470	0.9510	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\widehat{M}_{12}$	100	0.0480	0.4520	0.9220	0.9950	1.0000	1.0000	1.0000	1.0000	1.0000
	200	0.0570	0.7410	0.9990	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	300	0.0460	0.8800	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

also obtained the number of individuals in the workforce between the ages of 15 and 64 for all 81 countries, as well as average level of schooling (in years). We considered data of the following two periods: 1960–1969 and 1980–1987.

Our interest in the STARS dataset is to determine how real capital, labor supply and the educational level of workers affect GDP in all 81 countries during the two specified time periods. We also think that the educational level of the labor force interacts with real capital and labor supply. The three levels of correlation here are quite clear: the correlation between capital, education and labor; the correlations between all 81 countries; and the correlation between data from the '60 to '69 period, and that from the '80 to '87 period. It makes sense, therefore, to specify the following three-level clustered data varying-coefficient model to address our question of interest:

$$\begin{aligned} Y_{\text{sit}} &= \alpha_{s1}(U_{\text{sit}}) + X_{\text{sit}2}\alpha_{s2}(U_{\text{sit}}) + X_{\text{sit}3}\alpha_{s3}(U_{\text{sit}}) + \varepsilon_{\text{sit}}, \quad \text{for } s = 1, 2, \\ &\quad i = 1, \ldots, 81, \ t = 1, \ldots, T_s, \end{aligned} \tag{7.1}$$

with $T_1 = 10$ and $T_2 = 8$, where Y_{sit}, $X_{\text{sit}2}$, $X_{\text{sit}3}$ and U_{sit} represent, respectively, the log of GDP, the log of real capital, the log of labor supply, and the log of mean year of schooling of the workforce for country i in year t of period s ($s = 1$ for the period 1960–1969, and $s = 2$ for the period 1980–1987).

Using tests developed in Sect. 3, we look for correlations with and across the responses. Table 4, which summarizes the results, shows that the correlations both within and across responses are significant. The unknown coefficient functions are then estimated based on the two-stage procedure in Sect. 4. The estimated coefficient functions are plotted in Fig. 1a–c, in which the solid curve in each figure is for $s = 1$ and the dashed curve for $s = 2$. Figure 2a and 2b displays the fitted residuals for $s = 1$ and 2.

Figure 1a, which gives the fitted curves of $\alpha_{11}(\cdot)$ and $\alpha_{21}(\cdot)$, shows that log GDP generally increases with the log mean years of schooling during both the 1960s and

Table 4 Testing results of correlations within and across responses

Test statistic	Value of test statistic	Critical value	*P* value
$\widehat{M}_1$	3428.54	61.61	0.0000
$\widehat{M}_2$	2028.56	41.37	0.0000
$\widehat{M}$	8535.21	182.98	0.0000
$\widehat{M}_{12}$	3056.21	101.84	0.0000

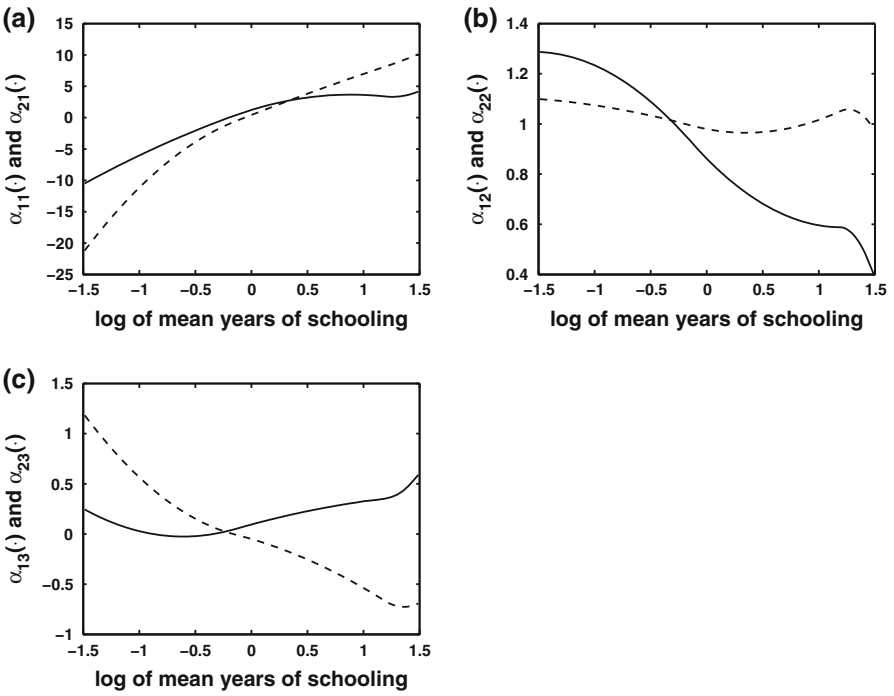

Fig. 1 *Panel a* the fitted curves of $\alpha_{11}(\cdot)$ (*solid curve*) and $\alpha_{21}(\cdot)$ (*dashed curve*). *Panel b* the fitted curves of $\alpha_{12}(\cdot)$ (*solid curve*) and $\alpha_{22}(\cdot)$ (*dashed curve*). *Panel c* the fitted curves of $\alpha_{13}(\cdot)$ (*solid curve*) and $\alpha_{23}(\cdot)$ (*dashed curve*)

1980s. It is also observed that log GDP is more elastic to changes in the years of schooling in the 1980s than in the 1960s; this is not entirely unexpected given the stronger demand for skilled labor as a result of the advances in technology experienced in the 1980s. Figure 1b displays the fitted curves of $\alpha_{12}(\cdot)$ and $\alpha_{22}(\cdot)$, showing the effects of the log of real capital on the log of GDP as education, the effect modifier, varies. These plots indicate that in both 1960s and 1980s, real capital has a positive effect on GDP, but as education level rises, a decrease in real capital's impact on GDP is observed for the 1960s, while the effect is relatively stable for the 1980s. The reason could again be related to technology; the educational level of the labor force was far more important to capital formation in the 1980s than in the 1960s. This explanation is corroborated by the plots of Fig. 1c, where it is shown that in the 1960s, labor supply had a positive effect on GDP and exhibited a generally mild positive correlation with

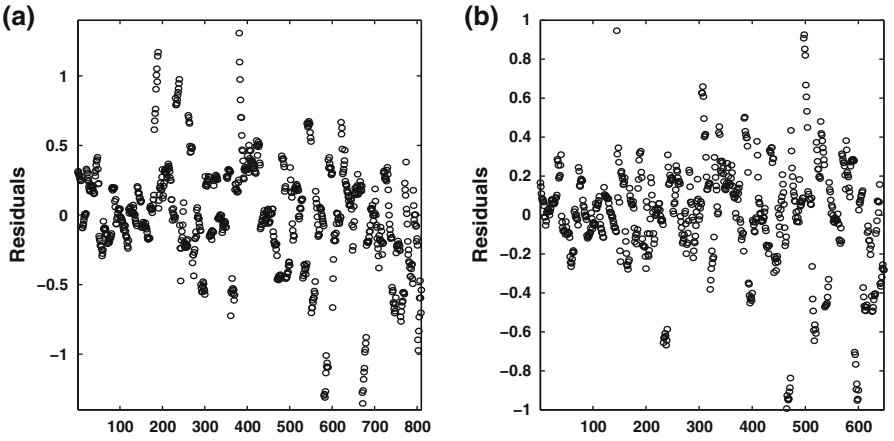

Fig. 2 *Panel a* the fitted residuals for period 1960–1969. *Panel b* the fitted residuals for period 1980–1987

educational level; in the 1980s, it was possible for a rise in the educational level of the labor force to have a negative effect on the contribution of labor supply to GDP because during the 1980s the US economy became less labor intensive, while the work force became more skilled. In this situation, a more educated labor force had more impact on GDP through its effect on capital formation than through the size of the labor force itself. Figure 2a and b shows that the fitted residuals are fairly randomly distributed around zero, providing support for the goodness of fit of the model.

8 Concluding remarks

Situations involving multi-level clustered data abound in practice, and the preceding analysis clearly shows that the proposed method is useful. Throughout the paper, we have made no assumption about the structure of the error covariance matrix Σ. This is a merit from a generality standpoint, but when s and T_s are large, the number of unknowns in Σ can quickly exceed the number of observations. How to reduce the problem's dimensionality while maintaining the flexibility of the error covariance is an important question to address in the future.

Acknowledgments We thank the editor, associate editor and two referees for helpful comments. Wan's work was based on a Strategic Research Grant from the City University of Hong Kong (No. 7008134). The usual disclaimer applies.

References

Baltagi BH (2008) Econometric analysis of panel data. Wiley, New York

Beierlein JG, Dunn JW, McConnon JR (1981) The demand for electricity and natural gas in the northeastern United States. Rev Econ Stat 63:403–408

Chapman AB, Guay-Woodford LM, Grantham JJ, Torres VE, Bae KT, Baumgarten DA, Kenney PJ, King BF, Glockner JF, Wetzel LH, Brummer ME, O'Neill WC, Robbin ML, Bennett WM, Klahr S, Hirschman GH, Kimmel PL, Thompson PA, Miller JP (2003) Renal structure in early autosomal-dominant polycystic kidney disease (ADPKD): the Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP) cohort. Kidney Int 64:1035–1045

Chen S, Zhong P (2011) ANOVA for longitudinal data with missing values. Ann Stat 38:3630–3659
Cleveland WS, Gross E, Shhyu WM (1991) Local regression models. In: Statistical models in S Chambers JM, Hastie TJ (eds) Wadsworth and Brooks, Pacific Grove, pp 309–376
Diggle PJ, Liang KY, Zeger SL (1994) Analysis of longitudinal data. Oxford University Press, Oxford
Fan J, Gijbels I (1996) Local polynomial modelling and its applications. Chapman and Hall, London
Fan J, Huang T, Li R (2007) Analysis of longitudinal data with semiparametric estimation of covariance function. J Am Stat Assoc 35:632–641
Fan J, Zhang J (2000) Two-step estimation of functional linear models with applications to longitudinal data. J R Stat Soc Ser B 62:303–322
Hastie T, Tibshiran R (1993) Varying-coefficient models. J R Stat Soc Ser B 55:757–796
Hoover DR, Rice JA, Wu CO, Yang L (1998) Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. Biometrika 85:809–822
Lin X, Carroll RJ (2006) Semiparametric estimation in general repeated measures problems. J R Stat Soc Ser B 68:69–88
Tsay WJ (2004) Testing for contemporaneous correlation of disturbances in seemingly unrelated regressions with serial dependence. Econ Lett 83:69–76
Wan G, Griffiths WE, Anderson JR (1992) Using panel data to estimate risk effects in seemingly unrelated production functions. Empir Econ 17:35–49
Wang N (2003) Marginal nonparametric kernel regression accounting for within-subject correlation. Biometrika 90:43–52
Welsh A, Lin X, Carroll RJ (2002) Marginal longitudinal nonparametric regression: locality and efficiency of Spline and Kernel methods. J Am Stat Assoc 97:482–493
Zellner A (1962) An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. J Am Stat Assoc 57:348–368
Zhang W, Fan J, Sun Y (2009) A semiparametric model for cluster data. Ann Stat 37:2377–2408
Zhou B, You J, Xu Q (2011) Efficient estimation for error component seemingly unrelated nonparametric regression models. Metrika 73:121–138