Nonseparable Panel Models with Index Structure and Correlated Random Effects

Cizek, Pavel; Sadikoglu, Serhan

Publication date:
2022

Document Version
Early version, also known as pre-print

CentER

Discussion Paper

No. 2022-009

NONSEPARABLE PANEL MODELS WITH INDEX STRUCTURE AND CORRELATED RANDOM EFFECTS

By

Pavel Čižek, Serhan Sadikoğlu

6 April 2022

ISSN 0924-7815
ISSN 2213-9532

Tilburg University

Nonseparable panel models with index structure and correlated random effects

Pavel Čížek* Serhan Sadikoğlu*

April 5, 2022

Abstract

To facilitate semiparametric estimation of general discrete-choice, censored, sample selection, and other complex panel data models, we study identification and estimation of nonseparable multiple-index models in the context of panel data with correlated random effects and a fixed number of time periods. The parameter vectors of interest are shown to be identified up to multiplicative constants and the average marginal effects are identified under the assumption that the distribution of individual effects depends on the explanatory variables only through their averages across time. Under this assumption, we propose to estimate the unknown parameters by the generalized method of moments based on the average and outer product of the difference of derivatives of the regression function. The rate of convergence and asymptotic distribution are established both for the proposed parameter estimates and the average marginal effects. We conduct Monte Carlo simulation study to assess finite-sample performance of the proposed estimator and provide an application demonstrating the use of the proposed methodology.

JEL codes: C13, C14, C23

Key words: correlated random effects, local polynomial smoothing, multiple-index model, nonlinear panel data, nonseparable models, outer product of gradients

*Department of Econometrics and Operation Research, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Phone: +31-13-466-8723, Fax: +31-13-466-3280, E-mail: P.Cizek@tilburguniversity.edu

1 Introduction

Many practically applied models such as linear, binary, or censored regression with heteroskedasticity, regression models with sample selection, multinomial-choice models, partially linear single-index models, and practically all multiple-equation models can be formulated as multiple-index models, explaining the response variables by means of several linear combinations of explanatory variables. We study their identification and estimation in the panel data context with correlated random effects (CRE). Hence, we consider a general nonseparable multiple-index panel data model for a random sample of n individuals observed for T time periods:

$$Y_{it} = \phi_t(X_{it}^\top \beta_1, \ldots, X_{it}^\top \beta_R, \alpha_i, U_{it}), \quad i = 1, ..., n, \quad t = 1, ..., T. \tag{1}$$

In this model, Y_{it} represents a vector of d_y dependent variables, X_{it} is a vector of d explanatory variables, ϕ_t is an unknown function specific to time period t, $\beta_1, \ldots, \beta_R$ denote the R linear combinations of the d explanatory variables ($R < d$), α_i is a vector of d_a individual effects correlated with X_{it} (d_a can be any finite number), and U_{it} represents all other unobservables, which are independent of α_i and X_{it}. Interest centers on the identification and estimation of parameter vectors $\beta_1, \ldots, \beta_R$ together with the average marginal effects of X_{it} on Y_{it} using panel data with a fixed finite number T of time periods.

More specifically, we identify the quantities of interest under a CRE assumption inspired by and analogous to the approach of Mundlak (1978) and Chamberlain (1982), which has been used in empirical research using various multiple-index panel models such as bivariate discrete-choice models (Schulz et al., 2014), censored hurdle regression (Christelis and Sanz-de-Galdeano, 2011), multinomial-choice problems (Boll et al., 2016), and sample selection models (Lechmann and Wunder, 2017). Model (1) encompasses all these cases along with many other models, including some not adapted to the panel data yet (e.g., generalized partial linear single-index models of Carroll et al., 1997). There are of course alternatives to the CRE assumption. For example in the specific case of univariate reponse ($d_y = 1$) and a single index ($R = 1$), there are several existing results for model (1) that require monotonicity of ϕ_t instead of the CRE assumption (e.g., Abrevaya, 2000; Botosaru and Muris, 2017; Freyberger, 2018). To highlight the main differences between the proposed model and existing results, we focus on two specific examples of multiple-index models: the heteroskedastic censored regression and sample-selection model.

Example: censored regression. Next to the parametric Tobit models, semiparametric censored-regression estimators were proposed for panel data by many authors (e.g., Honore, 1992; Honore and Hu, 2004; Abbrevaya and Shen, 2014). They are based on the latent linear response $Y_{it}^* = X_{it}^\top \beta_1 + \alpha_{1i} + U_{it}$ transformed by a known monotonic function such as $\max\{0, Y_{it}^*\}$ and the latent errors U_{it} having a general distribution independent of the covariates. While these assumptions allow estimation without restricting the relation of the individual effects α_{1i} and covariates X_{it}, they are violated once the error distribution changes over time, for example, under heteroskedasticity with the error variance depending on X_{it}. On the other hand, the proposed model (1) under CRE accommodates a general censored regression model with Y_{it}^* being an unknown nonlinear function of index $X_{it}^\top \beta_1$ and individual effects α_{1i} and latent errors having conditional variance depending on another linear combination $X_{it}^\top \beta_2$ and some other individual effects α_{2i}. For example, the censored model

$$Y_{it} = \max\{0, \phi_{1t}(X_{it}^\top \beta_1, \alpha_{1i}) + \phi_{2t}(X_{it}^\top \beta_2, \alpha_{2i}) U_{1it}\} \tag{2}$$

is a special case of (1) and can be estimated including the corresponding marginal effects, whereas the existing semiparametric methods require linear ϕ_{1t} and $\phi_{2t}(X_{it}^\top \beta_2, \alpha_{2i}) \equiv \phi_2(\alpha_{2i})$.

Example: sample selection. Apart from the parametrically specified models (e.g., Semykina and Wooldridge, 2013, 2018), semiparametric sample-selection estimators were proposed for panel data by Kyriazidou (1997, 2001) and Gayle and Viauroux (2007) for the linear outcome model, while Klein et al. (2015) focused on the binary outcomes. Most of these works do not restrict, thanks to linearity of the outcome equation, the relation of the individual effects α_i and covariates X_{it}, but do not address the estimation of all marginal effects. The exception exist only for much more restrictive pure random effects α_i independent of X_{it} (e.g., Klein et al., 2015). In comparison, the proposed model (1) under CRE accommodates a general sample selection model with $Y_{it} = (Y_{1it}, Y_{2it})$, the outcome variable Y_{2it}, and the selection variable Y_{1it},

$$Y_{2it} = \phi_{2t}(X_{it}^\top \beta_2, X_{it}^\top \beta_1, \alpha_{2i}, \alpha_{1i}, U_{2it}) \text{ observed if } Y_{1it} = \phi_{1t}(X_{it}^\top \beta_1, \alpha_{1i}, U_{1it}) > 0, \tag{3}$$

and facilitates identification of the coefficients and various types of average marginal effects (see Sections 2 and 5 for details). The typical case of the linear outcome equation then corresponds

to $\phi_{t2}(X_{it}^\top\beta_2, X_{it}^\top\beta_1, \alpha_{2i}, \alpha_{1i}, U_{2it}) = X_{it}^\top\beta_2 + \alpha_{2i} + g_t(X_{it}^\top\beta_1, \alpha_{1i}) + U_{2it}$ in model (3), where $g_t(X_{it}^\top\beta_1, \alpha_{1i})$ represents the sample-selection correction term (cf. Kyriazidou, 1997).

In the context of the above mentioned and other models, many existing semiparametric approaches rely on the monotonicity of the response Y_{it} as a function of an index $X_{it}^\top\beta$ to identify its coefficients without additional assumptions on the relationship between the individual effects and covariates and without modelling the error distribution and its relationship to covariates. However, the latter relationship is often important for the identification of the coefficients or marginal effects, for example in the limited dependent variable models with heteroskedasticity. Hence, the proposed multi-index model (1) explicitly models both the relationships of interest and auxiliary relationships, for example the error variance as a function of covariates, to be able to identify all coefficients and average marginal effects. To achieve this though, we have to impose an additional assumption – CRE – on the individual effects.

1.1 Overview of and links to existing literature

The identification and estimation of average marginal effects in short nonseparable panel models have been studied by several authors. Using time-homogeneity, Chernozhukov et al. (2013) derived bounds for marginal effects in static and dynamic models. Additionally under monotonicity, Ishihara (2020) identifies the regression function and Freyberger (2018) extends the analysis to models with interactive effects. On the other hand, Bester and Hansen (2009) showed that average marginal effects in a CRE model can be identified if the distribution of individual effects depends on explanatory variables only through an index function. Hoderlein and White (2012) established that the average marginal effects at $X_{it} = X_{it-1}$ can be identified by means of a generalized version of differencing in a static nonseparable model. Furthermore, Čížek and Lei (2018) studied nonseparable single-index panel data model and demonstrated that differencing average derivatives of a specific regression function can identify the index parameters and average marginal effects even in dynamic models. Let us also note that the mentioned identification assumptions such as time-homogeneity or CRE can be tested as discussed by Ghanem (2017) and that identification of quantities beyond the average marginal effects have been explored as well (e.g., Chernozhukov et al., 2013, 2015).

The mentioned nonparametric identification results have been studied and adapted to single-index models, $R = 1$ in (1), by Chen and Wang (2018) and Čížek and Lei (2018). These results

do not extend to multiple-index models, $R > 1$ in (1), as, for one dependent variable, the generalized differences or first derivatives taken with respect to each variable provide only d conditions to identify one vector of d parameters. For cross-sectional and longitudinal data, several estimation approaches were introduced though, where multiple linear combinations are typically identified by averaging functions of derivatives or variances of conditional expectations of responses or regression residuals. The average derivative estimation (Härdle and Stoker, 1989) was adapted to the multiple-index estimation by taking higher-order derivatives (Li, 1992), the outer product of gradients (Samarov, 1993), or both (Donkers and Schafgans, 2008). Alternative approaches include the minimum average variance estimation (Xia et al., 2002), the estimating equation approach (Xu et al., 2016) and the sliced inverse regression (Zhu et al., 2016), for instance. These techniques cannot be easily generalized to nonlinear panel data with a fixed number of time periods due to the presence of the unobserved heterogeneity represented by individual effects α_i in (1). Consequently, general multiple-index panel models have been studied so far only for panel data with large numbers of time periods, which facilitate consistent estimation of the individual specific effects (Xu et al., 2016).

Focusing on panel data with a limited number of time periods and one or more responses, our approach to nonseparable multiple-index models is built on the assumption employed by Bester and Hansen (2009) for marginal-effect identification and Čížek and Lei (2018) for single-index models: the distribution of unobserved individual effects depends on the observed covariates through their averages across time. This restricts the analysis to the CRE models, but allows for flexible time-varying specification as in Botosaru and Muris (2017), Freyberger (2018), and Ishihara (2020). To identify the multiple coefficient vectors $\beta_1, \ldots, \beta_R$, average differences of first derivatives in Čížek and Lei (2018) have to be replaced by the second-order derivatives or outer product of gradients. Given the benefits of the latter (see Xia et al., 2002), we propose to employ the outer product of differences of gradients (OPDG) and the generalized method of moments (GMM) to identify and estimate the parameters of a model with multiple linear indices, similarly to Donkers and Schafgans (2008) for cross-sectional data. The proposed estimation method retains several appealing features: (i) it delivers consistent results even with only two or three time periods, (ii) it applies directly to unbalanced data, and under some regularity conditions, (iii) it allows lagged dependent and discrete explanatory variables to enter the model. We also discuss how to estimate in the presence of functionally related regressors,

which are typically not allowed by methods based on nonparametrically estimated gradients (cf. Čížek and Lei, 2018; Donkers and Schafgans, 2008). Last but not least, the proposed model (1) is rather flexible as it covers a wide array of nonlinear panel data models used in applications mentioned earlier. Under the CRE assumption, the proposed approach offers a generally applicable semiparametric method for estimation of (non)linear sample selection models in panel data as well as for nonlinear panel models based on a latent partially-linear single-index structure (cf. Carroll et al., 1997); see Sections 2, 4, and 5.

The paper is organized as follows. For the simplicity of exposition, the key identification and asymptotic results are presented for two time periods, $T = 2$. The main identification result along with its assumptions are presented in Section 2. Next, the proposed semiparametric OPDG estimation procedure and the corresponding GMM estimator are introduced in Section 3. In that section, we also study the asymptotic properties of the proposed estimators and derive their rates of convergence and asymptotic distribution. In Sections 4 and 5, we assess the finite sample performance of the proposed method for nonlinear panel data models in a simulation study and real-data application. Simulation results for $T > 2$, details on estimation with discrete and functionally related regressors, and proofs are relegated to the Appendices.

2 Identification

To study the identification of the nonseparable panel data model with index structure given in equation (1), we consider panel data with n observations of time series $Y_i = (Y_{i1}, \ldots, Y_{iT})^\top$ and $X_i = (X_{i1}, \ldots, X_{iT})^\top$, which are independent and identically distributed across cross-sectional units $i \in \{1, \ldots, n\}$. The number T of time periods is assumed to be finite and fixed, and for the simplicity of exposition, the identification and estimation results are presented for two time periods. The two considered time periods are the current time period t and some past time period $t - \Delta$ with the typical choice being $\Delta = 1$, and they are labelled $\mathfrak{T} = (t, t - \Delta)$. Constructing the estimation procedure and its moment conditions for given two time periods will extend directly to more time periods since the moment conditions constructed for each available pair of time periods can be used jointly to estimate model (1). This approach with separate moment conditions for any two time periods also facilitates a straightforward application in unbalanced panel data. Alternative approaches to construction of the moment conditions using

more than two time periods are discussed in Supplementary Appendix D.

Let us demonstrate the identification principle using two time periods $t, t-\Delta$ and the standard linear regression model, $Y_{it} = X_{it}^\top \beta + \alpha_i + \varepsilon_{it}$, with strictly exogenous explanatory variables X_{it}, $E(\varepsilon_{it}|X_{it}, X_{i(t-\Delta)}) = 0$. Taking expectations $E(Y_{it}|X_{it}, X_{i(t-\Delta)}) = X_{it}^\top \beta + E(\alpha_i|X_{it}, X_{i(t-\Delta)})$, the standard approach to handle the individual effects α_i in this model is to eliminate them by taking differences: $E(Y_{it} - Y_{i(t-\Delta)}|X_{it}, X_{i(t-\Delta)}) = (X_{it} - X_{i(t-\Delta)})\beta$. Since this approach is not applicable in nonseparable models (1), we instead impose the following CRE assumption $E(\alpha_i|X_{it}, X_{i(t-\Delta)}) = E(\alpha_i|X_{it} + X_{i(t-\Delta)})$, where the individual effects depend on the covariates only through their sum or average across the time periods $t, t-\Delta$ (cf. Bester and Hansen, 2009; Čížek and Lei, 2018). This allow us to eliminate the individual effects by taking the derivatives with respect to the current and past values X_{it} and $X_{i(t-\Delta)}$ as

$$\frac{\partial E(Y_{it}|X_{it}, X_{i(t-\Delta)})}{\partial X_{it}} = \beta + \frac{\partial E(\alpha_i|X_{it} + X_{i(t-\Delta)})}{\partial X_{it}} \text{ and } \frac{\partial E(Y_{it}|X_{it}, X_{i(t-\Delta)})}{\partial X_{i(t-\Delta)}} = \frac{\partial E(\alpha_i|X_{it} + X_{i(t-\Delta)})}{\partial X_{i(t-\Delta)}}$$

imply that $DE(X_{it}, X_{i(t-\Delta)}) = \partial E(Y_{it}|X_{it}, X_{i(t-\Delta)})/\partial X_{it} - \partial E(Y_{it}|X_{it}, X_{i(t-\Delta)})/\partial X_{i(t-\Delta)} = \beta$, and subsequently, that $E\{DE(X_{it}, X_{i(t-\Delta)})DE(X_{it}, X_{i(t-\Delta)})^\top\} - \beta\beta^\top = 0$. As we will show, this moment equation, which eliminates the individual effects by taking the difference of the derivatives and which identifies the parameters by taking the outer product of this difference, applies also in the non-separable models with multiple linear combinations $X_{it}^\top \beta_1, \ldots, X_{it}^\top \beta_R$.

More specifically, the proposed methodology applies to and is presented here for one or more response variables forming a vector $Y_{it} \in \mathbb{R}^{d_y}$. The model (1) can be concisely expressed as

$$Y_{it} = \phi_t(X_{it}^\top \beta_1, \ldots, X_{it}^\top \beta_R, \alpha_i, U_{it}) = \phi_t(X_{it}^\top B, \alpha_i, U_{it}), \tag{4}$$

where $B = (\beta_1, \ldots, \beta_R)$ is $d \times R$ matrix containing the coefficients of the R linear combinations of d explanatory variables X_{it}. The number R of indices is assumed to be implied by the model and thus treated as known now. Finding the dimension R is discussed later in Section 3.3.

Examples. The multiple-index model (4) includes many panel-data models such as (i) heteroscedastic binary-choice models, $Y_{it} = \mathbb{1}\{X_{it}^\top \beta_1 + \alpha_{1i} + \sigma_t(X_{it}^\top \beta_2, \alpha_{2i})U_{it} > 0\}$, and (ii) censored models (2), $Y_{it} = \max\{0, X_{it}^\top \beta_1 + \alpha_{1i} + \sigma_t(X_{it}^\top \beta_2, \alpha_{2i})U_{it} > 0\}$, with an unknown standard deviation $\sigma_t(\cdot)$ and individual effects α_{1i} and α_{2i}, (iii) general transformation models with

partially-linear single-index structure $Y_{it} = g_t\{X_{it}^\top\beta_1 + \alpha_{1i} + h_t(X_{it}^\top\beta_2, \alpha_{2i}) + U_{it}\}$ with unknown functions $g_t(\cdot)$ and $h_t(\cdot)$, and (iv) multinomial choice models $Y_{it} = \text{argmax}_{j=1,\dots,J} g_{jt}(X_{it}^\top\beta_j + \alpha_{ji} + U_{jit})$ with individual effects α_{ji} and unknown link functions g_{jt}. It also covers (v) (non)linear sample selection models (3): for example, the linear model $Y_{2it} = X_{it}^\top\beta_2 + \alpha_{2i} + \varepsilon_{2it}$ observed when the selection variable $Y_{1it} = \mathbb{1}\{h_t(X_{it}^\top\beta_1, \alpha_{1i}, U_{1it}) > 0\}$ equals 1 can be formulated for the outcome variable as $Y_{2it} = X_{it}^\top\beta_2 + \alpha_{2i} + g_t(X_{it}^\top\beta_1, \alpha_{1i}) + U_{2it}$, where $g_t(X_{it}^\top\beta_1, \alpha_{1i}) = \text{E}_U\,\text{E}(\varepsilon_{2it}|h_t(X_{it}^\top\beta_1, \alpha_{1i}, U_{1it}) > 0)$ and α_{1i} and α_{2i} denote again individuals effects.

Since we focus on the identification of the parameters $B \subseteq \mathbb{R}^{d\times R}$ along with the average marginal effects of X_{it} on Y_{it} given α_i, we first introduce the required assumptions in Section 2.1. Later in Section 2.2, the key identification results are derived and linked to the average marginal effects defined by $E_\alpha[\partial\varphi_t(X_{it}^\top B, \alpha_i)/\partial X_{it}]$ with $\varphi_t(X_{it}^\top B, \alpha_i) = \text{E}_U[\phi_t(X_{it}^\top B, \alpha_i, U_{it})]$.

2.1 Identification assumptions

Here we state the assumptions for the identification of B, which mostly characterize the CRE structure and are thus multivariate extensions of the assumptions in Čížek and Lei (2018).

Assumption 1. *Let (Ω, F, P) be a complete probability space on which are defined the random vectors $\alpha_i : \Omega \to \mathcal{A}$, $X_{i(t-\Delta)} : \Omega \to \mathcal{X}$, and $(Y_{it}, X_{it}, U_{it}) : \Omega \to \mathcal{Y}\times\mathcal{X}\times\mathcal{U}$, $\mathcal{A} \subseteq \mathbb{R}^{d_a}, \mathcal{Y} \subseteq \mathbb{R}^{d_y}, \mathcal{X} \subseteq \mathbb{R}^d, \mathcal{U} \subseteq \mathbb{R}^{d_u}$, for any $i \in \mathbb{N}$, and finite integers d_y, d_a, d, and d_u. For all $i \in \mathbb{N}$, let (i) $\text{E}(\|Y_{it}\|^{\delta_y}) < \infty$ for some $\delta_y > 2$; (ii) $Y_{it} = \phi_t(X_{it}^\top B, \alpha_i, U_{it})$, where $B = (\beta_1, \dots, \beta_R) \in \mathcal{B} \subseteq \mathbb{R}^{d\times R}$ is a full-rank $d \times R$ matrix of parameters and ϕ_t is an unknown and nonconstant function on the support of $X_{it}^\top\beta_r$ for any $(\alpha_i, U_{it}) \in \mathcal{A}\times\mathcal{U}$, $r = 1, \dots, R$; and (iii) realizations of $(Y_{it}, X_{it}, X_{i(t-\Delta)})$ be observable, whereas those of (α_i, U_{it}) are unobservable.*

Assumption 2. *Unobservable U_{it} is independent of α_i, X_{it}, and $X_{i(t-\Delta)}$ and is identically distributed for all $i = 1, \dots, n$.*

While Assumption 1 just formalizes the data generating process (4) and confirms that there can be any finite number d_a of individual effects, the exogeneity Assumption 2 states that U_{it} should be uncorrelated with the covariates X_{it} and $X_{i(t-\Delta)}$ of the same individual unit i at two time periods t and $t - \Delta$. Let us discuss its implications in the context of an example. First, Assumption 2 in the sample-selection model (3) such as $Y_{2it} = X_{it}^\top\beta_2 + \alpha_{2i} + g_t(X_{it}^\top\beta_1, \alpha_{1i}) + U_{2it}$ above does not preclude correlation of the elements of U_{it}, and if

X_{it} contains only strictly exogenous variables, also of U_{it} across time. On the other hand, it permits lagged dependent variables to enter the model as explanatory variables provided that the unobserved U_{it} does not exhibit serial correlation (see Appendix I for further discussion). Next, Assumption 2 also allows the dependence between the traditional error term and the covariates. For example, the binary selection outcome in the sample-selection model (3) can have the form $Y_{1it} = \mathbb{1}\{h_t(X_{it}^\top \beta_1, \alpha_{1i}, U_{it}) > 0\} = \mathbb{1}\{X_{it}^\top \beta_1 + \alpha_{1i} + \sigma_t(X_{it}^\top \beta_1, \alpha_{3i}) U_{it} > 0\}$ and can explicitly model heteroskedasticity as a function $\sigma_t(X_{it}^\top \beta_1, \alpha_{3i})$ of a linear index and an individual effect, $\varepsilon_{it} = \sigma_t(X_{it}^\top \beta_1, \alpha_{3i}) U_{it}$, while it still satisfies Assumptions 1 and 2.

Assumption 3. *(i) X_{it} does not contain any time-invariant covariates: $P(X_{k,it} \neq X_{k,i(t-\Delta)}) > 0$ for all $k = 1, \ldots, d$. (ii) The joint distribution $F_{X_t, X_{t-\Delta}}$ of $(X_{it}, X_{i(t-\Delta)})$ is continuous and identical for all $i \in \mathbb{N}$. (iii) The conditional distribution $F_{\alpha|X_t, X_{t-\Delta}}$ of the individual effects α_i satisfies $F_{\alpha|X_t, X_{t-\Delta}}(\alpha_i|X_{it}, X_{i(t-\Delta)}) = F_{\alpha|X_t + X_{t-\Delta}}(\alpha_i|X_{it} + X_{i(t-\Delta)})$. (iv) $F_{X_t, X_{t-\Delta}}$ and $F_{\alpha|X_t, X_{t-\Delta}}$ are twice continuously differentiable with respect to X_t and $X_{t-\Delta}$ with uniformly bounded derivatives on $\mathcal{X}$.*

Assumption 3 contains the main assumptions for the identification of B and imposes the CRE structure as in Assumption 3 of Čížek and Lei (2018), restricting a general relationship between α_i and X_{it} or $X_{i(t-\Delta)}$ to an identical form for all individuals i (see Bester and Hansen, 2009, for an analysis of various CRE assumptions and Ghanem, 2017, for the test of the CRE assumptions). In particular, Assumption 3(iii) states that the conditional distribution $F_{\alpha|X_t, X_{t-\Delta}}(\alpha_i|X_{it}, X_{i(t-\Delta)})$ of α_i is assumed to be independent of i and its dependence on the explanatory variables X_{it} and $X_{i(t-\Delta)}$ occurs only through their sum $X_{it} + X_{i(t-\Delta)}$.[1] Although such an assumption is often employed in models with exogenous variables in the spirit of the Mundlak (1978) approach, it poses a constraint if X_{it} contains lagged dependent variables. In particular, it requires at least the stationary initial condition. We demonstrate this in Appendix I on the example of the dynamic sample selection model used also in the application in Section 5. Furthermore, analogously to other estimation methods based on differencing over time, Assumption 3(i) rules out the presence of time-invariant covariates in the model (4). Finally, the remaining Assumption 3(ii) imposes that the explanatory variables are continuously

[1] For two time periods, such a functional form assumption is required as Bester and Hansen (2009) showed that it is not possible to achieve identification of marginal effects in the CRE model if $F_{\alpha|X_t, X_{t-\Delta}}(\alpha_i|X_{it}, X_{i(t-\Delta)}) = F_{\alpha|X_t, X_{t-\Delta}}(\alpha_i|h(X_{it}, X_{i(t-\Delta)}))$ with a general unknown function h.

distributed; identification and estimation in the presence of discrete explanatory variables is discussed in Appendix G.

Next, we impose sufficient regularity on the function φ_t and relevant distribution functions and expectations. We adopt shorthand notations as $F(\alpha|x_t, x_{t-\Delta}) \equiv F_{\alpha|X_t,X_{t-\Delta}}(\alpha|X_{it} = x_t, X_{i(t-\Delta)} = x_{t-\Delta})$ and $f(\alpha|x_t, x_{t-\Delta}) \equiv f_{\alpha|X_t,X_{t-\Delta}}(\alpha|X_{it} = x_t, X_{i(t-\Delta)} = x_{t-\Delta})$.

Assumption 4. *Function $\varphi_t(v, \alpha) = \mathrm{E}_U[\phi_t(v, \alpha, U)]$ is an unknown twice continuously differentiable function with respect to $v \in \mathbb{R}^R$ for each $\alpha \in \mathcal{A}$. Moreover, $\mathrm{E}[\varphi'_{tr}(X_{it}^\top B, \alpha_i)] < \infty$, where $\varphi'_{tr}(x^\top B, \alpha) = \partial\varphi_t(x^\top B, \alpha)/\partial(x^\top \beta_r)$ for $r = 1, \dots R$.*

Assumption 5. *For each $(x_t, x_{t-\Delta}) \in \mathbb{R}^d \times \mathbb{R}^d$, there exists a σ-finite measure $\mu(.|x_t, x_{t-\Delta})$ that is absolutely continuous with respect to $F(\cdot|x_t, x_{t-\Delta})$ so that there exists a Radon-Nikodym density f such that $F(d\alpha|x_t, x_{t-\Delta}) = f(\alpha|x_t, x_{t-\Delta})\mu(d\alpha|x_t, x_{t-\Delta})$ for each $\alpha \in \mathcal{A}$.*

Assumption 6. *(i) Conditional expectation $\mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}]$ exists, is continuous in X_{it} and $X_{i(t-\Delta)}$, and its first derivatives exist almost surely;*
(ii) For each $(x_t, x_{t-\Delta}) \in \mathbb{R}^d \times \mathbb{R}^d$, there exists an integrable dominating function $D(\alpha_i|x_t, x_{t-\Delta})$ such that, for some $\epsilon > 0$ and any element x of x_t or $x_{t-\Delta}$,

$$\sup_{v \in \{x_s^\top \beta : \beta \in \mathcal{B}, s = t, t-\Delta\}} \max\left\{ \left|\varphi'_{tr}(v, \alpha_i) f(\alpha_i|x_t, x_{t-\Delta})\right|, \left|\varphi_t(v, \alpha_i) \frac{\partial f(\alpha_i|x_t, x_{t-\Delta})}{\partial x}\right| \right\} \leq D(\alpha_i|x_t, x_{t-\Delta}).$$

Assumption 7. *Matrices Γ_{1t} and Γ_{2t} are finite $d_y \times R$ and $R \times R$ full-rank matrices, respectively, where $\Gamma_{1t} = \left\{ \mathrm{E}\left[\varphi'_{tr}(X_{it}^\top B, \alpha_i)\right] \right\}_{r=1}^R$ and $\Gamma_{2t} = \left\{ \mathrm{E}\left[\varphi'_{tr}(X_{it}^\top B, \alpha_i)^\top \varphi'_{ts}(X_{it}^\top B, \alpha_i)\right] \right\}_{r,s=1}^R$.*

Assumption 4 simply imposes a sufficient degreee of smoothness on $\varphi_t(v, \alpha)$ together with integrability of its derivatives. Next, Assumptions 5 and 6 are essential for well-defined expectations $\mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}]$, $\partial\,\mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}]/\partial X_{it}$, and $\partial\,\mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}]/\partial X_{i(t-\Delta)}$. Additionally, Assumption 7 guarantees that all R linear indices are required in the model since the multicollinearity of the indices is ruled out. Finally, to identify B, Γ_{1t}, and Γ_{2t}, Assumption 7 has to be accompanied by some identification assumptions; examples from Donkers and Schafgans (2008) are given below.

Assumption 8. *One of the following conditions is satisfied: (i) each index $x_{it}^\top \beta_r$, $r = 1, \dots, R$, contains some explanatory variable that does not enter the other $R - 1$ indices and has a coefficient normalized to 1, that is, for each $r = 1, \dots, R$ there is some $k \in \{1, \dots, d\}$ such that*

$\beta_{rk} = 1$ and $\beta_{sk} = 0$, $s \neq r$, and $B\Gamma_{2t}B^\top$ has R distinct nonzero eigenvalues; or (ii) $B^\top B = I_R$ and Γ_{2t} is a diagonal matrix with unique nonzero elements sorted in the descending order.

Assumption 8 represents two classical identification assumptions that either rely on exclusion restrictions to uniquely identify B or order and normalize the parameters in a unique way. Although such assumptions suit for example the multinomial choice models, there are also other alternative forms of identification assumptions. For example, consider the linear sample selection model (3) with the selection variable $Y_{1it} = \mathbb{1}\{h_t(X_{it}^\top\beta_1, \alpha_{1i}, U_{1it}) > 0\}$ and the outcome variable $Y_{2it} = X_{it}^\top\beta_2 + \alpha_{2i} + g_t(X_{it}^\top\beta_1, \alpha_{1i}) + U_{2it}$. One can assume that all coefficients in β_1 and β_2 are nonzero and impose Assumption 8(ii), assuming the identification by the nonlinearity of the model (e.g., see Escanciano et al., 2016). On the other hand, one can impose a traditional exclusion restriction that one variable $X_{j,it}$ influences the selection variable Y_{1it}, $\beta_{1j} \neq 0$, but does not directly affect outcome Y_{2it}, $\beta_{2j} = 0$. Given the triangular and partially linear structure, no additional exclusion restriction is then needed because $\varphi'_{t2}(X_{it}^\top B, \alpha_i) = (0, 1)^\top$ and the complete Assumption 8 is thus not necessary.

2.2 Identification result

Now we state our main identification results under Assumptions 1–7 and 8. Note that the following theorem can be applied jointly for all responses Y_{it} or for each response separately.

Theorem 1. *Under Assumptions 1–7, B, Γ_{1t}, and Γ_{2t} satisfy the following moment equations:*

$$\delta_{\mathfrak{T}} = \mathrm{E}\left\{\frac{\partial}{\partial X_{it}^\top}\mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}] - \frac{\partial}{\partial X_{i(t-\Delta)}^\top}\mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}]\right\} = \Gamma_{1t}B^\top$$

$$\begin{aligned}\delta_{\mathfrak{T}\mathfrak{T}} = \mathrm{E}\Bigg[&\left\{\frac{\partial}{\partial X_{it}^\top}\mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}] - \frac{\partial}{\partial X_{i(t-\Delta)}^\top}\mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}]\right\}^\top \\ &\times \left\{\frac{\partial}{\partial X_{it}^\top}\mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}] - \frac{\partial}{\partial X_{i(t-\Delta)}^\top}\mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}]\right\}\Bigg] = B\Gamma_{2t}B^\top.\end{aligned}$$

Additionally, if Assumption 8 holds, θ denotes the corresponding unconstrained parameters of $(B, \Gamma_{1t}, \Gamma_{2t})$ or (B, Γ_{2t}), and $g_{\mathfrak{T}\mathfrak{T}}(\theta) = \left(g_{\mathfrak{T}}^1(\theta)^\top, g_{\mathfrak{T}\mathfrak{T}}^2(\theta)^\top\right)^\top$ with $g_{\mathfrak{T}}^1(\theta) = vec(\delta_{\mathfrak{T}} - \Gamma_{1t}B^\top)$ and $g_{\mathfrak{T}\mathfrak{T}}^2(\theta) = vec(\delta_{\mathfrak{T}\mathfrak{T}} - B\Gamma_{2t}B^\top)$, the true parameter values θ_0 are identified by $argmin_\theta g_{\mathfrak{T}\mathfrak{T}}(\theta)^\top g_{\mathfrak{T}\mathfrak{T}}(\theta)$ or $argmin_\theta g_{\mathfrak{T}\mathfrak{T}}^2(\theta)^\top g_{\mathfrak{T}\mathfrak{T}}^2(\theta)$, respectively.

According to Theorem 1, the parameters of interest can be identified by evaluating the average difference of gradients (ADG) $\delta_{\mathfrak{T}}$ and the average outer product of differences of gradients (OPDG) $\delta_{\mathfrak{T}\mathfrak{T}}$, where gradients refer in both cases to derivatives $\partial \mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}]/\partial X_{it}$ and $\partial \mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}]/\partial X_{i(t-\Delta)}$. Although we include ADG, which is essentially the multivariate form of Čížek and Lei (2018), the second part of Theorem 1 indicates that the moment conditions based on ADG are not necessary for identification and are obviously not sufficient if $d_y < R$. Hence, the identification result relies only on OPDG. As the coefficients B characterize the effects of X_{it} on responses Y_{it} in (4) for given α_i, Theorem 1 identifies them using the differences of gradients based on the following observation. Since X_{it} affects Y_{it} directly through indices $X_{it}^\top B$ and also indirectly through individual effects α_i, the derivative of $\mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}]$ with respect to X_{it} captures two effects of a change in X_{it} – via $X_{it}^\top B$ and via α_i. To eliminate the latter effect, we subtract the derivative of $\mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}]$ with respect to $X_{i(t-\Delta)}$ as $X_{i(t-\Delta)}$ influences Y_{it} only through the individual effects α_i, but given X_{it}, it does not affect $X_{it}^\top B$.

By estimating ADG $\delta_{\mathfrak{T}}$ and OPDG $\delta_{\mathfrak{T}\mathfrak{T}}$, the columns of B will be estimated up to scale under the exclusion restrictions imposed in Assumption 8(i) and up to an orthogonal transformation under Assumption 8(ii). In practice, other equivalent or additional identification assumptions and moment conditions can be constructed. For example in the sample-selection model (3), the outcome variable $Y_{2,it}$ depends on two linear combinations $X_{it}^\top B = X_{it}^\top(\beta_1, \beta_2)$, whereas the selection variable $Y_{1,it}$ depends on one linear combination $X_{it}^\top \beta_1$. In such a case, although the moment conditions suggested in Theorem 1 apply directly, it can be preferable to apply Theorem 1 and to construct the moment conditions separately for each response variable $Y_{1,it}$ and $Y_{2,it}$ rather than the whole vector $Y_{it} = (Y_{1,it}, Y_{2,it})^\top$. One can then easily impose that the moment conditions for $Y_{1,it}$ do not depend on the second index $X_{it}^\top \beta_2$, for instance.

Apart from the coefficients B, Theorem 1 and its proof also facilitate identification of various marginal effects. It is known that the individual marginal effects $\partial \mathrm{E}[Y_{it}|X_{it}, \alpha_i]/\partial X_{it}$ with the individual heterogeneity α_i kept constant cannot be estimated for only two time periods. Bester and Hansen (2009) and Wooldridge (2010, Section 2.2.5) therefore suggest to average this marginal effect $\partial \mathrm{E}[Y_{it}|X_{it}, \alpha_i]/\partial X_{it}$ over the distribution of the individual-specific effects α_i. From definition $\varphi_t(X_{it}^\top B, \alpha_i) = \mathrm{E}_U[\phi_t(X_{it}^\top B, \alpha_i, U_{it})]$, it follows that $\partial \mathrm{E}[Y_{it}|X_{it}, \alpha_i]/\partial X_{it} = \partial \varphi_t(X_{it}^\top B, \alpha_i)/\partial X_{it}$ and the suggested marginal effect (ME) can be written as

$$\int \frac{\partial \varphi_t(X_{it}^\top B, \alpha_i)}{\partial X_{it}} f(\alpha | X_{it}, X_{i(t-\Delta)}) d\alpha = \frac{\partial \, \mathrm{E}[Y_{it} | X_{it}, X_{i(t-\Delta)}]}{\partial X_{it}^\top} - \frac{\partial \, \mathrm{E}[Y_{it} | X_{it}, X_{i(t-\Delta)}]}{\partial X_{i(t-\Delta)}^\top}, \quad (5)$$

where the equality is verified in equations (A.2) and (A.5) in the proof of Theorem 1. The ME are thus equal to the difference of the two derivatives of the conditional expectations on the right-hand side of (5). Averaging them with respect to covariates results in the average marginal effect (AME) equal to $\delta_{\mathfrak{T}}$ by Theorem 1. As described later in Section 3, estimation of the derivatives in (5) is the first step required to estimate $\delta_{\mathfrak{T}}$ in Theorem 1 and estimates of (5) and AME are thus a result of the estimation procedure. We refer to $\delta_{\mathfrak{T}}$ as the total AME since it characterizes the effect of covariates via all linear combinations $X_{it}^\top B$. As the identification of these total ME and AME relies only on the first part of Theorem 1, it does not require the identification Assumption 8, which is used only to decompose the total AME $\delta_{\mathfrak{T}}$ to the scaling matrices Γ_{1t} and Γ_{2t} and the coefficient matrix B in a unique way.

We are also interested in the marginal effects characterizing the effects of covariates on Y_{it} via a particular linear combination. For example in the sample-selection model (3), variables X_{it} influence the outcome variable Y_{2it} directly through the linear combination $X_{it}^\top \beta_2$ and we can refer to this particular marginal effect as the direct AME or the AME specific to $X_{it}^\top \beta_2$. However, X_{it} also influences the average outcome Y_{2it} indirectly by means of the sample-selection correction, which is characterized by the linear combination $X_{it}^\top \beta_1$, and we can be also interested in this indirect AME specific to $X_{it}^\top \beta_1$. A similar situation arises in the multinomial-choice model $Y_{it} = \mathrm{argmax}_{j=1,\dots,J} g_{jt}(X_{it}^\top \beta_j + \alpha_{ji} + U_{jit})$, where we are interested in the probability of a particular choice $Y_{it} = j$. The covariates influence this probability either directly via the linear combination $X_{it}^\top \beta_j$ affecting the corresponding utility $g_{jt}(X_{it}^\top \beta_j + \alpha_{ji} + U_{jit})$ of option j or indirectly via the linear combinations $X_{it}^\top \beta_l$, $l \neq j$, affecting the utilities of the alternative options $l \neq j$. These marginal effects specific to particular linear combinations can be obtained on average by decomposing the total AME $\delta_{\mathfrak{T}}$: since $\partial \, \mathrm{E}[Y_{it} | X_{it}, \alpha_i] / \partial X_{it} = \partial \varphi_t(X_{it}^\top B, \alpha_i) / \partial X_{it} = \sum_{r=1}^R \varphi_{tr}'(X_{it}^\top B, \alpha_i) \beta_r^\top$, taking expectation results in $\delta_{\mathfrak{T}} = \mathrm{E}[\partial \varphi_t(X_{it}^\top B, \alpha_i) / \partial X_{it}^\top] = \sum_{r=1}^R \mathrm{E}\left[\varphi_{tr}'(X_{it}^\top B, \alpha_i)\right] \beta_r^\top = \Gamma_{1t} B^\top$ by (5) and Theorem 1. Once the matrices B and Γ_{1t} are estimated, it is thus possible to obtain the R index-specific contributions $\left\{ \mathrm{E}\left[\varphi_{tr}'(X_{it}^\top B, \alpha_i)\right] \beta_r^\top \right\}_{r=1}^R = \{\Gamma_{1t,\cdot r} B^\top\}_{r=1}^R$ to the total AME

$\delta_{\mathfrak{T}} = \mathrm{E}[\partial\varphi_t(X_{it}^\top B, \alpha_i)/\partial X_{it}^\top]$, which are again independent of the normalization Assumption 8.

3 Estimation approach

Based on Theorem 1, the estimation method is described in two steps: first, estimating the ADG and OPDG expectations is discussed in Section 3.1, and then the GMM estimator of the model parameters Γ_{1t}, Γ_{2t}, and B is introduced in Section 3.2.

3.1 Average and outer product of differences of gradients

Under Assumptions 1–7, $\delta_{\mathfrak{T}}$ and $\delta_{\mathfrak{T}\mathfrak{T}}$ in Theorem 1 are both based on the differences of the derivatives of the conditional expectation $\mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}]$ and can be estimated for given time periods $\mathfrak{T} = (t, t-\Delta)$ in the following way. First, the conditional expectation $m_{\mathfrak{T}}(X_{it}, X_{i(t-\Delta)}) = \mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}]$ and its derivatives $\partial\,\mathrm{E}[Y_{it}^\top|X_{it}, X_{i(t-\Delta)}]/\partial X_{it}$ and $\partial\,\mathrm{E}[Y_{it}^\top|X_{it}, X_{i(t-\Delta)}]/\partial X_{i(t-\Delta)}$ are estimated using the local polynomial regression, and then their differences, products, and outer expectations are averaged. Subsequently, the asymptotic distributions of the proposed estimators $\hat{\delta}_{\mathfrak{T}}$ and $\hat{\delta}_{\mathfrak{T}\mathfrak{T}}$ of $\delta_{\mathfrak{T}}$ and $\delta_{\mathfrak{T}\mathfrak{T}}$ are established as well as the corresponding asymptotic results for the generalized method of moments estimator suggested in Theorem 1.

As we perform the local polynomial regression with respect to X_{it} and its lag $X_{i(t-\Delta)}$, let us denote the conditioning variables by $Z_{i\mathfrak{T}} = (X_{it}^\top, X_{i(t-\Delta)}^\top)^\top$ and the non-negative kernel weights by $K_h(u) = K(u/h_n)/h_n^{2d}, u \in \mathbb{R}^{2d}$. For simplicity, the same bandwidth h_n is used for each dimension of $Z_{i\mathfrak{T}}$ (its choice is discussed in Section 4). Let us first consider a single component $Y_{c,it}$, $c \in \{1, \ldots, d_y\}$, of the response vector $Y_{it} = (Y_{1,it}, \ldots, Y_{d_y,it})^\top$ and estimate the expectation $m_{c,\mathfrak{T}}(z) = \mathrm{E}[Y_{c,it}|Z_{i\mathfrak{T}} = z]$ together with its derivatives $\delta_{c,\mathfrak{T},1}(z) = m'_{c,\mathfrak{T},1}(z) = \partial\,\mathrm{E}[Y_{c,it}|Z_{i\mathfrak{T}} = z]/\partial X_{it}$ and $\delta_{c,\mathfrak{T},2}(z) = m'_{c,\mathfrak{T},2}(z) = \partial\,\mathrm{E}[Y_{c,it}|Z_{i\mathfrak{T}} = z]/\partial X_{i(t-\Delta)}$ by the local polynomial regression. If $|\underline{k}| = k_1 + \ldots + k_{2d}$ denotes the length of a vector $\underline{k} = (k_1, \ldots, k_{2d})^\top \in \mathbb{N}_0^{2d}$ and $z^{\underline{k}} = z_1^{k_1} \times \ldots \times z_{2d}^{k_{2d}}$, the local polynomial regression of order p minimizes

$$\sum_{i=1}^{n} \left[Y_{c,it} - \sum_{|\underline{k}|=0}^{p} (Z_{i\mathfrak{T}} - z)^{\underline{k}} b_{c,\underline{k},\mathfrak{T}}(z) \right]^2 K_h\left(Z_{i\mathfrak{T}} - z\right). \tag{6}$$

The estimated parameters $\hat{b}_{c,\mathfrak{T}}(z) = (\hat{b}_{c,\underline{k},\mathfrak{T}}(z))_{|k|=0}^{p}$ contain the estimates of the first-order derivatives of $m_{\mathfrak{T}}(z)$ represented by the $2d$ elements of $\hat{b}_{c,1,\mathfrak{T}}(z) = \{\hat{b}_{c,\underline{k},\mathfrak{T}}(z)\}_{|\underline{k}|=1}$. The vector

$\hat{b}_{c,1,\mathfrak{T}}(z)$ thus estimates $m'_{c,\mathfrak{T}}(z) = (\delta^\top_{c,\mathfrak{T},1}(z), \delta^\top_{c,\mathfrak{T},2}(z))^\top$.

Note that we can write the minimizer $\hat{b}_{c,\mathfrak{T}}(z)$ of (6) in a convenient matrix form since it minimizes the weighted least-squares criterion (6). More specifically,

$$\hat{b}_{c,\mathfrak{T}}(z) = (\hat{b}^\top_{c,0,\mathfrak{T}}(z), \hat{b}^\top_{c,1,\mathfrak{T}}(z), \ldots, \hat{b}^\top_{c,p,\mathfrak{T}}(z))^\top = [Z^\top_{\mathfrak{T}}(z) W_{\mathfrak{T}}(z) Z_{\mathfrak{T}}(z)]^{-1} Z^\top_{\mathfrak{T}}(z) W_{\mathfrak{T}}(z) Y_c, \quad (7)$$

where $Y_c = (Y^\top_{c,1t}, \ldots, Y^\top_{c,nt})^\top$, $Z_{\mathfrak{T}}(z) = \{Z^\top_{i\mathfrak{T}}(z)\}_{i=1}^n$ with $Z_{i\mathfrak{T}}(z) = \{(Z_{i\mathfrak{T}} - z)^{\underline{k}}\}_{|\underline{k}|=0}^p$, and the weight matrix $W_{\mathfrak{T}}(z) = \text{diag}\{K_h(Z_{i\mathfrak{T}} - z)\}_{i=1}^n$.

Furthermore, the difference of derivatives $\delta_{c,\mathfrak{T}}(z) = \delta_{c,\mathfrak{T},1}(z) - \delta_{c,\mathfrak{T},2}(z)$ can be expressed as $\delta_{c,\mathfrak{T}}(z) = \delta_{c,\mathfrak{T},1}(z) - \delta_{c,\mathfrak{T},2}(z) = m'_{c,\mathfrak{T},1}(z) - m'_{c,\mathfrak{T},2}(z) = L_1 m'_{c,\mathfrak{T}}(z)$, where L_1 is a submatrix formed by columns $2, \ldots, d+1$ of matrix $L = (e_2 - e_{d+2}, \ldots, e_{d+1} - e_{2d+1})^\top$ with e_j representing the unit vector such that its jth element is 1, all other elements are 0, and its length equals the length of $\hat{b}_{c,\mathfrak{T}}(z)$ for $j = 2, ..., 2d+1$. Then the local derivative estimator of $\delta_{c,\mathfrak{T}}(z)$ is written as

$$\hat{\delta}_{c,\mathfrak{T}}(z) = L\hat{b}_{c,\mathfrak{T}}(z) = L \cdot [Z^\top_{\mathfrak{T}}(z) W_{\mathfrak{T}}(z) Z_{\mathfrak{T}}(z)]^{-1} Z^\top_{\mathfrak{T}}(z) W_{\mathfrak{T}}(z) Y_c. \quad (8)$$

Finally, to combine the differences of derivatives for different components of the response vector Y_{it}, we denote the matrix of response observations $Y = (Y_1, \ldots, Y_{d_y})^\top$, the conditional expectation $m_{\mathfrak{T}}(z) = \mathrm{E}[Y_{it} | Z_{i\mathfrak{T}} = z] = (m_{1,\mathfrak{T}}(z), \ldots, m_{d_y,\mathfrak{T}}(z))^\top$, and the difference of derivatives $\delta_{\mathfrak{T}}(z) = (\delta_{1,\mathfrak{T}}(z), \ldots, \delta_{d_y,\mathfrak{T}}(z))^\top = (m'_{1,\mathfrak{T},1}(z) - m'_{c,\mathfrak{T},2}(z), \ldots, m'_{d_y,\mathfrak{T},1}(z) - m'_{d_y,\mathfrak{T},2}(z))^\top$. Given (8), this difference of derivatives can be estimated at z by

$$\hat{\delta}_{\mathfrak{T}}(z) = \left(\hat{\delta}_{1,\mathfrak{T}}(z), \ldots, \hat{\delta}_{d_y,\mathfrak{T}}(z)\right)^\top = \left(L \cdot [Z^\top_{\mathfrak{T}}(z) W_{\mathfrak{T}}(z) Z_{\mathfrak{T}}(z)]^{-1} Z^\top_{\mathfrak{T}}(z) W_{\mathfrak{T}}(z) Y\right)^\top. \quad (9)$$

By Theorem 1, B can be identified using $\delta_{\mathfrak{T}} = \mathrm{E}[\delta_{\mathfrak{T}}(Z_{i\mathfrak{T}})]$ and $\delta_{\mathfrak{T}\mathfrak{T}} = \mathrm{E}[\delta_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top \delta_{\mathfrak{T}}(Z_{i\mathfrak{T}})]$, which can be estimated by the corresponding sample averages of (9):

$$\hat{\delta}_{\mathfrak{T}} = \frac{1}{n} \sum_{i=1}^n \hat{\delta}_{\mathfrak{T}}(Z_{i\mathfrak{T}}), \quad (10)$$

$$\hat{\delta}_{\mathfrak{T}\mathfrak{T}} = \frac{1}{n} \sum_{i=1}^n \hat{\delta}^\top_{\mathfrak{T}}(Z_{i\mathfrak{T}}) \hat{\delta}_{\mathfrak{T}}(Z_{i\mathfrak{T}}). \quad (11)$$

The first estimator $\hat{\delta}_{\mathfrak{T}}$ is the multivariate version of the average difference of derivatives

estimator, proposed in the univariate case by Čížek and Lei (2018) and labelled here ADG, and the second estimator $\hat{\delta}_{\mathfrak{T}\mathfrak{T}}$ evaluates the average outer product of differenced gradients (OPDG). Before proceeding to the GMM estimation based on $\hat{\delta}_{\mathfrak{T}}$ and $\hat{\delta}_{\mathfrak{T}\mathfrak{T}}$ in Section 3.2, we establish the consistency and derive the asymptotic distribution of the ADG and OPDG estimators $\hat{\delta}_{\mathfrak{T}}$ and $\hat{\delta}_{\mathfrak{T}\mathfrak{T}}$ based on the local polynomial regression (6) under the following assumptions.

Assumption 9. *1. The bandwidth h_n satisfies $nh_n^{2p+2} \to 0$, $nh_n^{4d+2} \to 0$, $nh_n^{2d+3}/\ln n \to \infty$, and $n^{1-2/\delta_y} h_n^{2d}/[\ln n\{\ln n(\ln\ln n)^{1+\delta_y}\}^{2/\delta_y}] \to \infty$ for $n \to +\infty$.*

2. The kernel function K is a symmetric density function with a compact support, functions $u^{\underline{k}}K(u)$ are Lipschitz for any $\underline{k} \in \mathbb{N}^{2d}, 0 \leq |\underline{k}| \leq 2p+1$, and $\int \|u\|^{4p} K(u)du < \infty$.

3. $Z_{i\mathfrak{T}} = (X_{it}^\top, X_{i(t-\Delta)}^\top)^\top$ has a compact support $\mathcal{D} \subset \mathbb{R}^{2d}$. Additionally, the density function $f_{\mathfrak{T}}$ of $Z_{i\mathfrak{T}}$ exists, satisfies $\inf_{z\in\mathcal{D}} f_{\mathfrak{T}}(z) \geq \epsilon_1$ and $\sup_{z\in\mathcal{D}} f_{\mathfrak{T}}(z) \leq \epsilon_2$ for some $\epsilon_1, \epsilon_2 > 0$, and is twice continuously differentiable with uniformly bounded derivatives. Let $f'_{\mathfrak{T}}$ and $f'_{\mathfrak{T};j}$ denote the first and jth partial derivative of $f_{\mathfrak{T}}$, respectively, $j = 1, \ldots, 2d$.

4. For $\mathfrak{T} = (t, t-\Delta)$, $m_{\mathfrak{T}}(z) = \mathrm{E}[Y_{it}|Z_{i\mathfrak{T}} = z]$ is $(p+1)$-times differentiable with uniformly bounded and Lipschitz partial derivatives on $\mathcal{D} \subset \mathbb{R}^{2d}$. Additionally, $m_{\mathfrak{T}}(Z_{i\mathfrak{T}})$ and its $(p+1)$ derivatives as functions of $Z_{i\mathfrak{T}}$ have finite second moments.

5. Errors $V_{i\mathfrak{T}} = Y_{it} - \mathrm{E}(Y_{it}|Z_{i\mathfrak{T}}) = Y_{it} - m_{\mathfrak{T}}(Z_{i\mathfrak{T}})$ have finite fourth moments. For pairs of time periods $\mathfrak{T} = (t, t-\Delta)$ and $\mathfrak{S} = (s, s-\Delta')$, let $\Sigma_{\mathfrak{T}\mathfrak{T}}(z) = \mathrm{E}(V_{i\mathfrak{T}}V_{i\mathfrak{T}}^\top|Z_{i\mathfrak{T}} = z)$ and $\Sigma_{\mathfrak{T}\mathfrak{S}}(z_1, z_2) = \mathrm{E}(V_{i\mathfrak{T}}V_{i\mathfrak{S}}^\top|Z_{i\mathfrak{T}} = z_1, Z_{i\mathfrak{S}} = z_2)$ be continuous in z and (z_1, z_2), respectively.

6. The conditional distributions of $Y_{it}|Z_{i\mathfrak{T}}$ and of $Z_{i\mathfrak{T}}|Y_{it}$ are continuous and have bounded densities.

Assumption 9 introduces typical assumptions on the bandwidth h_n, the kernel function K, and data generating process for the average derivative estimators (cf., Li et al., 2003). This includes the existence of $p+1$ derivatives in Assumption 9.4 ($p \geq d+1$), which is usually imposed for average derivative estimation (e.g., Cattaneo et al., 2013, Čížek and Lei, 2018, and Härdle and Stoker, 1989). It can be relaxed via the iterative procedure of Hristache et al. (2001b) developed for multiple-index models. Specifically, they show that their iterative procedure achieves $\sqrt{n}$-consistency under the mild assumption that the second derivatives exist and are bounded regardless of the dimension d if the number R of the indices is less than 4.

Next, we establish the asymptotic distribution of the average derivative $\hat{\delta}_{\mathfrak{T}}$ and outer product $\hat{\delta}_{\mathfrak{T}\mathfrak{T}}$ under the stated assumptions.

Theorem 2. *Define the vector* $MM_{\mathfrak{T}}(Z_{i\mathfrak{T}}) = vec(L_1 m'_{\mathfrak{T}}(Z_{i\mathfrak{T}}), L_1 m'_{\mathfrak{T}}(Z_{i\mathfrak{T}}) m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top L_1^\top)$. *Under Assumptions 1–7 and 9, the average derivative* $\hat{\delta}_{\mathfrak{T}}$ *and outer product* $\hat{\delta}_{\mathfrak{T}\mathfrak{T}}$ *estimators defined in equations* (10)–(11) *for one given* $\mathfrak{T}$ *are consistent and jointly asymptotically normal:*

$$\sqrt{n}\left(vec(\hat{\delta}_{\mathfrak{T}}, \hat{\delta}_{\mathfrak{T}\mathfrak{T}}) - h_n^p vec(LA^1_{\mathfrak{T}}, L[A^2_{\mathfrak{T}\mathfrak{T}} + A^{2\top}_{\mathfrak{T}\mathfrak{T}}]L^\top) - \mathrm{E}\, MM_{\mathfrak{T}}(Z_{i\mathfrak{T}})\right) \to N\left(0, \Omega_{\mathfrak{T}} + \Phi_{\mathfrak{T}}\right)$$

in distribution as $n \to +\infty$, *where* $A^1_{\mathfrak{T}}$ *and* $A^2_{\mathfrak{T}\mathfrak{T}}$ *are defined in* (A.16) *and* $\Omega_{\mathfrak{T}} + \Phi_{\mathfrak{T}}$ *is positive semidefinite and defined by* $\Omega_{\mathfrak{T}} = \mathrm{Var}\left[MM_{\mathfrak{T}}(Z_{i\mathfrak{T}})\right]$ *and* $\Phi_{\mathfrak{T}} = \mathrm{E}\left[GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})\Sigma_{\mathfrak{T}\mathfrak{T}}(Z_{i\mathfrak{T}})GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top\right]$. *Further,* $GM_{\mathfrak{T}}(z) = (-[I_{d_y} \otimes \{LG_{\mathfrak{T};\cdot 1}(z)\}]^\top, \iota_{d_y} vec[-2L_1 m''_{\mathfrak{T}}(Z_{i\mathfrak{T}})L_1^\top - LG_{\mathfrak{T};\cdot 1}(Z_{i\mathfrak{T}})m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top L_1^\top - L_1 m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})\, G_{\mathfrak{T};\cdot 1}(Z_{i\mathfrak{T}})^\top L^\top]^\top)^\top$, *where* $G_{\mathfrak{T};\cdot 1}(z) = G_{\mathfrak{T}}(z)e_1$ *is the first column of matrix* $G_{\mathfrak{T}}(z) = [Mf_{\mathfrak{T}}(z)]^{-1}\,[\sum_{j=1}^{2d} f'_{\mathfrak{T};j}(z)Q_s]$, I_{d_y} *represents the* $d_y \times d_y$ *identity matrix,* ι_{d_y} *is the* $d_y \times 1$ *vector of ones, and the matrices of kernel weights* M, B, *and* Q_s *are defined in Appendix B.*

Theorem 2 establishes that the ADG $\hat{\delta}_{\mathfrak{T}}$ and OPDG $\hat{\delta}_{\mathfrak{T}\mathfrak{T}}$ estimators are consistent and jointly asymptotically normal for any given $\mathfrak{T}$. In the univariate setting, the result for ADG entails Theorem 2 of Čížek and Lei (2018), but we see that the asymptotic distribution of OPDG clearly differs from ADG, for example, by its dependence on the second derivative of $m_{\mathfrak{T}}$. Similar to other average derivative estimators, their $\sqrt{n}$-consistency is a consequence of taking the sample average of n nonparametric estimates at each $Z_{i\mathfrak{T}}$, see (10)–(11), as long as the asymptotic bias terms $h_n^p LA^1_{\mathfrak{T}}$ and $h_n^p L[A^2_{\mathfrak{T}\mathfrak{T}} + A^{2\top}_{\mathfrak{T}\mathfrak{T}}]L^\top$ are negligible relative to $n^{-1/2}$. As seen in Theorem 2, the bias terms $h_n^p LA^1_{\mathfrak{T}}$ and $h_n^p L[A^2_{\mathfrak{T}\mathfrak{T}} + A^{2\top}_{\mathfrak{T}\mathfrak{T}}]L^\top$ can be eliminated if $\sqrt{n}h_n^p \to 0$ by employing local polynomial estimation with higher order polynomials, by using an undersmoothing bandwidth, or by the generalized jackknife procedure described in Supplementary Appendix F.

If data have $T > 2$ periods, there are multiple pairs $\mathfrak{T} = (t, t-\Delta)$ of time periods as $\Delta \in \{1, \ldots, T-1\}$ and $t \in \{\Delta+1, \ldots, T\}$. Although Theorem 2 can be applied to any pair of time periods $\mathfrak{T}$, the moment conditions suggested in Theorem 1 can be constructed for multiple pairs $\mathfrak{T} \in \mathscr{S}$ of time periods, where $\mathscr{S}$ denotes the set of employed pairs $(t, t-\Delta)$. They can be all jointly incorporated in a GMM criterion (see Section 3.2) with the aim of improving accuracy of estimation by adding extra moment conditions. To facilitate such estimation, we provide here their joint distribution for a given number $|\mathscr{S}|$ and set $\mathscr{S}$ of time-period pairs.

Theorem 3. *Let Assumptions 1–7 and 9 hold for every $\mathfrak{T} \in \mathscr{S}$ and let us define the vectors $MM_{\mathfrak{T}}(Z_{i\mathfrak{T}}) = vec(L_1 m'_{\mathfrak{T}}(Z_{i\mathfrak{T}}), L_1 m'_{\mathfrak{T}}(Z_{i\mathfrak{T}}) m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top L_1^\top)$ and $MM(Z_{i\cdot}) = \{MM_{\mathfrak{T}}(Z_{i\mathfrak{T}})\}_{\mathfrak{T}\in\mathscr{S}}$, where $Z_{i\cdot} = \{Z_{i\mathfrak{T}}\}_{\mathfrak{T}\in\mathscr{S}}$. Denoting $\hat{\delta} = \{vec(\hat{\delta}_{\mathfrak{T}}, \hat{\delta}_{\mathfrak{T}\mathfrak{T}})\}_{\mathfrak{T}\in\mathscr{S}}$ and $Bias = \{vec(LA^1_{\mathfrak{T}}, L[A^2_{\mathfrak{T}\mathfrak{T}} + A^{2\top}_{\mathfrak{T}\mathfrak{T}}]L^\top)\}_{\mathfrak{T}\in\mathscr{S}}$, where $A^1_{\mathfrak{T}}$ and $A^2_{\mathfrak{T}\mathfrak{T}}$ are defined in* (A.16)*, $\hat{\delta}$ is asymptotically normal:*

$$\sqrt{n}\left(\hat{\delta} - h_n^p Bias - \mathrm{E}\, MM(Z_{it,\cdot})\right) \to N(0, \Omega + \Phi)$$

in distribution as $n \to +\infty$, where $\Omega + \Phi$ is positive semidefinite and Ω and Φ matrices consist of $|\mathscr{S}| \times |\mathscr{S}|$ blocks of dimensions $(d+d^2)\times(d+d^2)$; the blocks with coordinates $(\mathfrak{T}, \mathfrak{S}) \in \mathscr{S}\times\mathscr{S}$ within matrices Ω and Φ have the forms

$$\Omega^{(\mathfrak{T},\mathfrak{S})} = \mathrm{Cov}\left[MM_{\mathfrak{T}}(Z_{i\mathfrak{T}}), MM_{\mathfrak{T}}(Z_{i\mathfrak{S}})\right],$$

$$\Phi^{(\mathfrak{T},\mathfrak{S})} = \mathrm{E}\left[GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})\Sigma_{\mathfrak{T}\mathfrak{S}}(Z_{i\mathfrak{T}}, Z_{i\mathfrak{S}})GM_{\mathfrak{S}}(Z_{i\mathfrak{S}})^\top\right].$$

Further $GM_{\mathfrak{T}}(z) = (-[I_{d_y}\otimes\{LG_{\mathfrak{T};\cdot 1}(z)\}]^\top, \iota_{d_y} vec[-2L_1 m''_{\mathfrak{T}}(Z_{i\mathfrak{T}})L_1^\top - LG_{\mathfrak{T};\cdot 1}(Z_{i\mathfrak{T}})m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top L_1^\top - L_1 m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})G_{\mathfrak{T};\cdot 1}(Z_{i\mathfrak{T}})^\top L^\top]^\top)^\top$, where $G_{\mathfrak{T};\cdot 1}(z) = G_{\mathfrak{T}}(z)e_1$ is the first column of matrix $G_{\mathfrak{T}}(z) = [Mf_{\mathfrak{T}}(z)]^{-1}[\sum_{j=1}^{2d} f'_{\mathfrak{T};j}(z)Q_s]$, I_{d_y} is the $d_y \times d_y$ identity matrix, ι_{d_y} is the $d_y \times 1$ vector of ones, and the matrices of kernel weights M and Q_s are defined in Appendix B.

Finally, note that the asymptotic variance matrices in Theorems 2 and 3 depend on the first and second derivatives of the regression function $m_{\mathfrak{T}}(Z_{i\mathfrak{T}})$ similarly to Samarov (1993), which are obtained during estimation as a product of the local polynomial estimation. After estimating $\hat{\delta}$, the only quantity that needs to be estimated to compute the asymptotic variance is thus the density function $f_{\mathfrak{T}}(z)$ and its derivatives since all other quantities L, M, Q_s are known and fully determined by the kernel used in estimation (see Appendix B).

3.2 GMM

In Theorem 1, we showed that the true parameter values θ_0 of the unconstrained parameters θ of matrices $(B, \Gamma_{1t}, \Gamma_{2t})$ can be identified by minimizing $g_{\mathfrak{T}\mathfrak{T}}(\theta)^\top g_{\mathfrak{T}\mathfrak{T}}(\theta)$, where $g_{\mathfrak{T}\mathfrak{T}}(\theta) = \left(g^1_{\mathfrak{T}}(\theta)^\top, g^2_{\mathfrak{T}\mathfrak{T}}(\theta)^\top\right)^\top$, $g^1_{\mathfrak{T}}(\theta) = \mathrm{vec}(\delta_{\mathfrak{T}} - \Gamma_{1t}B^\top)$, and $g^2_{\mathfrak{T}\mathfrak{T}}(\theta) = \mathrm{vec}(\delta_{\mathfrak{T}\mathfrak{T}} - B\Gamma_{2t}B^\top)$. Given the estimators $\hat{\delta}_{\mathfrak{T}}$ and $\hat{\delta}_{\mathfrak{T}\mathfrak{T}}$ introduced in (10)–(11), the moment conditions $g_{\mathfrak{T}\mathfrak{T}}(\theta) = 0$ have natural sample analogs $\hat{g}_{\mathfrak{T}\mathfrak{T}}(\theta) = \left(\hat{g}^1_{\mathfrak{T}}(\theta)^\top, \hat{g}^2_{\mathfrak{T}\mathfrak{T}}(\theta)^\top\right)^\top = 0$, where $\hat{g}^1_{\mathfrak{T}}(\theta) = \mathrm{vec}(\hat{\delta}_{\mathfrak{T}} - \Gamma_{1t}B^\top)$ and $\hat{g}^2_{\mathfrak{T}\mathfrak{T}}(\theta) =$

$\text{vec}(\hat{\delta}_{\mathfrak{T}\mathfrak{T}} - B\Gamma_{2t}B^\top)$. An initial GMM estimate of θ can thus be obtained by

$$\hat{\theta}_n^0 = \text{argmin}_\theta \, \hat{g}_{\mathfrak{T}\mathfrak{T}}(\theta)^\top \hat{g}_{\mathfrak{T}\mathfrak{T}}(\theta), \tag{12}$$

or, if multiple pairs $\mathfrak{T} = (t, t-\Delta)$ from the set $\mathscr{S}$ of time-period pairs are used, by

$$\hat{\theta}_n^0 = \text{argmin}_\theta \, \hat{g}(\theta)^\top \hat{g}(\theta), \tag{13}$$

where $\hat{g}(\theta) = \{\hat{g}_{\mathfrak{T}\mathfrak{T}}(\theta)\}_{\mathfrak{T}\in\mathscr{S}}$ and $g(\theta) = \{g_{\mathfrak{T}\mathfrak{T}}(\theta)\}_{\mathfrak{T}\in\mathscr{S}}$. This initial estimator can be further improved by applying weights to the moment conditions depending on their precision or desirability in the estimation. Given a square matrix W_n of dimension $|\mathscr{S}|(d+d^2)$, the weighted GMM estimator can be defined by

$$\hat{\theta}_n = \text{argmin}_\theta \, \hat{g}(\theta)^\top W_n \hat{g}(\theta), \tag{14}$$

where the initial choice corresponds to $W_n = I_{|\mathscr{S}|(d+d^2)}$ and the second step weighting matrix W_n would typically be based on the inverse of the variance matrix of the moment conditions $\hat{g}(\theta)$, which equals the variance matrix $\Omega + \Phi$ of $\hat{\delta} = \{\text{vec}(\hat{\delta}_{\mathfrak{T}}, \hat{\delta}_{\mathfrak{T}\mathfrak{T}})\}_{\mathfrak{T}\in\mathscr{S}}$ derived in Theorem 3. Note that this variance matrix does not have the full rank and it is thus necessary to replace the standard inverse by a consistent estimator as discussed in Donkers and Schafgans (2008).

The proposed GMM estimators (12)–(14) will be shortly labelled as GMM-ADG-OPDG, where ADG and OPDG refer to the moment conditions based on $g^1_{\mathfrak{T}}(\theta)$ and $g^2_{\mathfrak{T}\mathfrak{T}}(\theta)$, respectively. To derive their asymptotic distributions, additional assumptions have to be introduced regarding the properties of the parameters and the weighting matrix.

Assumption 10. *The weighting matrix W_n is such that $W_n \to W$ in probability for $n \to \infty$, where W is positive semidefinite.*

Assumption 11. *The true value parameter θ_0 minimizing $g(\theta)^\top W g(\theta)$ is in the interior of the compact parameter space Θ. Moreover, the matrix $\Pi^\top W \Pi$ has full rank, where $\Pi = \partial g(\theta_0)/\partial\theta^\top$.*

Since we cannot impose the positive definiteness of the weighting matrix W as usual, the full rank Assumption 11 has to be explicitly imposed to prevent an invalid weighting matrix (e.g., $W = 0$ or W selecting a single moment condition). For the weighting matrix equal to the

identity matrix, it boils down to a full-rank assumption on Π, which had been already established in Lemma 3 in Donkers and Schafgans (2008). For the choice of the efficient weighting matrix W_{eff}, it can be shown that $\Pi^\top W_{eff}\Pi$ has full rank analogously to Lemma 4 in Donkers and Schafgans (2008).

Theorem 4. *Using the notation of Theorem 3, it holds under Assumptions 1–11 that*

$$\sqrt{n}(\hat{\theta}_n - \theta_0 - h_n^p \left(\Pi^\top W \Pi\right)^{-1} \Pi^\top W Bias) \to N\left(0, \left(\Pi^\top W \Pi\right)^{-1} \Pi^\top W(\Omega + \Phi) W \Pi \left(\Pi^\top W \Pi\right)^{-1}\right)$$

in distribution as $n \to \infty$.

3.3 Dimension selection

The proposed estimation procedure assumes that the number of indices required to model the responses is known, at least to some extent. This assumption is valid in many applications, which rely on a more specific model than (4). For example, the heteroscedastic binary-choice model $Y_{it} = \mathbb{1}\{X_{it}^\top \beta_1 + \alpha_{1i} + \sigma(X_{it}^\top \beta_2 + \alpha_{2i})U_{it} > 0\}$ could be assumed to require two indices $X_{it}^\top \beta_1$ and $X_{it}^\top \beta_2$. Although such an assumption is not necessarily correct, it provides a baseline model and one can then test whether the heteroskedasticity is present (i.e., whether a single index $X_{it}^\top \beta_1$ is sufficient) or whether the heteroskedasticity has a more general structure (i.e., more than two indices are needed to describe the data generating process) by the standard likelihood ratio and Lagrange multiplier tests for GMM.

On the other hand, if there is no prior knowledge about the number of indices required, a sequential test for determining the dimension of the parameter space B is required. It follows from Theorem 1 that the correct dimension of B can be determined by finding the rank of the matrix $\delta_{\mathfrak{T}\mathfrak{T}}$, which is defined and can be estimated independently of B, Γ_{1t}, and Γ_{2t}. Hence, the rank of $\delta_{\mathfrak{T}\mathfrak{T}}$ and thus the rank R of B can be determined prior to the GMM estimation based on the estimator $\hat{\delta}_{\mathfrak{T}\mathfrak{T}}$ proposed in Section 3.1 and the matrix rank estimator proposed by Chen and Fang (2019, Appendix C). For the consistent estimate $\hat{R}$, the GMM estimation procedure in Section 3.2 can be then applied, and given that the considered number of indices is finite ($R < d$), the asymptotic distribution in Theorem 4 applies.

4 Simulation study

In this section, we document the finite sample performance of the GMM estimators based on the moment conditions proposed in Theorem 1 for various panel models with correlated random effects. Therefore, we consider the data generating processes which are characterized by individual effects and explanatory variables having the same joint distribution across cross-sectional units with nonzero correlation between individual effects and explanatory variables. We compare the GMM-ADG-OPDG estimates with the existing estimators for each panel data model. General simulation and implementation details are introduced in Section 4.1, and specific models and the corresponding results are presented later in Sections 4.2–4.4.

4.1 Implementation

In the following sections, we present the results for the baseline GMM estimator (12), that is, the GMM estimator based on the ADG and OPDG moment conditions defined in Theorem 1 for $T = 2$ time periods, the first-order differences $\Delta = 1$, and the identity weighting matrix. The results for more time periods $T > 2$ and higher-order differences $\Delta > 1$ are in Supplementary Appendix D. To obtain the GMM estimates (12), the following procedure is used:

- The ADG and OPDG estimates $\hat{\delta}_{\mathfrak{T}}$ and $\hat{\delta}_{\mathfrak{T}\mathfrak{T}}$ defined in (10)–(11) are obtained by the local quadratic regression estimation with the Gaussian product kernel and the bandwidth is chosen by the leave-one-out cross-validation (a common bandwidth is used for all variables). The second-order polynomial regression is chosen to facilitate and demonstrate the easy applicability of the method. Although this can lead to a non-negligible asymptotic bias, we eliminate it by a generalized jackknife with the bandwidth multiples $1.3, 1.6, 1.9$, which can be used as a bias-correction procedure as described in Supplementary Appendix F. Additionally, we show in finite-sample simulations that the bias is practically negligible in various nonlinear models and there is often no need for the bias-correction procedure.

- Given the ADG and OPDG estimates $\hat{\delta}_{\mathfrak{T}}$ and $\hat{\delta}_{\mathfrak{T}\mathfrak{T}}$ and the normalization and dimension of B, we obtain an initial value of parameter matrix B from $\hat{\delta}_{\mathfrak{T}\mathfrak{T}}$ by imposing the normalization and exclusion constraints of a given model. For this initial values of B, the initial values of matrices Γ_{1t} and Γ_{2t} are chosen so that $\hat{\delta}_{\mathfrak{T}} = \Gamma_{1t}B$ and $\hat{\delta}_{\mathfrak{T}\mathfrak{T}} = B^{\top}\Gamma_{2t}B$

hold and all these initial values are used as a starting point for the standard Newton-type algorithm sequentially minimizing the GMM criterion (12).

To assess the performance of the described estimator in finite-samples, we generate $S = 1000$ panel data samples with the number of individuals $n = 1000$ and the number of time periods $T = 2$. The models contain three regressors X_{1it}, X_{2it}, and X_{3it}, which are drawn independently from the normal distribution $N(0, 1)$ trimmed at -3 and 3. The individual effects $\alpha_i = (\alpha_{1i}, \alpha_{2i})^\top$ are then created as a weighted average of randomly distributed errors independent of covariates X_{1it}, X_{2it}, and X_{3it} and time averages $\bar{X}_{1i} = T^{-1} \sum_{t=1}^{T} X_{1it}$ and $\bar{X}_{2i} = T^{-1} \sum_{t=1}^{T} X_{2it}$ in the cases of α_{1i} and α_{2i}, respectively. Given these individual effects and covariates $X_{it} = (X_{1it}, X_{2it}, X_{3it})^\top$, the response variables depend on two linear combinations $\alpha_{1i} + X_{it}^\top \beta_1$ and $\alpha_{2i} + X_{it}^\top \beta_2$; specific functional dependence is going to be specified for particular regression models in the respective sections. For each model, the parameter estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ obtained by the proposed as well as existing methods are always normalized in the same way defined by the identification assumptions of each model: for example, the first coefficient $\hat{\beta}_{11}$ of $\hat{\beta}_1$ is normalized to 1. After normalization, the bias and root mean squared errors (RMSE) are computed for each parameter and reported for every considered estimation method. Further, we report the average asymptotic standard errors (ASE) based on Theorem 4 (the estimation approach of Härdle and Stoker, 1989, is used) and the empirical coverage rates (COV) for the 95% confidence intervals based on the asymptotic distribution of each estimator. For all experiments, we also estimate the total AME (TME) $\delta_{\mathfrak{T}}$ of X_{it} on the response and the AME $\Gamma_{1t,1}\beta_1$ and $\Gamma_{1t,2}\beta_2$ of X_{it} specific to linear combinations $X_{it}^\top \beta_1$ and $X_{it}^\top \beta_2$, respectively, which are labelled SME1 and SME2; see Section 2 for details. In all cases, we report the marginal effects averaged across the simulated samples and their RMSE; the biases of the marginal effects are always very small and thus not reported.

In the following sections, we will analyze the estimation of the coefficients and AME in the context of the binary partially linear single-index model (Section 4.2), in the heteroskedastic censored regression model (Section 4.3), and in the sample selection model (Section 4.4).

4.2 Binary-choice model

Let us first explore the performance of the proposed method in the binary partially-linear single-index model in the case of a logistic panel data regression; such a model in the cross-sectional

Table 1: The bias, RMSE, asymptotic standard errors (ASE), coverage rates (COV) in the upper half, and marginal effect averages (total TME and SME1 and SME2 specific to indices $x_{it}^\top\beta_1$ and $x_{it}^\top\beta_2$) and RMSE for $P(Y_{it}=1|X_{it},\alpha_{1i},\alpha_{2i})$ in the lower half of all estimators in the binary partially linear single-index model for the sample size $n=1000$ and $T=2$.

	β_{13}				β_{23}			
	Bias	RMSE	ASE	COV	Bias	RMSE	ASE	COV
FE Logit	0.060	0.135	0.113	0.876				
GMM-OPDG	-0.004	0.172	0.174	0.951	0.012	0.210	0.204	0.949
Jackknife	-0.009	0.194	0.186	0.932	0.019	0.208	0.183	0.930
GMM-ADG-OPDG	-0.006	0.174	0.168	0.937	0.012	0.210	0.202	0.947
Jackknife	-0.014	0.205	0.185	0.930	0.019	0.209	0.181	0.930
SMS	-0.079	0.268	0.153	0.826				
$P(Y_{it}=1\|X_{it},\alpha_i)$	TME:	RMSE	CFN:	RMSE	SME1:	RMSE	SME2:	RMSE
X_{1it}	0.168	0.017	0.168	0.016	0.168	0.017	—	—
X_{2it}	0.020	0.017	0.020	0.018	—	—	0.020	0.002
X_{3it}	-0.158	0.011	-0.158	0.018	-0.168	0.020	0.010	0.001

setting was studied by Carroll et al. (1997), for instance. More specifically, the simulated binary-choice model is defined by $Y_{it} = \mathbb{1}\{X_{it}^\top\beta_1 + \alpha_{1i} + g(X_{it}^\top\beta_2, \alpha_{2i}) + U_{it} > 0\}$, where $g(t_1,t_2) = (1-0.75t_1^2-0.25t_2^2)^2/2$. The coefficients are $\beta_1 = (1,0,-1)^\top$ and $\beta_2 = (0,1,0.5)^\top$, the individual effects are defined by $\alpha_{1i} = \bar{X}_{1i} + \epsilon_{1i}$, $\epsilon_{1i} \sim U(-1,1)$, and $\alpha_{2i} = \bar{X}_{2i} + \epsilon_{2i}$, $\epsilon_{2i} \sim U(-1,1)$, respectively, and the error term follows the logistic distribution, $U_{it} \sim \Lambda(0, \sqrt{3}/\pi)$. In this model, there is no exclusion restriction imposed during the estimation. Hence, the parameters are normalized so that $\beta_{11}=1$, $\beta_{12}=0$, $\beta_{21}=0$, and $\beta_{22}=1$.

Given the lack of existing methods, the proposed estimator is compared to the fixed-effect estimators of the standard binary-choice models: the FE logit and smoothed maximum score (SMS) estimator of Charlier et al. (1995) are used to provide a benchmark for the magnitude of RMSE. Although there are no other estimators of SME1 and SME2, TME in this specific model can be also estimated by the method of Chernozhukov et al. (2019) labelled CFN here. All results for $T=2$ are summarized in Table 1 (see Table A.5 for $T>2$). The results in Table 1 confirm that both the FE logit and SMS exhibit a bias and cannot estimate the coefficient β_{23}, whereas the proposed GMM(-ADG)-OPDG estimators provide relatively precise estimates with negligible finite-sample biases. There is thus no need for bias correction and the reported GMM(-ADG)-OPDG estimates and their jackknife bias corrections lead to rather similar results.

Furthermore, the ASEs are relatively close to the finite-sample RMSEs, and consequently,

Table 2: The bias, RMSE, asymptotic standard errors (ASE), coverage rates (COV) in the upper half, and marginal effect averages (total TME and SME1 and SME2 specific to indices $x_{it}^\top\beta_1$ and $x_{it}^\top\beta_2$) and RMSE for $E(Y_{it}|X_{it},\alpha_{1i},\alpha_{2i})$ in the lower half of all estimators in the heteroskedastic censored regression model for the sample size $n = 1000$ and $T = 2$.

	β_{13}				β_{23}			
	Bias	RMSE	ASE	COV	Bias	RMSE	ASE	COV
Pooled Tobit	1.125	1.137	1.664	0.852				
TLS	0.746	13.535	0.838	0.092				
GMM-OPDG	-0.023	0.256	0.330	0.986	0.053	0.287	0.407	0.989
Jackknife	-0.027	0.289	0.322	0.968	0.053	0.292	0.392	0.975
GMM-ADG-OPDG	-0.023	0.255	0.309	0.979	0.053	0.287	0.407	0.989
Jackknife	-0.027	0.282	0.308	0.963	0.053	0.291	0.392	0.972
SMS	-0.090	0.368	0.213	0.802				
$E(Y_{it}\|X_{it},\alpha_{1i},\alpha_{2i})$	TME:	RMSE	TLSME:	RMSE	SME1:	RMSE	SME2:	RMSE
X_{1it}	0.657	0.106	-0.002	0.122	0.658	0.106	—	—
X_{2it}	0.029	0.157	0.676	0.101	—	—	0.029	0.157
X_{3it}	-0.640	0.123	0.022	0.113	-0.656	0.121	0.017	0.093

the empirical coverage rates for all coefficients are smaller, but generally close to the 95% nominal level. Although all empirical results are similar for GMM-OPDG and GMM-ADG-OPDG and the ADG moment conditions are not particularly useful for the precision of estimation, they will be always included as they directly provide estimates of TME, for example, those in the lower half of Table 1. Comparing TME with those by CFN, we see these total AME and their RMSEs are practically equal. The GMM-ADG-OPDG however also allows identification of the AME specific to each index $X_{it}^\top\beta_1$ and $X_{it}^\top\beta_2$: see SME1 and SME2 in Table 1, which are all precisely estimated as documented by the corresponding RMSEs.

4.3 Censored regression model

We now study the performance of the proposed method in the heteroskedastic censored regression model. More specifically, the simulated censored model is defined by $Y_{it} = \max\{0, X_{it}^\top\beta_1 + \alpha_{1i} + g(X_{it}^\top\beta_2, \alpha_{2i})U_{it}\}$, where $g(t_1, t_2) = 1.5t_1^2$. The coefficients are $\beta_1 = (1, 0, -1)^\top$ and $\beta_2 = (0, 1, 0.5)^\top$, the individual effects are defined by $\alpha_{1i} = \bar{X}_{1i} + \epsilon_{1i}$, $\epsilon_{1i} \sim U(-1, 1)$, and $\alpha_{2i} = \bar{X}_{2i} + \epsilon_{2i}$, $\epsilon_{2i} \sim U(-1, 1)$, respectively, and the error term follows the standard normal distribution, $U_{it} \sim N(0, 1)$. As there are no exclusion restrictions imposed during the estimation, the parameter normalization $\beta_{11} = 1$, $\beta_{12} = 0$, $\beta_{21} = 0$, and $\beta_{22} = 1$ is used.

Given that the heteroskedasticity is a function of $X_{it}^\top \beta_2$ and thus varies over time, the standard censored regression estimators such as the pooled Tobit or trimmed least squares (TLS) and trimmed least absolute deviation (TLAD) of Honore (1992) are inconsistent. As the TLAD estimates cannot be reliably computed in every sample, but are rather close to TLS, we report the pooled Tobit and the TLS estimates with their marginal effects (TLSME). To provide a consistent benchmark estimator, we report also the binary-choice SMS (Charlier et al., 1995) for responses $\mathbb{1}(Y_{it} > 0)$, which is applicable under a general form of heteroskedasticity.

The results for $T = 2$ are summarized in Table 2 (see Table A.6 for $T > 2$); the slope coefficient β_{13} is reported for all methods, whereas the coefficient β_{23} determining the conditional variance can be estimated and is thus reported only for the proposed GMM(-ADG)-OPDG estimators. The results in Table 2 confirm that the pooled Tobit and TLS exhibit a large bias, and in the latter case, also a large RMSE. The reported SMS estimator is characterized by only a small bias and reasonable RMSE, but it is outperformed by GMM(-ADG)-OPDG, which are characterized by smaller biases, smaller RMSE, and the ability to estimate both the indices $X_{it}^\top \beta_1$ and $X_{it}^\top \beta_2$. This is important to obtain the marginal effects, which cannot be consistently by the existing methods: both SME1 and SME2 corresponding to the marginal effects due to $X_{it}^\top \beta_1$ and $X_{it}^\top \beta_2$, respectively, and TME are estimated with negligible biases, which are therefore not reported, and small standard errors. In contrast, TLSME provides incorrect MEs as the only significant ME corresponds to variable X_{2it} with zero coefficient β_{13}. Finally, let us note that the asymptotic standard errors and coverages are higher than expected since we use for simplicity the same bandwidth for the estimation of the regression parameters and the standard errors. Given the shape of the conditional variance, the estimated asymptotic standard errors would be closer to the simulated ones if a (smaller) bandwidth chosen specifically for the variance estimation was used.

4.4 Sample selection model

Finally, we consider a linear sample-selection model, noting that the performance of the ADG estimator in the standard linear regression model is close to the first-difference least squares (Čížek and Lei, 2018). Although the proposed GMM estimator can handle also more complicated models, for example, including interactions between the individual effects and the linear indices, the simple linear structure allows us to compare the results with the existing sam-

Table 3: The bias, RMSE, asymptotic standard errors (ASE), coverage rates (COV) in the upper half, and marginal effect averages (total TME and SME1 and SME2 specific to indices $x_{it}^\top\beta_1$ and $x_{it}^\top\beta_2$) and RMSE for $P(Y_{it}=1|X_{it},\alpha_{1i})$ and $E(Y_{2it}|X_{it},\alpha_{1i},\alpha_{2i})$ in the lower half of all estimators in the sample-selection model for the sample size $n=1000$ and $T=2$.

	β_{12}			β_{13}			β_{23}		
	RMSE	ASE	COV	RMSE	ASE	COV	RMSE	ASE	COV
GMM-ADG-OPDG	0.109	0.119	0.973	0.157	0.154	0.957	0.102	0.105	0.959
Jackknife	0.111	0.120	0.973	0.165	0.153	0.957	0.110	0.105	0.959
FE-Logit+KYR	0.123	0.132	0.966	0.121	0.132	0.966	0.115	0.110	0.941
$P(Y_{1it}=1\|X_{it},\alpha_{1i})$	TME:	RMSE	CFN:	RMSE		SME1:	RMSE	SME2:	RMSE
X_{1it}	0.180	0.014	0.179	0.014		0.179	0.015	—	—
X_{2it}	0.180	0.013	0.179	0.015		0.179	0.015	—	—
X_{3it}	0.179	0.014	0.179	0.014		0.180	0.018	—	—
$E(Y_{2it}\|X_{it},\alpha_{1i},\alpha_{2i})$	TME:	RMSE	SME1:	RMSE		KYR:	RMSE	SME2:	RMSE
X_{1it}	-0.122	0.083	-0.123	0.083		—	—	—	—
X_{2it}	0.872	0.085	-0.123	0.084		0.999	0.121	0.994	0.122
X_{3it}	0.375	0.083	-0.124	0.087		0.499	0.124	0.500	0.118

ple selection estimators: $Y_{2it} = X_{it}^\top\beta_2 + \alpha_{2i} + U_{2it}$ is observed when the selection variable $Y_{1it} = \mathbb{1}\{X_{it}^\top\beta_1 + \alpha_{1i} + U_{1it} > 0\}$ equals 1. The coefficients $\beta_1 = (1,1,1)^\top$ and $\beta_2 = (0,1,0.5)^\top$. The individual effects $\alpha_{1i} = 0.5 + 0.5\bar{X}_{1i} + \epsilon_{1i}$, $\epsilon_{1i} \sim U(0,1)$, and $\alpha_{2i} = \bar{X}_{2i} + \epsilon_{2i}$, $\epsilon_{2i} \sim N(0,1)$, respectively, and errors follow the logistic and Gaussian distributions $U_{1it} \sim \Lambda(0, \sqrt{3}/(2\pi))$ and $U_{2it} \sim N(0, 0.8)$ with their mutual correlation equal to 0.75, which leads to approximately 68% selected observations. The coefficient identification is obtained by the normalization of $\beta_{11} = 1$ and $\beta_{22} = 1$ and the exclusion restriction that X_{1it} does not enter the mean equation for Y_{2it}.

The proposed GMM-ADG-OPDG estimator is compared to the sample-selection estimator of Kyriazidou (1997) based on the fixed-effect logit (FE-Logit) modelling of the selection equation; note that the logit is correctly specified as U_{1it} follows the logistic distribution and that the auxiliary parameters of this sample-selection estimator are chosen as in Kyriazidou (1997, Section 4). The results are summarized in Table 3 (the finite-sample bias of all estimators is negligible, see Table A.7) and there is thus no need for bias correction (the GMM-ADG-OPDG estimates and their jackknife bias corrections are practically equivalent). This could possibly be caused by relatively smooth link functions in the standard models. The RMSEs reported in Table 3 show that the FE logit estimates of the selection equation parameters β_{12} and β_{13} are overall more precise than GMM-ADG-OPDG, but this does not have a substantial effect on the

precision of estimates of the mean equation parameter β_{23}. The RMSE of GMM-ADG-OPDG estimates of β_{23} is slightly smaller than that of Kyriazidou (1997). In all cases, ASEs are close to RMSE and the coverage rates are close to the nominal level of 95%.

The marginal effects in Table 3 are compared to existing methods where possible. In the selection equation, TME equals SME1 as there is just one linear combination involved and it is compared to the AME obtained by the Chernozhukov et al. (2019) approach labelled CFN. In the linear outcome equation, SME2 can be compared to the regression coefficients obtained by Kyriazidou (1997), but SME1 provides additionally the AME due to sample selection driven by linear combination $X_{it}^{\top}\beta_1$ and TME equals the sum of SME1 and SME2. All marginal effects are estimated without a substantial bias and with a good precision.

5 Empirical application

We now illustrate the use of the proposed estimator by applying it to the estimation of dynamic earnings of females. We use the data studied by Semykina and Wooldridge (2013) that originate from the Panel Study of Income Dynamics (PSID) in years 1980–1992. The sample contains 486 women below 60 years of age and observed for 12 years, both in and out of the labor force. We compare the results by the proposed GMM-ADG-OPDG estimator with the results of Semykina and Wooldridge (2013), labelled SW, and Kyriazidou (2001), labelled KYR.

In this application, the sample selection model discussed in Sections 2 and 4.4 is used. Contrary to the specific data-generating process used in simulated examples, the estimation procedure does not however impose the linearity of the outcome equation in any way. Recall that the sample selection model (3) thus obeys the following outcome and selection equations:

$$Y_{2it} = \phi_{2t}(X_{it}^{\top}\beta_2, X_{it}^{\top}\beta_1, \alpha_{2i}, \alpha_{1i}, U_{2it}) \text{ observed if } Y_{1it} = \phi_{1t}(X_{it}^{\top}\beta_1, \alpha_{1i}, U_{1it}) > 0.$$

Here the dependent outcome variable Y_{2it} equals the natural logarithm of the average annual hourly earnings, which are observed or not depending on whether the person was employed in a given year or not. This response variable is complemented by the binary dependent variable Y_{1it} corresponding to the selection equation and indicating the employment status (i.e., the earnings are observed or not). Given that the sample contains only women with completed education, many individual characteristics such as race or education are parts of time-invariant

individual-specific effects α_{1i} and α_{2i}. Therefore, the explanatory variables X_{it} only include the lag of the log-earnings $Y_{2i(t-1)}$, the labor market experience Exp_{it} (in years, constructed from the employment status data Y_{1it}) and its square Exp_{it}^2, and as a control for the selection into employment, the number of children $Kids_{it}$. Because of the lagged dependent variable, we verify the validity of Assumption 3(iii) in this model in Appendix I.

For the identification of the coefficients, it is assumed that the number of children does not directly affect the earnings, and as in Semykina and Wooldridge (2013), that the past earnings do not affect the current labor force participation. Additionally, we normalize the coefficients of the lagged dependent variable in the outcome equation and of the number of children in the selection equation to 1 and -1, respectively. Finally, note that, given that the cross-validated bandwidth for the normalized explanatory variables is much larger than a unit increase in $Kids_{it}$, we treat the number of children as a continuous variable in this application.

Including the experience and its square in the estimation is typically not accommodated by the methods relying on locally estimated derivatives. For the proposed method, estimation with an explanatory variable and its square can be performed as described in Supplementary Appendix H. Since Exp_{it}^2 was however not significant for any of the considered estimators, the results presented here contain only Exp_{it} and are obtained by the GMM-ADG-OPDG method implemented in the same way as in Section 4.1 with the following exception. Since the lagged dependent variable is included in the estimation, the expectations in Theorem 1 require the dependent variable Y_{2it} and its two lags $Y_{2i(t-1)}$ and $Y_{2i(t-\Delta)}$. For the first-order differencing, $\Delta = 1$, every woman used in estimation has to work for the three consecutive periods, and thus the change in their labor market experiences between times t and $t - \Delta = t - 1$ is always 1. A similar argument applies also to $\Delta = 2$ if we want to identify both the effect of the labor market experience and its square. Hence, we use only lags $\Delta \geq 3$. We thus proceed as discussed in detail in Supplementary Appendix D and use the orders of differencing $\Delta = 3, \ldots, 6$ (higher orders of differencing are not used as they result in less than 500 observed wages for $\Delta > 6$).

All estimation results are summarized in Table 4, which contains both the identification constraints and the corresponding coefficient and ME estimates. While SW and KYR provide coefficient estimates of the linear outcome equation, we can estimate using GMM-ADG-OPDG the coefficients and the marginal effects similarly to Section 4.1: the total AME (TME) X_{it} on the response and the AME $\Gamma_{1t,1}\beta_1$ and $\Gamma_{1t,2}\beta_2$ of X_{it} specific to linear combinations $X_{it}^\top\beta_1$ and

Table 4: Estimation results for the dynamic model of the logarithm of the hourly earnings. Superscripts ^{a,b,c} represent the significance at the 10%, 5%, and 1% levels, respectively, based on the bootstrapped standard errors, and $^{=}$ indicates the identification constraint.

Dependent variable	Explanatory variable	KYR	SW	GMM-ADG-OPDG				
				β_1	β_2	SME1	SME2	TME
Y_{2it}	Y_{2it-1}	0.043	0.579^{c}	$0.000^{=}$	$1.000^{=}$	$0.000^{=}$	0.434^{c}	0.434^{c}
	Exp_{it}	0.033^{b}	0.000	3.993^{c}	0.385	-0.156^{a}	0.167	0.010
	$Kids_{it}$	$0.000^{=}$	$0.000^{=}$	$-1.000^{=}$	$0.000^{=}$	0.039^{a}	$0.000^{=}$	0.039^{a}
Y_{1it}	Exp_{it}	—	—	3.993^{c}	—	0.193^{c}	—	0.193^{c}
	$Kids_{it}$	—	—	$-1.000^{=}$	—	-0.046^{c}	—	-0.048^{c}

$X_{it}^{\top}\beta_2$, respectively, which are labelled SME1 and SME2; see Section 2 and 4.1 for details. For example, we can identify separately the direct effect of a higher experience on the wages and the sample-selection effect of a higher experience on wages due to an increased employment probability. More specifically, while the selection variable Y_{1it} is determined by just one linear combination $X_{it}^{\top}\beta_1$ and SME1 and TME are thus identical, the outcome Y_{2it} depends on two linear combinations $X_{it}^{\top}\beta_1$ and $X_{it}^{\top}\beta_2$. In the outcome equation, the total marginal effect of X_{it} on Y_{2it} is thus decomposed into two parts: (i) $\Gamma_{1t,1}\beta_1$, which represents the "indirect" effect of X_{it} on Y_{2it} through the index $X_{it}^{\top}\beta_1$ determining the selection into employment and entering the outcome equation due to the sample-selection correction, and (ii) $\Gamma_{1t,2}\beta_2$, which quantifies the "direct" effect of X_{it} on Y_{2it} via the index $X_{it}^{\top}\beta_2$ entering only the outcome equation.

The estimation results in Table 4 indicate that both the labor market experience and the number of children have significant and expected impacts – positive and negative, respectively – on the labor force participation. In the outcome equation conditionally on working in the previous period, the labor market experience does not have a significant direct or total impact on the female log-earnings for all estimators except of KYR discussed later (see columns SW, SME2, and TME in Table 4). Interestingly, there is a weakly significant indirect effect SME1 that indicates a small effect of the number of children on the outcome due to selection into employment, which can be estimated only by the GMM-ADG-OPDG method. This confirms the presence of sample selection, which was also significant in the case of SW at 1% level. Furthermore, the coefficient of the lagged dependent variable in the outcome equation is normalized to 1, but we can judge its significance by the corresponding direct marginal effect SME2, which

is significant at the 1% level and provides evidence for the dependence in the earnings. Its value is a bit smaller than that obtained by SW, which could be related to possible nonlinearity of the outcome equation or specification of the selection equation. Finally, let us note that – as argued by Semykina and Wooldridge (2013) – the KYR estimate of the autoregressive coefficient is imprecise and insignificant, which likely leads to the significant KYR coefficient of Exp_{it}, capturing the missing effect of the lagged dependent variable. This result is robust to the number of lags used as instruments and to the employed sets of moment conditions.

6 Conclusion

In this paper, we show that the parameters of nonseparable multiple-index models with correlated random effects are identified by the average difference of gradients and the outer product of the difference of gradients. We propose a GMM estimator based on the local polynomial regression and establish its consistency and asymptotically normality. For a number of nonlinear panel data models, the GMM estimator seems to perform adequately well in a simulation study and application. Future research should focus on improving the selection of the number of indices, and in large models, on the variable selection.

A Proof of Theorem 1

Proof. [Proof of Theorem 1] The first part of the proof is similar to the proof of Theorem 1 in Čížek and Lei (2018). We write the expectation $\mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}]$ at $(x_t, x_{t-\Delta})$ as

$$\begin{aligned}&\mathrm{E}[Y_{it}|X_{it} = x_t, X_{i(t-\Delta)} = x_{t-\Delta}] = \mathrm{E}[\phi_t(X_{it}^\top B, \alpha_i, U_{it})|X_{it} = x_t, X_{i(t-\Delta)} = x_{t-\Delta}] \\ &= \int \phi_t(x_t^\top B, \alpha, u_t) F_{U_t,\alpha|X_t,X_{t-\Delta}}(du_t, d\alpha|X_{it} = x_t, X_{i(t-\Delta)} = x_{t-\Delta}).\end{aligned} \tag{A.1}$$

For notational convenience, we use $F(u_t, \alpha|x_t, x_{t-\Delta}))$ and $f(\alpha|x_t, x_{t-\Delta})$ as a shorthand for $F_{U_t,\alpha|X_t,X_{t-\Delta}}(u_t, \alpha|X_{it} = x_t, X_{i(t-\Delta)} = x_{t-\Delta})$ and $f_{\alpha|X_t,X_{t-\Delta}}(\alpha|X_{it} = x_t, X_{i(t-\Delta)} = x_{t-\Delta})$, respectively. Recall that $\varphi_t(X_{it}^\top B, \alpha_i) = \mathrm{E}_U(\phi_t(X_{it}^\top B, \alpha_i, U_{it}))$ and U_{it} is independent of α_i,

X_{it}, and $X_{i(t-\Delta)}$ by Assumption 2. We use successive conditioning to write (A.1) as

$$\begin{aligned}
&\mathrm{E}[Y_{it}|X_{it}=x_t, X_{i(t-\Delta)}=x_{t-\Delta}] \\
&= \int \left[\int \phi_t(x_t^\top B, \alpha, u_t) F_{U_t|\alpha, X_t, X_{t-\Delta}}(du_t|\alpha, x_t, x_{t-\Delta})\right] f(\alpha|x_t, x_{t-\Delta}) d\alpha \\
&= \int \left[\int \phi_t(x_t^\top B, \alpha, u_t) F_{U_t}(du_t)\right] f(\alpha|x_t, x_{t-\Delta}) d\alpha \\
&= \int \varphi_t(x_t^\top B, \alpha) f(\alpha|x_t, x_{t-\Delta}) d\alpha,
\end{aligned}$$

where F_{U_t} represents the distribution function of U_{it}. By Assumptions 4, 5, and 6, the derivatives of the above expectation exist and interchanging the order of integration and derivative can be thus justified. Hence, the derivative of the above expectation equals

$$\begin{aligned}
\frac{\partial}{\partial x_t^\top} \mathrm{E}[Y_{it}|X_{it}=x_t, X_{i(t-\Delta)}=x_{t-\Delta}] = &\int \left[\frac{\partial}{\partial x_t^\top} \varphi_t(x_t^\top B, \alpha)\right] f(\alpha|x_t, x_{t-\Delta}) d\alpha \\
&+ \int \varphi_t(x_t^\top B, \alpha) \frac{\partial}{\partial x_t^\top} f(\alpha|x_t, x_{t-\Delta}) d\alpha.
\end{aligned} \tag{A.2}$$

We can rewrite the first part of the right handside of (A.2) as

$$\sum_{r=1}^{R} \int \varphi_{tr}'(x_t^\top B, \alpha) f(\alpha|x_t, x_{t-\Delta}) d\alpha \cdot \beta_r^\top = \sum_{r=1}^{R} \mathrm{E}\left[\varphi_{tr}'(X_{it}^\top B, \alpha_i)|X_{it}=x_t, X_{i(t-\Delta)}=x_{t-\Delta}\right] \beta_r^\top. \tag{A.3}$$

since $\frac{\partial}{\partial x_t^\top}\varphi_t(x_t^\top \beta, \alpha) = \sum_{r=1}^{R} \varphi_{tr}'(x_t^\top B, \alpha)\beta_r^\top$. As Assumption 3 implies $f(\alpha|x_t, x_{t-\Delta}) = f(\alpha|x_t + x_{t-\Delta})$, the second part of (A.2) can be rewritten as

$$\int \varphi_t(x_t^\top B, \alpha) \frac{\partial}{\partial (x_t + x_{t-\Delta})^\top} f(\alpha|x_t + x_{t-\Delta}) d\alpha. \tag{A.4}$$

As argued in Čížek and Lei (2018), there are two components in the marginal effects (A.2): (A.3) exhibits the direct effect of a change in X_{it} averaged over α_i while (A.4) characterizes the indirect effect of a change in X_{it} on Y_{it} induced by the change of the individual effects α_i. On the other hand, the marginal effects of $X_{i(t-\Delta)}$ on Y_{it} contains only the indirect effect:

conditionally on X_{it} and α_i, Y_{it} is independent of $X_{i(t-\Delta)}$. Therefore, we obtain

$$\begin{aligned}
&\frac{\partial}{\partial x_{t-\Delta}^\top} \mathrm{E}[Y_{it}|X_{it}=x_t, X_{i(t-\Delta)}=x_{t-\Delta}] \\
&= \int \left[\int \phi_t(x_t^\top B, \alpha, u_t) F_{U_t}(du_t) \right] \frac{\partial}{\partial x_{t-\Delta}^\top} f(\alpha|x_t, x_{t-\Delta}) d\alpha \\
&= \int \varphi_t(x_t^\top B, \alpha) \frac{\partial}{\partial (x_t + x_{t-\Delta})^\top} f(\alpha|x_t + x_{t-\Delta}) d\alpha.
\end{aligned} \tag{A.5}$$

Note that the last expressions in (A.5) and (A.4) are identical, so we can rewrite (A.2) as

$$\begin{aligned}
&\frac{\partial}{\partial x_t^\top} \mathrm{E}[Y_{it}|X_{it}=x_t, X_{i(t-\Delta)}=x_{t-\Delta}] \\
&= \sum_{r=1}^{R} \mathrm{E}\left[\varphi_{tr}'(X_{it}^\top B, \alpha_i) | X_{it}=x_t, X_{i(t-\Delta)}=x_{t-\Delta} \right] \beta_r^\top + \frac{\partial}{\partial x_{t-\Delta}^\top} \mathrm{E}[Y_{it}|X_{it}=x_t, X_{i(t-\Delta)}=x_{t-\Delta}].
\end{aligned}$$

After taking the expectation with respect to X_{it} and $X_{i(t-\Delta)}$ and rearranging the equation,

$$\delta_{\mathfrak{T}} = \mathrm{E}\left\{ \frac{\partial}{\partial x_t^\top} \mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}] - \frac{\partial}{\partial x_{t-\Delta}^\top} \mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}] \right\} = \sum_{r=1}^{R} \mathrm{E}\left[\varphi_{tr}'(X_{it}^\top B, \alpha_i) \right] \beta_r^\top.$$

Next, denoting the $p \times R$ matrix $\Gamma_{1t} = \left\{ \mathrm{E}\left[\varphi_{tr}'(X_{it}^\top B, \alpha_i) \right] \right\}_{r=1}^{R}$, it follows that

$$\mathrm{E}\left\{ \frac{\partial}{\partial x_t^\top} \mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}] - \frac{\partial}{\partial x_{t-\Delta}^\top} \mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}] \right\} = \Gamma_{1t} B^\top. \tag{A.6}$$

Similarly for the $R \times R$ matrix $\Gamma_{2t} = \left\{ \mathrm{E}\left[\varphi_{tr}'(X_{it}^\top B, \alpha_i)^\top \varphi_{ts}'(X_{it}^\top B, \alpha_i) \right] \right\}_{r,s=1}^{R}$, we obtain by the law of iterated expectations

$$\begin{aligned}
\delta_{\mathfrak{T}\mathfrak{T}} =& \mathrm{E}\left[\left\{ \frac{\partial}{\partial x_t^\top} \mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}] - \frac{\partial}{\partial x_{t-\Delta}^\top} \mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}] \right\}^\top \right. \\
&\left. \times \left\{ \frac{\partial}{\partial x_t^\top} \mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}] - \frac{\partial}{\partial x_{t-\Delta}^\top} \mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}] \right\} \right] \\
=& \mathrm{E}\left[\left\{ \sum_{r=1}^{R} \mathrm{E}\left[\varphi_{tr}'(X_{it}^\top B, \alpha_i) | X_{it}=x_t, X_{i(t-\Delta)}=x_{t-\Delta} \right] \beta_r^\top \right\}^\top \right. \\
&\left. \times \left\{ \sum_{s=1}^{R} \mathrm{E}\left[\varphi_{ts}'(X_{it}^\top B, \alpha_i) | X_{it}=x_t, X_{i(t-\Delta)}=x_{t-\Delta} \right] \beta_s^\top \right\} \right] \\
=& \sum_{r=1}^{R} \sum_{s=1}^{R} \beta_r \, \mathrm{E}\left[\varphi_{tr}'(X_{it}^\top B, \alpha_i) \varphi_{ts}'(X_{it}^\top B, \alpha_i) \right] \beta_s^\top = B \Gamma_{2t} B^\top.
\end{aligned} \tag{A.7}$$

Finally, given that $\delta_{\mathfrak{T}}$ and $\delta_{\mathfrak{T}\mathfrak{T}}$ are identified, let Assumption 8 hold and let θ_1, θ_2, and θ denote all unconstrained parameters of (B, Γ_{1t}), (B, Γ_{2t}), and $(B, \Gamma_{1t}, \Gamma_{2t})$, respectively, and $g_{\mathfrak{T}\mathfrak{T}}(\theta) = \left(g^1_{\mathfrak{T}}(\theta), g^2_{\mathfrak{T}\mathfrak{T}}(\theta)\right)$ with

$$g^1_{\mathfrak{T}}(\theta) = \text{vec}(\delta_{\mathfrak{T}} - \Gamma_{1t}B^\top)$$

and

$$g^2_{\mathfrak{T}\mathfrak{T}}(\theta) = \text{vec}(\delta_{\mathfrak{T}\mathfrak{T}} - B\Gamma_{2t}B^\top).$$

As we have derived in equations (A.6) and (A.7), these moment conditions $g_{\mathfrak{T}\mathfrak{T}}(\theta) = 0$ are satisfied at the true parameter value θ_0 of θ. Additionally, Lemma 3 in Donkers and Schafgans (2008) together with Assumption 8 imply that $g_{\mathfrak{T}\mathfrak{T}}(\theta)^\top g_{\mathfrak{T}\mathfrak{T}}(\theta)$ and $g^2_{\mathfrak{T}\mathfrak{T}}(\theta)^\top g^2_{\mathfrak{T}\mathfrak{T}}(\theta)$ have a local minimum equal to 0 at the true parameter value of θ_0. To show that the minimum at θ_0 is unique and thus global, we focus on the subset $g^2_{\mathfrak{T}\mathfrak{T}}(\theta) = 0$ of the moment conditions $g_{\mathfrak{T}\mathfrak{T}}(\theta) = 0$, which is sufficient for the identification. Given that matrix B under Assumption 8(i) and B under Assumption 8(ii) are equivalent up to the multiplication by a fixed full-rank matrix, we verify the uniqueness only under Assumption 8(ii).

Let us first note that the moment equation $g^2_{\mathfrak{T}\mathfrak{T}}(\theta_0) = 0$ being satisfied at θ_0 and Assumption 8(ii) imply that matrix $\delta_{\mathfrak{T}\mathfrak{T}}$ has rank R. Further under Assumption 8(ii), any solution of the moment conditions $g^2_{\mathfrak{T}\mathfrak{T}}(\theta) = 0$ has to consist of an orthonormal matrix B and a full-rank diagonal matrix Γ_{2t}. Hence for a solution of $g^2_{\mathfrak{T}\mathfrak{T}}(\theta) = 0$, it has to hold $\delta_{\mathfrak{T}\mathfrak{T}} - B\Gamma_{2t}B^\top = 0$, or equivalently after multiplication by B from the right, $\delta_{\mathfrak{T}\mathfrak{T}}B - B\Gamma_{2t} = 0$. For any solution $B = (\beta_1, \ldots, \beta_R)$ and $\Gamma_{2t} = (\gamma_1, \ldots, \gamma_R)$ of the moment equations, it thus holds by Assumption 8(ii) and $\delta_{\mathfrak{T}\mathfrak{T}}$ being positive semidefinite that $\delta_{\mathfrak{T}\mathfrak{T}}\beta_r - \gamma_r\beta_r = 0, r = 1, \ldots, R$, $B^\top B = I$, and $\gamma_1 > \gamma_2 > \ldots > \gamma_R > 0$. Scalars γ_r and columns β_r are thus non-zero eigenvalues and the corresponding eigenvectors of $\delta_{\mathfrak{T}\mathfrak{T}}$. Given that $\delta_{\mathfrak{T}\mathfrak{T}}$ has rank R, no eigenvalues γ_r are degenerate, and vectors $\beta_1, \ldots, \beta_R$ are orthonormal, the eigen decomposition of $\delta_{\mathfrak{T}\mathfrak{T}}$ is unique, and consequently, there is only one solution of moment conditions $g^2_{\mathfrak{T}\mathfrak{T}}(\theta) = 0$ with Γ_{2t} having its diagonal elements sorted in the descending order. As $g^2_{\mathfrak{T}\mathfrak{T}}(\theta_0) = 0$, this solution thus has to coincide with θ_0. ■

B Notation and auxiliary lemmas

In this appendix, we introduce the notation used in the main theorems and auxiliary lemmas. We also state important auxiliary results, which are either found in the existing works of in Čížek and Lei (2018), Li et al. (2003), and Masry (1996) or are formally proved in Supplementary Appendix E. These auxiliary results are then used in the proofs of Theorems 2–4 in Appendix C.

The auxiliary results of this section concern the properties of the local polynomial estimator (9) and the corresponding averages (10) and (11). As this estimation is performed for each response variable separately, we assume for simplicity of notation in this section that Y_{it} represents one particular (scalar) response variable, denoted in the main text as $Y_{c,it}$ for $c \in \{1, \dots, d_y\}$. The same notation, omitting the subscript indicating the response component considered, is then applied also to the conditional expectations $m_{\mathfrak{T}}(z)$, their derivative differences $\delta_{\mathfrak{T}}(z)$, residuals $V_{i\mathfrak{T}}$, and their (co)variances $\sigma_{\mathfrak{T}\mathfrak{S}}(z)$ replacing matrices $\Sigma_{\mathfrak{T}\mathfrak{S}}(z)$. Given one particular response, we now introduce notation and some theorems of Čížek and Lei (2018), Masry (1996), and Li et al. (2003); the assumptions of those theorems are included in Assumptions 1–9.

Since $m_{\mathfrak{T}}(\cdot)$ is $(p+1)$-times differentiable with uniformly bounded derivatives by Assumption 9, $m_{\mathfrak{T}}(z)$ can be locally approximated at some z_0 by a polynomial of order p:

$$m_{\mathfrak{T}}(z) \approx \sum_{0 \leq |\underline{k}| \leq p} \frac{1}{\underline{k}!} D^{\underline{k}} m_{\mathfrak{T}}(v)|_{v=z_0} (z - z_0)^{\underline{k}},$$

where $\underline{k} = (k_1, \dots, k_{2d}) \in \mathbb{N}^{2d}$, $\underline{k}! = k_1! \times \cdots \times k_{2d}!$, $|\underline{k}| = \sum_{i=1}^{2d} k_i$, $z^{\underline{k}} = z_1^{k_1} \times \cdots \times z_{2d}^{k_{2d}}$,

$$\sum_{0 \leq |\underline{k}| \leq p} = \sum_{j=0}^{p} \sum_{k_1=0}^{j} \cdots \sum_{k_{2d}=0; k_1 + \cdots + k_{2d} = j}, \quad \text{and} \quad (D^{\underline{k}} m_{\mathfrak{T}})(z) = \frac{\partial^{\underline{k}} m_{\mathfrak{T}}(z)}{\partial z_1^{k_1} \dots \partial z_{2d}^{k_{2d}}}.$$

Further, define for $\underline{j} = (j_1, \dots, j_{2d}) \in \mathbb{N}^{2d}$ and $z \in \mathbb{R}^{2d}$

$$\begin{aligned} \bar{\tau}^{e}_{\mathfrak{T},\underline{j}}(z) &= \frac{1}{n} \sum_{i=1}^{n} \left(Y_{it} - m_{\mathfrak{T}}(Z_{i\mathfrak{T}}) \right) \left(\frac{Z_{i\mathfrak{T}} - z}{h_n} \right)^{\underline{j}} K_h \left(Z_{i\mathfrak{T}} - z \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} V_{i\mathfrak{T}} \left(\frac{Z_{i\mathfrak{T}} - z}{h_n} \right)^{\underline{j}} K_h \left(Z_{i\mathfrak{T}} - z \right), \end{aligned} \tag{A.8}$$

and

$$\bar{s}^e_{\mathfrak{T},\underline{j}}(z) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Z_{i\mathfrak{T}}-z}{h_n}\right)^{\underline{j}} K_h\left(Z_{i\mathfrak{T}}-z\right), \tag{A.9}$$

where $V_{i\mathfrak{T}} = Y_{it} - m_{\mathfrak{T}}(Z_{i\mathfrak{T}})$ and $K_h(u) = h_n^{-2d}K\left(u/h_n\right)$. We indicate the dependence of $\bar{\tau}^e_{\mathfrak{T},j}$,$\bar{s}^e_{\mathfrak{T},j}(z)$, and other averages on the sample size n by the bar above the letters for notational convenience. As we take the "short" T approach, the asymptotic expressions should be understood for $n \to +\infty$ with $T > 1$ being fixed in what follows.

As the moment conditions of the proposed GMM estimator are based on the local polynomial estimator, we have to first characterize its Bahadur-type representation following Masry (1996) and Li et al. (2003). Adapting their conventional notation to our setting, we first express $\bar{\tau}^e_{\mathfrak{T},\underline{j}}$ in a matrix form in a lexicographical order. Let $N_l = (l+2d-1)!/(l!(2d-1)!)$ be the number of distinct $2d$-tuples with $|\underline{j}| \equiv j_1 + \cdots + j_{2d} = l$. In local polynomial estimation, N_l denotes the number of the distinct lth order partial derivatives of $m_{\mathfrak{T}}(z)$. We arrange these $2d$-tuples as a sequence in a lexicographical order. The highest priority is given to the last position so that $(0,\ldots,0,l)$ is the first element in the sequence and $(l,0,\ldots,1)$ is the last element. Let $g_l^{-1} \equiv g^{-1}_{|\underline{j}|}$ denote this one-to-one mapping. We arrange the $N_l = N_{|\underline{j}|}$ values of $\bar{\tau}^e_{\mathfrak{T},\underline{j}}(z)$ in a column vector $\bar{\tau}_{\mathfrak{T},l}(z) = (\bar{\tau}^e_{\mathfrak{T},g_l(k)}(z))_{k=1}^{N_l}$ in the lexicographical order. We further define the column vector $\bar{\tau}_{\mathfrak{T}}(z) = (\bar{\tau}^\top_{\mathfrak{T},0}(z), \bar{\tau}^\top_{\mathfrak{T},1}(z), \ldots, \bar{\tau}^\top_{\mathfrak{T},p}(z))^\top$, where $\bar{\tau}_{\mathfrak{T},l}(z)$ is an $N_l \times 1$ vector with elements $\bar{\tau}^e_{\mathfrak{T},\underline{j}}(z)$, $|\underline{j}| = l$, arranged according to the lexicographical order; $\bar{\tau}_{\mathfrak{T}}(z)$ has thus dimension $N \times 1$ with $N = \sum_{l=0}^{p} N_l$.

Furthermore, by arranging $\bar{s}^e_{\mathfrak{T},\underline{j}+\underline{k}}(z)$ in a matrix $\bar{S}_{\mathfrak{T},|\underline{j}|,|\underline{k}|}(z)$ in the lexicographical order with the (l_1,l_2)th element given by $[\bar{S}_{\mathfrak{T},|\underline{j}|,|\underline{k}|}(z)]_{l_1 l_2} = \bar{s}^e_{\mathfrak{T},g_{|\underline{j}|}(l_1)+g_{|\underline{k}|}(l_2)}(z)$, we define the $N \times N$ matrix $\bar{S}_{\mathfrak{T}}(z)$ and $N \times N_{p+1}$ matrix $\bar{B}_{\mathfrak{T}}(z)$ by

$$\bar{S}_{\mathfrak{T}}(z) = \begin{pmatrix} \bar{S}_{\mathfrak{T},0,0}(z) & \bar{S}_{\mathfrak{T},0,1}(z) & \ldots & \bar{S}_{\mathfrak{T},0,p}(z) \\ \bar{S}_{\mathfrak{T},1,0}(z) & \bar{S}_{\mathfrak{T},1,1}(z) & \ldots & \bar{S}_{\mathfrak{T},1,p}(z) \\ \vdots & \vdots & \ddots & \vdots \\ \bar{S}_{\mathfrak{T},p,0}(z) & \bar{S}_{\mathfrak{T},p,1}(z) & \ldots & \bar{S}_{\mathfrak{T},p,p}(z) \end{pmatrix} \quad \text{and} \quad \bar{B}_{\mathfrak{T}}(z) = \begin{pmatrix} \bar{S}_{\mathfrak{T},0,p+1}(z) \\ \bar{S}_{\mathfrak{T},1,p+1}(z) \\ \vdots \\ \bar{S}_{\mathfrak{T},p,p+1}(z) \end{pmatrix}.$$

Let $\mu_{\underline{j}} = \int_{\mathbb{R}^{2d}} u^{\underline{j}} K(u)du$ and $v_{s,\underline{j}} = \int_{\mathbb{R}^{2d}} u_s u^{\underline{j}} K(u)du$, where u_s is the sth element of vector u. Then we define $N_i \times N_j$ dimensional matrices $M_{i,j}$ and $Q_{s,i,j}$ to have their (l_1,l_2)th elements

given by $\mu_{g_i(l_1)+g_j(l_2)}$ and $v_{s,g_i(l_1)+g_j(l_2)}$, respectively, for $s = 1, \ldots, 2d$, and we further define

$$M = \begin{pmatrix} M_{0,0} & M_{0,1} & \ldots & M_{0,p} \\ M_{1,0} & M_{1,1} & \ldots & M_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ M_{p,0} & M_{p,1} & \ldots & M_{p,p} \end{pmatrix}, B = \begin{pmatrix} M_{0,p+1} \\ M_{1,p+1} \\ \vdots \\ M_{p,p+1} \end{pmatrix}, Q_s = \begin{pmatrix} Q_{s,0,0} & Q_{s,0,1} & \ldots & Q_{s,0,p} \\ Q_{s,1,0} & Q_{s,1,1} & \ldots & Q_{s,1,p} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{s,p,0} & Q_{s,p,1} & \ldots & Q_{s,p,p} \end{pmatrix}.$$

Let $f'_{\mathfrak{T},s}(z)$ denote the sth element of the first derivative $f'_{\mathfrak{T}}(z)$ of the density function $f_{\mathfrak{T}}(z)$ of $Z_{i\mathfrak{T}}$, $s = 1, \ldots, 2d$, and define $M^{f_{\mathfrak{T}}}(z) = Mf_{\mathfrak{T}}(z)$, $Q^{f_{\mathfrak{T}}}(z) = \sum_{s=1}^{2d} f'_{\mathfrak{T},s}(z)Q_s$, and $G^{f_{\mathfrak{T}}}(z) = [M^{f_{\mathfrak{T}}}(z)]^{-1}Q^{f_{\mathfrak{T}}}(z)[M^{f_{\mathfrak{T}}}(z)]^{-1}$.

By Masry (1996, equation (2.13)) and Li et al. (2003, equation (A.9)), as $n \to +\infty$

$$\hat{\beta}_{\mathfrak{T}}(z) - \beta_{\mathfrak{T}}(z) = \bar{S}_{\mathfrak{T}}^{-1}(z)\bar{\tau}_{\mathfrak{T}}(z) + h_n^{p+1}\bar{S}_{\mathfrak{T}}^{-1}\bar{B}_{\mathfrak{T}}(z)m_{\mathfrak{T}}^{(p+1)}(z) + O_p(h_n^{p+2}), \quad 0 \leq |k| \leq p, \text{ (A.10)}$$

where $\hat{\beta}_{\mathfrak{T}}(z) = (h_n^0\hat{b}_{0,\mathfrak{T}}^\top(z), \ldots, h_n^p\hat{b}_{p,\mathfrak{T}}^\top(z))^\top$ and $\hat{b}_{k,\mathfrak{T}}(z)$ are the estimates of parameter vectors $(b_{\underline{j},\mathfrak{T}}(z))_{|\underline{j}|=k}$ in objective function (6) and $m_{\mathfrak{T}}^{(p+1)}(z)$ is the N_{p+1} elements of derivatives $(D^{\underline{j}}m_{\mathfrak{T}})(z)/\underline{j}!$ for $|\underline{j}| = p+1$ arranged in the lexicographical order.

Recall that local derivative estimator $\hat{\delta}_{\mathfrak{T}}(z) = L\hat{b}_{\mathfrak{T}}(z) = h_n^{-1}L\hat{\beta}_{\mathfrak{T}}(z)$ in equation (8) is defined as the difference of the first d and last d elements of $\hat{b}_{1,\mathfrak{T}}(z)$. The average derivative estimator and average outer product of gradients are thus defined as in equations (10) and (11) by

$$\hat{\delta}_{\mathfrak{T}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\delta}_{\mathfrak{T}}(Z_{i\mathfrak{T}}) = \frac{1}{nh_n}\sum_{i=1}^{n}L\hat{\beta}_{\mathfrak{T}}(Z_{i\mathfrak{T}}).$$
$$\hat{\delta}_{\mathfrak{T}\mathfrak{T}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\delta}_{\mathfrak{T}}(Z_{i\mathfrak{T}})\hat{\delta}_{\mathfrak{T}}^\top(Z_{i\mathfrak{T}}) = \frac{1}{nh_n^2}\sum_{i=1}^{n}L\hat{\beta}_{\mathfrak{T}}(Z_{i\mathfrak{T}})\hat{\beta}_{\mathfrak{T}}^\top(Z_{i\mathfrak{T}})L^\top.$$

These averages will be now decomposed and analyzed using the representation (A.10). In particular, the decompositions used in Supplementary Appendix E rely on the following averages:

$$\bar{A}_{\mathfrak{T}}^{11} = \frac{1}{n}\sum_{i=1}^{n}\bar{S}_{\mathfrak{T}}^{-1}(Z_{i\mathfrak{T}})\bar{\tau}_{\mathfrak{T}}(Z_{i\mathfrak{T}}),$$
$$\bar{A}_{\mathfrak{T}}^{12} = \frac{1}{n}\sum_{i=1}^{n}\bar{S}_{\mathfrak{T}}^{-1}(Z_{i\mathfrak{T}})\bar{B}_{\mathfrak{T}}(Z_{i\mathfrak{T}})m_{\mathfrak{T}}^{(p+1)}(Z_{i\mathfrak{T}}),$$
$$\bar{J}_{\mathfrak{T}}^{11} = \frac{1}{n}\sum_{i=1}^{n}(M^{f_{\mathfrak{T}}}(Z_{i\mathfrak{T}}))^{-1}\bar{\tau}_{\mathfrak{T}}(Z_{i\mathfrak{T}}),$$

$$\bar{J}^{12}_{\mathfrak{T}} = \frac{1}{n}\sum_{i=1}^{n} G^{f_{\mathfrak{T}}}(Z_{i\mathfrak{T}})\bar{\tau}_{\mathfrak{T}}(Z_{i\mathfrak{T}}),$$
$$\bar{J}^{21}_{\mathfrak{T}\mathfrak{T}} = \frac{1}{n}\sum_{i=1}^{n} (M^{f_{\mathfrak{T}}}(Z_{i\mathfrak{T}}))^{-1}\bar{\tau}_{\mathfrak{T}}(Z_{i\mathfrak{T}})m^{'}_{\mathfrak{T}}(Z_{i\mathfrak{T}})^{\top},$$
$$\bar{J}^{22}_{\mathfrak{T}\mathfrak{T}} = \frac{1}{n}\sum_{i=1}^{n} G^{f_{\mathfrak{T}}}(Z_{i\mathfrak{T}})\bar{\tau}_{\mathfrak{T}}(Z_{i\mathfrak{T}})m^{'}_{\mathfrak{T}}(Z_{i\mathfrak{T}})^{\top},$$

and finally, $\bar{J}^{1}_{\mathfrak{T}} = \bar{J}^{11}_{\mathfrak{T}}/h_n - \bar{J}^{12}_{\mathfrak{T}}$ and $\bar{J}^{2}_{\mathfrak{T}\mathfrak{T}} = \bar{J}^{21}_{\mathfrak{T}\mathfrak{T}}/h_n - \bar{J}^{22}_{\mathfrak{T}\mathfrak{T}}$.

Using this notation, the following results are derived in Supplementary Appendix E.

Lemma 1. *Under Assumptions 1–9,* $\bar{A}^{12}_{\mathfrak{T}} = A^{1}_{\mathfrak{T}} + O(h_n)$ *and* $\bar{A}^{22}_{\mathfrak{T}\mathfrak{T}} = A^{2}_{\mathfrak{T}\mathfrak{T}} + O(h_n)$ *almost surely as* $n \to \infty$*, where* $A^{1}_{\mathfrak{T}} = M^{-1}B\,\mathrm{E}[m^{(p+1)}_{\mathfrak{T}}(Z_{i\mathfrak{T}})]$ *and* $A^{2}_{\mathfrak{T}\mathfrak{T}} = M^{-1}B\,\mathrm{E}[m^{(p+1)}_{\mathfrak{T}}(Z_{i\mathfrak{T}})m^{'}_{\mathfrak{T}}(Z_{i\mathfrak{T}})^{\top}]$.

Lemma 2. *Under Assumptions 1–9,* $\bar{J}^{11}_{\mathfrak{T},r} = O_p\left((nh^d_n)^{-1}\right)$ *and* $\sqrt{n}\bar{J}^{21}_{\mathfrak{T}\mathfrak{T},r}/h_n \to N(0,\Phi^{21}_{\mathfrak{T},r})$ *as* $n \to \infty$ *for* $r = 2,\ldots,2d+1$*, where*

$$\Phi^{21}_{\mathfrak{T}\mathfrak{T},r} = \mathrm{E}\left[\sigma_{\mathfrak{T}\mathfrak{T}}(Z_{i\mathfrak{T}})m^{''}_{\mathfrak{T}}(Z_{i\mathfrak{T}})e^{\top}_{r-1}e_{r-1}m^{''}_{\mathfrak{T}}(Z_{i\mathfrak{T}})^{\top}\right],$$

and $m^{''}_{\mathfrak{T}}(z) = \partial m^{'}_{\mathfrak{T}}(z)/\partial z^{\top}$.

Lemma 3. *Under Assumptions 1–9,* $\sqrt{n}\bar{J}^{12}_{\mathfrak{T},r} \to N(0,\Phi^{12}_{\mathfrak{T},r})$ *and* $\sqrt{n}\bar{J}^{22}_{\mathfrak{T},r} \to N(0,\Phi^{22}_{\mathfrak{T},r})$ *in distribution as* $n \to \infty$ *for* $r = 2,\ldots,2d+1$*, where*

$$\Phi^{12}_{\mathfrak{T},r} = \mathrm{E}\left[\sigma_{\mathfrak{T}\mathfrak{T}}(Z_{i\mathfrak{T}})G_{\mathfrak{T};r,1}(Z_{i\mathfrak{T}})G_{\mathfrak{T};r,1}(Z_{i\mathfrak{T}})\right],$$

$$\Phi^{22}_{\mathfrak{T},r} = \mathrm{E}\left[\sigma_{\mathfrak{T}\mathfrak{T}}(Z_{i\mathfrak{T}})G_{\mathfrak{T};r,1}(Z_{i\mathfrak{T}})G_{\mathfrak{T};r,1}(Z_{i\mathfrak{T}})m^{'}_{\mathfrak{T}}(Z_{i\mathfrak{T}})m^{'}_{\mathfrak{T}}(Z_{i\mathfrak{T}})^{\top}\right],$$

and the matrix $G_{\mathfrak{T}}(z) = G^{f_{\mathfrak{T}}}(z)M^{f_{\mathfrak{T}}}(z) = [M^{f_{\mathfrak{T}}}(z)]^{-1}Q^{f_{\mathfrak{T}}}(z)$.

Lemma 4. *Let* $G_{\mathfrak{T};\cdot 1}(z) = G_{\mathfrak{T}}(z)e_1$ *be the first column of* $G_{\mathfrak{T}}(z) = [M^{f_{\mathfrak{T}}}(z)]^{-1}Q^{f_{\mathfrak{T}}}(z)$ *and* $GM_{\mathfrak{T}}(z) = vec(-LG_{\mathfrak{T};\cdot 1}(z), -2L_1m^{''}_{\mathfrak{T}}(Z_{i\mathfrak{T}})L^{\top}_1 - LG_{\mathfrak{T};\cdot 1}(Z_{i\mathfrak{T}})m^{'}_{\mathfrak{T}}(Z_{i\mathfrak{T}})^{\top}L^{\top}_1 - L_1m^{'}_{\mathfrak{T}}(Z_{i\mathfrak{T}})G_{\mathfrak{T};\cdot 1}(Z_{i\mathfrak{T}})^{\top}L^{\top})$. *Under Assumptions 1–9,*

$$\sqrt{n}vec(L\bar{J}^{1}_{\mathfrak{T}}, L[\bar{J}^{2}_{\mathfrak{T}\mathfrak{T}} + \bar{J}^{2\top}_{\mathfrak{T}\mathfrak{T}}]L^{\top}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[V_{i\mathfrak{T}}GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})] \to N(0,\Phi_{\mathfrak{T}}) \tag{A.11}$$

in distribution as $n \to +\infty$*, where* $\Phi_{\mathfrak{T}} = \mathrm{E}\left[\sigma_{\mathfrak{T}\mathfrak{T}}(Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}})GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})^{\top}\right]$.

The final result derived in Supplementary Appendix E using Lemmas 1–4 can be stated for $n \to \infty$ as (see equations (S.30)–(S.31))

$$\frac{1}{n}\sum_{i=1}^{n}[\hat{\beta}_{\mathfrak{T}}(Z_{i\mathfrak{T}}) - \beta_{\mathfrak{T}}(Z_{i\mathfrak{T}})] = h_n \bar{J}^1_{\mathfrak{T}} + o_p(h_n^{3/2}n^{-1/2}) + h_n^{p+1}A^1_{\mathfrak{T}} + O_p(h_n^{p+2}) \tag{A.12}$$

and

$$\begin{aligned}&\frac{1}{n}L\sum_{i=1}^{n}[\hat{\beta}_{\mathfrak{T}}(Z_{i\mathfrak{T}})\hat{\beta}^{\top}_{\mathfrak{T}}(Z_{i\mathfrak{T}}) - \beta_{\mathfrak{T}}(Z_{i\mathfrak{T}})\beta^{\top}_{\mathfrak{T}}(Z_{i\mathfrak{T}})]L^{\top} \\ &= h_n^2\left\{\bar{J}^2_{\mathfrak{T}\mathfrak{T}} + \bar{J}^{2\top}_{\mathfrak{T}\mathfrak{T}}\right\}\{1+o_p(1)\} + o_p(h_n^{5/2}n^{-1/2}) + h_n^{p+2}\left\{A^2_{\mathfrak{T}\mathfrak{T}} + A^{2\top}_{\mathfrak{T}\mathfrak{T}}\right\} + O_p(h_n^{p+3}).\end{aligned} \tag{A.13}$$

C Proofs of Theorems 2–4

In Appendix B, (A.12)–(A.13) were derived for one particular response $Y_{c,it}$ rather than for the vector of responses. Since $\frac{\partial}{\partial x_t^{\top}}\mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}] = \left(\frac{\partial}{\partial x_t^{\top}}\mathrm{E}[Y_{c,it}|X_{it}, X_{i(t-\Delta)}]\right)_{c=1}^{R}$ and $\frac{\partial}{\partial x_t}\mathrm{E}[Y_{it}^{\top}|X_{it}, X_{i(t-\Delta)}]\frac{\partial}{\partial x_t^{\top}}\mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}] = \sum_{c=1}^{R}\frac{\partial}{\partial x_t}\mathrm{E}[Y_{c,it}|X_{it}, X_{i(t-\Delta)}]\frac{\partial}{\partial x_t^{\top}}\mathrm{E}[Y_{c,it}|X_{it}, X_{i(t-\Delta)}]$, denoting $\hat{\beta}_{c,\mathfrak{T}}(z) = (h_n^0\hat{b}^{\top}_{c,0,\mathfrak{T}}(z), \ldots, h_n^p\hat{b}^{\top}_{c,p,\mathfrak{T}}(z))^{\top}$ and $\hat{\beta}_{\mathfrak{T}}(z) = (\hat{\beta}^{\top}_{1,\mathfrak{T}}(z), \ldots, \hat{\beta}_{d_y,\mathfrak{T}}(z))$ allows us to rewrite (A.12)–(A.13) for all responses Y_{it} and $n \to \infty$ as

$$\frac{1}{n}\sum_{i=1}^{n}[\hat{\beta}_{\mathfrak{T}}(Z_{i\mathfrak{T}}) - \beta_{\mathfrak{T}}(Z_{i\mathfrak{T}})] = h_n \bar{J}^1_{\mathfrak{T}} + o_p(h_n^{3/2}n^{-1/2}) + h_n^{p+1}A^1_{\mathfrak{T}} + O_p(h_n^{p+2}) \tag{A.14}$$

and

$$\begin{aligned}&\frac{1}{n}L\sum_{i=1}^{n}[\hat{\beta}_{\mathfrak{T}}(Z_{i\mathfrak{T}})\hat{\beta}^{\top}_{\mathfrak{T}}(Z_{i\mathfrak{T}}) - \beta_{\mathfrak{T}}(Z_{i\mathfrak{T}})\beta^{\top}_{\mathfrak{T}}(Z_{i\mathfrak{T}})]L^{\top} \\ &= h_n^2\left\{\bar{J}^2_{\mathfrak{T}\mathfrak{T}} + \bar{J}^{2\top}_{\mathfrak{T}\mathfrak{T}}\right\}\{1+o_p(1)\} + o_p(h_n^{5/2}n^{-1/2}) \\ &+ h_n^{p+2}\left\{A^2_{\mathfrak{T}\mathfrak{T}} + A^{2\top}_{\mathfrak{T}\mathfrak{T}}\right\} + O_p(h_n^{p+3}),\end{aligned} \tag{A.15}$$

where

$$A^1_{\mathfrak{T}} = (A^1_{c,\mathfrak{T}})_{c=1}^{d_y}, \bar{J}^1_{\mathfrak{T}} = (\bar{J}^1_{c,\mathfrak{T}})_{c=1}^{d_y} \qquad \text{and} \qquad A^2_{\mathfrak{T}\mathfrak{T}} = (A^2_{c,\mathfrak{T}\mathfrak{T}})_{c=1}^{d_y}, \bar{J}^2_{c,\mathfrak{T}\mathfrak{T}} = (\bar{J}^2_{c,\mathfrak{T}\mathfrak{T}})_{c=1}^{d_y} \tag{A.16}$$

and the symbols with subscript c refer to the analogously labelled objects defined in Appendix B: $A^1_{c,\mathfrak{T}} = M^{-1}B\,\mathrm{E}[m^{(p+1)}_{c,\mathfrak{T}}(Z_{i\mathfrak{T}})]$ and $A^2_{c,\mathfrak{T}\mathfrak{T}} = M^{-1}B\,\mathrm{E}[m^{(p+1)}_{c,\mathfrak{T}}(Z_{i\mathfrak{T}})m'_{c,\mathfrak{T}}(Z_{i\mathfrak{T}})^\top]$ in Lemma 1, $\bar{J}^1_{c,\mathfrak{T}} = \frac{1}{n}\sum_{i=1}^n\{(M^{f_\mathfrak{T}}(Z_{i\mathfrak{T}}))^{-1}\bar{\tau}_{c,\mathfrak{T}}(Z_{i\mathfrak{T}})/h_n - G^{f_\mathfrak{T}}(Z_{i\mathfrak{T}})\bar{\tau}_{c,\mathfrak{T}}(Z_{i\mathfrak{T}})\}$ and $\bar{J}^2_{c,\mathfrak{T}\mathfrak{T}} = \frac{1}{n}\sum_{i=1}^n\{(M^{f_\mathfrak{T}}(Z_{i\mathfrak{T}}))^{-1}\bar{\tau}_{c,\mathfrak{T}}(Z_{i\mathfrak{T}})m'_{c,\mathfrak{T}}(Z_{i\mathfrak{T}})^\top/h_n - G^{f_\mathfrak{T}}(Z_{i\mathfrak{T}})\bar{\tau}_{c,\mathfrak{T}}(Z_{i\mathfrak{T}})m'_{c,\mathfrak{T}}(Z_{i\mathfrak{T}})^\top\}$.

Proof. [Proof of Theorem 2] Let $\tilde{\delta}_\mathfrak{T} = \frac{1}{n}\sum_{i=1}^n L_1 m'_\mathfrak{T}(Z_{i\mathfrak{T}})$ and $\tilde{\delta}_{\mathfrak{T}\mathfrak{T}} = \frac{1}{n}\sum_{i=1}^n L_1 m'_\mathfrak{T}(Z_{i\mathfrak{T}})m'_\mathfrak{T}(Z_{i\mathfrak{T}})^\top L_1^\top$. Let us also define the vector $MM_\mathfrak{T}(Z_{i\mathfrak{T}}) = \mathrm{vec}(L_1 m'_\mathfrak{T}(Z_{i\mathfrak{T}}), L_1 m'_\mathfrak{T}(Z_{i\mathfrak{T}})m'_\mathfrak{T}(Z_{i\mathfrak{T}})^\top L_1^\top)$. We will study the asymptotic distribution of

$$\begin{aligned}
&\sqrt{n}\left(\mathrm{vec}(\hat{\delta}_\mathfrak{T},\hat{\delta}_{\mathfrak{T}\mathfrak{T}}) - h_n^p\mathrm{vec}(LA^1_\mathfrak{T}, L[A^2_{\mathfrak{T}\mathfrak{T}} + A^{2\top}_{\mathfrak{T}\mathfrak{T}}]L^\top) - \mathrm{E}\,MM_\mathfrak{T}(Z_{i\mathfrak{T}})\right)\\
=&\sqrt{n}\left\{\mathrm{vec}(\hat{\delta}_\mathfrak{T},\hat{\delta}_{\mathfrak{T}\mathfrak{T}}) - \mathrm{vec}(\tilde{\delta}_\mathfrak{T},\tilde{\delta}_{\mathfrak{T}\mathfrak{T}}) - h_n^p\mathrm{vec}(LA^1_\mathfrak{T}, L[A^2_{\mathfrak{T}\mathfrak{T}} + A^{2\top}_{\mathfrak{T}\mathfrak{T}}]L^\top)\right\} \qquad \text{(A.17)}\\
+&\sqrt{n}\left\{\mathrm{vec}(\tilde{\delta}_\mathfrak{T},\tilde{\delta}_{\mathfrak{T}\mathfrak{T}}) - \mathrm{E}\,MM_\mathfrak{T}(Z_{i\mathfrak{T}})\right\}.
\end{aligned}$$

Next, it holds almost surely that

$$\begin{aligned}
&\sqrt{n}\left\{\mathrm{vec}(\hat{\delta}_\mathfrak{T},\hat{\delta}_{\mathfrak{T}\mathfrak{T}}) - \mathrm{vec}(\tilde{\delta}_\mathfrak{T},\tilde{\delta}_{\mathfrak{T}\mathfrak{T}}) - h_n^p\mathrm{vec}(LA^1_\mathfrak{T}, L[A^2_{\mathfrak{T}\mathfrak{T}} + A^{2\top}_{\mathfrak{T}\mathfrak{T}}]L^\top)\right\}\\
=&\sqrt{n}\mathrm{vec}\left(\frac{1}{nh_n}\sum_{i=1}^n[L\hat{\beta}_\mathfrak{T}(Z_{i\mathfrak{T}}) - L\beta_\mathfrak{T}(Z_{i\mathfrak{T}}) - h_n^{p+1}LA^1_\mathfrak{T}],\right.\\
&\qquad\left.\frac{1}{nh_n^2}\sum_{i=1}^n[L\{\hat{\beta}_\mathfrak{T}(Z_{i\mathfrak{T}}) - \beta_\mathfrak{T}(Z_{i\mathfrak{T}})\}\{\hat{\beta}^\top_\mathfrak{T}(Z_{i\mathfrak{T}}) - L\beta^\top_\mathfrak{T}(Z_{i\mathfrak{T}})\}L^\top - h_n^{p+2}L(A^2_{\mathfrak{T}\mathfrak{T}} + A^{2\top}_{\mathfrak{T}\mathfrak{T}})L^\top]\right)\\
=&\sqrt{n}\mathrm{vec}\left(L\bar{J}^1_\mathfrak{T} + o_p(h_n^{1/2}n^{-1/2}) + O_p(h_n^{p+1}),\right.\\
&\qquad\left.L\left\{\bar{J}^2_{\mathfrak{T}\mathfrak{T}} + \bar{J}^{2\top}_{\mathfrak{T}\mathfrak{T}}\right\}L^\top\{1+o_p(1)\} + o_p(h_n^{1/2}n^{-1/2}) + O_p(h_n^{p+1})\right)\\
=&\sqrt{n}\mathrm{vec}\left(L\bar{J}^1_\mathfrak{T}, L\left\{\bar{J}^2_{\mathfrak{T}\mathfrak{T}} + \bar{J}^{2\top}_{\mathfrak{T}\mathfrak{T}}\right\}L^\top\right) + o_p(1) + O_p(n^{1/2}h_n^{p+1})\\
=&\sqrt{n}\mathrm{vec}\left(L\bar{J}^{12}_\mathfrak{T}, L\left\{\bar{J}^{22}_{\mathfrak{T}\mathfrak{T}} + \bar{J}^{22\top}_{\mathfrak{T}\mathfrak{T}}\right\}L^\top\right) + o_p(1),
\end{aligned} \qquad \text{(A.18)}$$

where the first equality follows from the definitions of $\hat{\delta}_\mathfrak{T}(z) = h_n^{-1}L\hat{\beta}_\mathfrak{T}(z)$, $\hat{\delta}_{\mathfrak{T}\mathfrak{T}}(z) = h_n^{-2}L\hat{\beta}_\mathfrak{T}(z)\hat{\beta}_\mathfrak{T}(z)^\top L^\top$ and the latter equalities are due to (A.14), (A.15), and the conditions imposed on the bandwidth h_n by Assumption 9.1. Now define the vector "analog" $GM_\mathfrak{T}(z) = (-[I_{d_y}\otimes\{LG_{\mathfrak{T};\cdot 1}(z)\}]^\top,$ $\iota_{d_y}\mathrm{vec}[-2L_1 m''_\mathfrak{T}(Z_{i\mathfrak{T}})L_1^\top - LG_{\mathfrak{T};\cdot 1}(Z_{i\mathfrak{T}})m'_\mathfrak{T}(Z_{i\mathfrak{T}})^\top L_1^\top - L_1 m'_\mathfrak{T}(Z_{i\mathfrak{T}})G_{\mathfrak{T};\cdot 1}(Z_{i\mathfrak{T}})^\top L^\top]^\top)^\top$ of the summands in (A.11) in Lemma 4 along with $G_{\mathfrak{T};\cdot 1}(z) = G_\mathfrak{T}(z)e_1$ being the first column of $G_\mathfrak{T}(z) = [M^{f_\mathfrak{T}}(z)]^{-1}Q^{f_\mathfrak{T}}(z)$, I_{d_y} the identity matrix, and ι_{d_y} the vector of ones. Also recall that $MM_\mathfrak{T}(Z_{i\mathfrak{T}}) =$

$\text{vec}(L_1 m'_{\mathfrak{T}}(Z_{i\mathfrak{T}}), L_1 m'_{\mathfrak{T}}(Z_{i\mathfrak{T}}) m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top L_1^\top)$. For $n \to \infty$, we obtain

$$\begin{aligned}
&\sqrt{n}\left(\text{vec}(\hat{\delta}_{\mathfrak{T}}, \hat{\delta}_{\mathfrak{T}\mathfrak{T}}) - h_n^p \text{vec}(LA^1_{\mathfrak{T}}, L[A^2_{\mathfrak{T}\mathfrak{T}} + A^{2\top}_{\mathfrak{T}\mathfrak{T}}]L^\top) - \text{E}\, MM_{\mathfrak{T}}(Z_{i\mathfrak{T}})\right) \\
=& -\sqrt{n}\text{vec}\left(L\bar{J}^1_{\mathfrak{T}}, L\left\{\bar{J}^2_{\mathfrak{T}\mathfrak{T}} + \bar{J}^{2\top}_{\mathfrak{T}\mathfrak{T}}\right\}L^\top\right) + o_p(1) \\
&+\sqrt{n}\left\{\text{vec}(\tilde{\delta}_{\mathfrak{T}}, \tilde{\delta}_{\mathfrak{T}\mathfrak{T}}) - \text{E}\,\text{vec}[L_1 m'_{\mathfrak{T}}(Z_{i\mathfrak{T}}), L_1 m'_{\mathfrak{T}}(Z_{i\mathfrak{T}}) m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top L_1^\top]\right\} \\
=& \frac{1}{\sqrt{n}}\sum_{i=1}^n [GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})V_{i\mathfrak{T}} + \{MM_{\mathfrak{T}}(Z_{i\mathfrak{T}}) - \text{E}\, MM_{\mathfrak{T}}(Z_{i\mathfrak{T}})\}] + o_p(1)
\end{aligned} \tag{A.19}$$

by substituting (A.18) into equation (A.17) and using (A.11). Now we apply the multivariate central limit theorem to (A.19) consisting of identically distributed random variables with zero means and variance matrix

$$\begin{aligned}
&\text{Var}\,[GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})V_{i\mathfrak{T}} + \{MM_{\mathfrak{T}}(Z_{i\mathfrak{T}}) - \text{E}\, MM_{\mathfrak{T}}(Z_{i\mathfrak{T}})\}] \\
=&\Phi_{\mathfrak{T}} + \Omega_{\mathfrak{T}} + 2\,\text{Cov}\,[GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})V_{i\mathfrak{T}}; \{MM_{\mathfrak{T}}(Z_{i\mathfrak{T}}) - \text{E}\, MM_{\mathfrak{T}}(Z_{i\mathfrak{T}})\}] \\
=&\Phi_{\mathfrak{T}} + \Omega_{\mathfrak{T}} + 2\,\text{E}\,[GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})V_{i\mathfrak{T}} \times \{MM_{\mathfrak{T}}(Z_{i\mathfrak{T}}) - \text{E}\, MM_{\mathfrak{T}}(Z_{i\mathfrak{T}})\}] \\
=&\Phi_{\mathfrak{T}} + \Omega_{\mathfrak{T}} + 2\,\text{E}\,[GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})\,\text{E}(V_{i\mathfrak{T}}|Z_{i1}, \ldots, Z_{iT})\ \times \{MM_{\mathfrak{T}}(Z_{i\mathfrak{T}}) - \text{E}\, MM_{\mathfrak{T}}(Z_{i\mathfrak{T}})\}] = \Phi_{\mathfrak{T}} + \Omega_{\mathfrak{T}}
\end{aligned}$$

because $\text{Cov}[GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})V_{i\mathfrak{T}}, GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})V_{i\mathfrak{T}}] = \text{E}[GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})\,\text{E}(V_{i\mathfrak{T}}V_{i\mathfrak{T}}^\top|Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}})$ $GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top] = \text{E}[GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})\Sigma_{\mathfrak{T}\mathfrak{T}}(Z_{i\mathfrak{T}})GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top]$. Therefore,

$$\sqrt{n}\left(\hat{\delta}_{\mathfrak{T}} - h_n^p \text{vec}(LA^1_{\mathfrak{T}}, L[A^2_{\mathfrak{T}\mathfrak{T}} + A^{2\top}_{\mathfrak{T}\mathfrak{T}}]L^\top) - \text{E}\, MM_{\mathfrak{T}}(Z_{i\mathfrak{T}})\right) \to N(0, \Phi_{\mathfrak{T}} + \Omega_{\mathfrak{T}})$$

in distribution as $n \to +\infty$. ■

Proof. [Proof of Theorem 3] Let $\tilde{\delta}_{\mathfrak{T}} = \frac{1}{n}\sum_{i=1}^n L_1 m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})$ and $\tilde{\delta}_{\mathfrak{T}\mathfrak{T}} = \frac{1}{n}\sum_{i=1}^n L_1 m'_{\mathfrak{T}}(Z_{i\mathfrak{T}}) m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top L_1^\top$. For $Z_{i\cdot} = \{Z_{i\mathfrak{T}}\}_{\mathfrak{T}\in\mathscr{S}}$, let us also define the vector $MM_{\mathfrak{T}}(Z_{i\mathfrak{T}}) = \text{vec}(L_1 m'_{\mathfrak{T}}(Z_{i\mathfrak{T}}), L_1 m'_{\mathfrak{T}}(Z_{i\mathfrak{T}}) m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top L_1^\top)$ and $MM(Z_{i\cdot}) = \{MM_{\mathfrak{T}}(Z_{i\mathfrak{T}})\}_{\mathfrak{T}\in\mathscr{S}}$. Also $GM_{\mathfrak{T}}(z) = (-[I_{d_y} \otimes \{LG_{\mathfrak{T};\cdot 1}(z)\}]^\top, \iota_{d_y}\text{vec}[-2L_1 m''_{\mathfrak{T}}(Z_{i\mathfrak{T}})L_1^\top - LG_{\mathfrak{T};\cdot 1}(Z_{i\mathfrak{T}}) m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top L_1^\top - L_1 m'_{\mathfrak{T}}(Z_{i\mathfrak{T}}) G_{\mathfrak{T};\cdot 1}(Z_{i\mathfrak{T}})^\top L^\top]^\top)^\top$ (vector $G_{\mathfrak{T};\cdot 1}(z) = [M^{f_{\mathfrak{T}}}(z)]^{-1} Q^{f_{\mathfrak{T}}}(z) e_1$ was defined in Lemma 4, I_{d_y} is the identity matrix, and ι_{d_y} is the vector of ones), $GM(Z_{i\cdot}) = \{GM_1^\top(Z_{i\mathfrak{T}})\}_{\mathfrak{T}\in\mathscr{S}}$, and

$$VGM(V_{i\cdot}, Z_{i\cdot}) = \{V_{i\mathfrak{T}}^\top GM_{\mathfrak{T}}^\top(Z_{i\mathfrak{T}})\}_{\mathfrak{T}\in\mathscr{S}}.$$

Finally, let $\hat{\delta} = \{\text{vec}(\hat{\delta}_{\mathfrak{T}}, \hat{\delta}_{\mathfrak{T}\mathfrak{T}})\}_{\mathfrak{T}\in\mathscr{S}}$ and $Bias = \{MM_{\mathfrak{T}}(Z_{i\mathfrak{T}}) + h_n^p \text{vec}(LA^1_{\mathfrak{T}}, L[A^2_{\mathfrak{T}\mathfrak{T}} + A^{2\top}_{\mathfrak{T}\mathfrak{T}}]L^\top)\}_{\mathfrak{T}\in\mathscr{S}}$. Since equation (A.19) holds for any fixed $\mathfrak{T} \in \mathscr{S}$, we have

$$\sqrt{n}(\hat{\delta} - Bias) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [VGM(V_{i\cdot}, Z_{i\cdot}) + \{MM(Z_{i\cdot}) - \mathrm{E}\, MM(Z_{i\cdot})\}] + o_p(1). \tag{A.20}$$

Note that $VGM(V_{i\mathfrak{T}}, Z_{i\mathfrak{T}})$, $\{MM_{\mathfrak{T}}(Z_{i\mathfrak{T}}) - \mathrm{E}\, MM_{\mathfrak{T}}(Z_{i\mathfrak{T}})\}$, and their sums are identically distributed random variables with zero means and finite second moments by Assumption 9. Hence, we can use the multivariate central theorem to obtain the asymptotic distribution of $n^{-1/2}\sum_{i=1}^{n}[VGM(V_{i\cdot}, Z_{i\cdot}) + \{MM(Z_{i\cdot}) - \mathrm{E}\, MM(Z_{i\cdot})\}]$ in (A.20). Since the covariance of the elements of $VGM(V_{i\cdot}, Z_{i\cdot})$ corresponding to the $\mathfrak{S}$th and $\mathfrak{T}$th pairs of time periods is

$$\begin{aligned}\mathrm{Cov}[V_{i\mathfrak{T}} GM_{\mathfrak{T}}(Z_{i\mathfrak{T}}), V_{i\mathfrak{S}} GM_{\mathfrak{S}}(Z_{i\mathfrak{S}})] &= \mathrm{E}[\mathrm{E}(V_{i\mathfrak{T}} V_{i\mathfrak{S}} | Z_{i\mathfrak{T}}, Z_{i\mathfrak{S}}) GM_{\mathfrak{T}}(Z_{i\mathfrak{T}}) GM_{\mathfrak{S}}(Z_{i\mathfrak{S}})^\top] \\ &= \mathrm{E}[\mathrm{E}[\Sigma_{\mathfrak{T}\mathfrak{S}}(Z_{i\mathfrak{T}}, Z_{i\mathfrak{S}}) GM_{\mathfrak{T}}(Z_{i\mathfrak{T}}) GM_{\mathfrak{S}}(Z_{i\mathfrak{S}})^\top] = \Phi^{(\mathfrak{T},\mathfrak{S})},\end{aligned}$$

$\mathrm{Var}(VGM_i) = \Phi$. Furthermore, we have $\mathrm{Cov}\{MM_{\mathfrak{T}}(Z_{i\mathfrak{T}}) - \mathrm{E}\, MM_{\mathfrak{T}}(Z_{i\mathfrak{T}}), MM_{\mathfrak{S}}(Z_{it,\mathfrak{S}}) - \mathrm{E}\, MM_{\mathfrak{S}}(Z_{it,\mathfrak{S}})\} = \Omega^{(\mathfrak{T},\mathfrak{S})}$ and thus $\mathrm{Var}\{MM(Z_{it,\cdot}) - \mathrm{E}\, MM(Z_{it,\cdot})\} = \Omega$. Since for any $\mathfrak{S}, \mathfrak{T} \in \mathscr{S}$

$$\begin{aligned}&\mathrm{Cov}(VGM(V_{i\mathfrak{T}}, Z_{i\mathfrak{T}}), MM_{\mathfrak{S}}(Z_{i\mathfrak{S}}) - \mathrm{E}\, MM_{\mathfrak{S}}(Z_{i\mathfrak{S}})) \\ =& \mathrm{E}[VGM(V_{i\mathfrak{T}}, Z_{i\mathfrak{T}})\{MM_{\mathfrak{S}}(Z_{i\mathfrak{S}}) - \mathrm{E}\, MM_{\mathfrak{S}}(Z_{i\mathfrak{S}})\}^\top] \\ =& \mathrm{E}[GM_{\mathfrak{T}}(Z_{i\mathfrak{T}}) V_{i\mathfrak{T}}\{MM_{\mathfrak{S}}(Z_{i\mathfrak{S}}) - \mathrm{E}\, MM_{\mathfrak{S}}(Z_{i\mathfrak{S}})\}^\top] \\ =& \mathrm{E}[GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})\, \mathrm{E}(V_{i\mathfrak{T}}|Z_{i\cdot})\{MM_{\mathfrak{S}}(Z_{i\mathfrak{S}}) - \mathrm{E}\, MM_{\mathfrak{S}}(Z_{i\mathfrak{S}})]\}^\top] \\ =& 0,\end{aligned}$$

we obtain $\mathrm{Var}[VGM(V_{i\cdot}, Z_{i\cdot}) + \{MM(Z_{i\cdot}) - \mathrm{E}\, MM(Z_{i\cdot})\}] = \Phi + \Omega$. Therefore, $\sqrt{n}(\hat{\delta} - Bias) \to N(0, \Phi + \Omega)$ in distribution as $n \to +\infty$. ■

Proof. [Proof of Theorem 4] Assuming that $W_n \to W$ in probability for $n \to \infty$ is such that $\theta_0 = \text{argmin}_\theta g(\theta)^\top W g(\theta)$ and θ_0 is in the interior of Θ, where $g(\theta) = \{g_{\mathfrak{T}\mathfrak{T}}(\theta)\}_{\mathfrak{T}\in\mathscr{S}}$, the consistency of $\hat{\theta}_n$ follows from Theorem 3 and Newey and McFadden (Theorem 2.1). Then the first order conditions are

$$\frac{\partial \hat{g}_n^\top(\hat{\theta}_n)}{\partial \theta} W_n \hat{g}_n(\hat{\theta}_n) = 0,$$

The Taylor expansion of $\hat{g}_n(\hat{\theta}_n)$ further leads to

$$\frac{\partial \hat{g}_n^\top(\hat{\theta}_n)}{\partial \theta^\top} W_n \left[\hat{g}_n(\theta_0) + \frac{\partial \hat{g}_n(\tilde{\theta}_n)}{\partial \theta^\top}(\hat{\theta}_n - \theta_0) \right] = 0,$$

where $\tilde{\theta}_n \in \left[\theta_0, \hat{\theta}_n\right]$ lies in a multidimensional interval. Since $\tilde{\theta}_n \to \theta_0$ in probability, $\partial \hat{g}_n(\tilde{\theta}_n)/\partial \theta^\top \to \partial g(\theta_0)/\partial \theta^\top = \Pi$ since this derivative is differentiable and does not depend on the data.

Since $\Pi^\top W \Pi$ is a full rank matrix, it follows that

$$\Pi^\top W \left[\hat{g}_n(\theta_0) + \Pi(\hat{\theta}_n - \theta_0) \right] = o_p(\hat{g}_n(\theta_0)) + o_p(\hat{\theta}_n - \theta_0)$$

and

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \left(\Pi^\top W \Pi \right)^{-1} \Pi^\top W \sqrt{n} \hat{g}_n(\theta_0)(1 + o_p(1)). \tag{A.21}$$

Since $\sqrt{n}\hat{g}_n(\theta_0) = \{\text{vec}(\hat{\delta}_{\mathfrak{T}}, \hat{\delta}_{\mathfrak{T}\mathfrak{T}})^\top - \text{vec}(\Gamma_{1t(\mathfrak{T})} B, B\Gamma_{2t(\mathfrak{T})} B^\top)^\top\}_{\mathfrak{T} \in \mathscr{S}}$ for $t(\mathfrak{T})$ representing the later time period in $\mathfrak{T}$, Theorem 3 implies that

$$\sqrt{n}(\hat{g}_n(\theta_0) - h_n^p Bias) \to N(0, \Omega + \Phi)$$

in distribution, where $\Omega + \Phi$ represents the asymptotic variance of $\hat{\delta} = \{\text{vec}(\hat{\delta}_{\mathfrak{T}}, \hat{\delta}_{\mathfrak{T}\mathfrak{T}})\}_{\mathfrak{T} \in \mathscr{S}}$. Hence,

$$\sqrt{n}(\hat{\theta}_n - \theta_0 - h_n^p \left(\Pi^\top W \Pi \right)^{-1} \Pi^\top W Bias) \to N\left(0, \left(\Pi^\top W \Pi \right)^{-1} \Pi^\top W (\Omega + \Phi) W \Pi \left(\Pi^\top W \Pi \right)^{-1} \right)$$

in distribution as $n \to \infty$. ■

References

[1] Abrevaya, J. (2000). Rank estimation of a generalized fixed-effects regression model. *Journal of Econometrics* 95, 1–23.

[2] Abrevaya, J., and S. Shen (2014). Estimation of censored panel-data with slope heterogeneity. *Journal of Applied Econometrics* 29, 523–548.

[3] Arcones, M. A. (1995). A Bernstein-type inequaltiy for U-statistics and U-processes. *Statistics and Probability Letters 22 (3),* 239–247.

[4] Bester, C., and C. Hansen (2009). Idenitification of marginal effects in a non-parametric correlated random effects model. *Journal of Business and Economics Statistics* 27 (2), 235–250.

[5] Boll, C., Leppin, J. S. and K. Schömann (2016). Who is overeducated and why? Probit and dynamic mixed multinomial logit analyses of vertical mismatch in East and West Germany. *Education Economics* 24(6), 639–662.

[6] Botosaru, I. and C. Muris (2017). Binarization for panel models with fixed effects. Cemmap working paper No. CWP31/17.

[7] Carroll, R. J., Fan, J., Gijbels, I. and P. M. Wand (1997). Generalized partially linear single index models. *Journal of the American Statistical Associatio*n 92 (438), 477–489.

[8] Cattaneo, M. D., Crump, R. K. and M. Jansson (2013). Generalized jackknife estimators of weighted average derivatives. *Journal of the American Statistical Associatio*n 108 (504), 1243–1256.

[9] Chamberlein, G. (1982) Multivariate regression models for panel data. *Journal of Econometrics* 18, 5–46.

[10] Charlier, E., Melenberg, B. and A. H. O. van Soest (1995). A smoothed maximum score estimator for the binary choice panel data model with an application to labour force participation. *Statistica Neerlandica* 49 (3), 324–342.

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

[12] Chen, Q. and Y. Fang (2019). Improved inference on the rank of a matrix. *Quantitative Economics* 10, 1787–1824.

[13] Chen, S. and X. Wang (2018). Semiparametric estimation of panel data models without monotonicity or separability. *Journal of Econometrics* 206, 515–530.

[14] Chernozhukov, V., Fernandez-Val, I., Hahn, J. and W. Newey (2013). Average and quantile effects in nonseparable panel models. *Econometrica* 81 (2), 535–580.

[15] Chernozhukov, V., Fernández-Val, I., Hoderlein, S., Holzmann, H., and W. K. Newey (2015). Nonparametric identification in panels using quantiles. *Journal of Econometrics* 188(2), 378–392.

[16] Chernozhukov, V., Fernández-Val, I. and W. K. Newey (2019). Nonseparable multinomial choice models in cross-section and panel data. Journal *of Econometrics* 211(1), 104–116.

[17] Čížek, P. and J. Lei (2018). Identification and estimation of nonseparable single-index models in panel data with correlated random effects. *Journal of Econometrics* 203 (1), 113–128.

[18] Christelis, D. and A. Sanz-de-Galdeano (2011). Smoking persistence across countries: a panel data analysis. *Journal of Health Economics* 30(5), 1077–1093.

[19] Donkers, B. and M. Schafgans (2008). Specification and estimation of semiparametric multiple-index models. *Econometric Theory* 24 (6), 1584–1606.

[20] Escanciano, J. C., Jacho-Chavez, D. and A. Lewbel (2016). Identification and estimationof semiparametric two-step models. *Quantitative Economics* 7 (2), 561–589.

[21] Fox, J. T. (2007). Semiparametric estimation of multinomial discrete-choice models using a subset of choices. *The RAND Journal of Economics* 38 (4), 1002–1019.

[22] Freyberger, J. (2018). Non-parametric panel data models with interactive fixed effects. *Review of Economic Studies* 85, 1824–1851.

[23] Fukumizu, K. and C. Leng (2014). Gradient-based kernel dimension reduction for regression. *Journal of the American Statistical Associatio*n 109 (505), 359–370.

[24] Gao, W. Y. and M. Li (2020). Robust semiparametric estimation in panel multinomial choice models. arXiv:2009.00085 [Econ EM].

[25] Gayle, G.-L. and C. Viauroux (2007). Root-N consistent semiparametric estimators of a dynamic panel-sample-selection model. *Journal of Econometrics* 141(1), 179–212.

[26] Ghanem, D. (2017). Testing identifying assumptions in nonseparable panel data models. *Journal of* Econometrics 197 (2), 202–217.

[27] Härdle, W., and T. M. Stoker (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association 84* (408), 986–995.

[28] Horowitz, J. L., and W. Härdle (1996). Direct semiparametric estimation of single-index mo dels with discrete covariates. *Journal of the American Statistical Association* 91, 1632–1640.

[29] Hoderlein, S. and H. White (2012). Nonparametric identification in nonseparable panel data models with generalized fixed effects. *Journal of Econometrics* 168 (2), 300–314.

[30] Honore, B. E., and L. Hu (2002). Estimation of cross-sectional and panel data censored regression models with endogeneity. *Journal of Econometrics* 122, 293–316.

[31] Hristache, M., Juditsky, A., Polzehl, J. and V. Spokoiny (2001). Structure adaptive approach for dimension reduction. *The Annals of Statistics* 29 (6), 1537–1566.

[32] Ishihara, T. (2020). Identification and estimation of time-varying nonseparable panel data models without stayers. *Journal of Econometrics* 215, 184–208.

[33] Klein, R., Shen, C. and F. Vella (2015). Estimation of marginal effects in semiparametric selection models with binary outcomes. *Journal of Econometrics* 185(1), 82–94.

[34] Kong, E., Linton, O. and Y. Xia (2010). Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model. *Econometric Theory 26 (5),* 1529–1564.

[35] Kong, E. and Y. Xia (2014). An adaptive composite quantile approach to dimension reduction. *The Annals of Statistics* 42 (4), 1657–1688.

[36] Kyriazidou, E. (1997). Estimation of a panel data sample selection model. *Econometrica* 65 (6), 1335–1364.

[37] Kyriazidou, E. (2001). Estimation of dynamic panel data sample selection models. *Review of Economic Studies* 68, 543–572.

[38] Lechmann, S. J. and C. Wunder (2017). The dynamics of solo self-employment: persistence and transition to employership. *Labour Economics* 49, 95–105.

[39] Li, K-C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association 87* (420), 1025–1039.

[40] Li, Q., Lu, X. and A. Ullah (2003). Multivariate local polynomial regression for estimating average derivatives. *Journal of Nonparametric Statistics* 15 (4-5), 607–624.

[41] Manski, C. F. (1987). Semiparametric analysis of random effects linear models from binary panel data. *Econometrica* 55 (2), 357–362.

[42] Masry, E. (1996). Multivariate local polynomial regression for time-series: uniform strong consistency and rates. *Journal of Time Series Analysis* 17 (6), 571–599.

[43] Mundlak, Y. (1978). On the Pooling of Time Series and Cross Section Data. *Econometrica* 46(1), 69–85.

[44] Newey, W. and T. Stoker (1993). Efficiency of weighted average derivative estimatrs and index models. *Econometrica* 61 (5), 1199–1223.

[45] Racine, J. and Q. Li (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics* 119(1), 99–130.

[46] Samarov, A. M. (1993). Exploring regression structure using nonparametric functional estimation. *Journal of the American Statistical Association 88* (423), 836–847.

[47] Schulz, R., Wersing, M. and A. Werwatz (2014). Renting versus owning and the role of human capital: evidence from Germany. *Journal of Real Estate Finance and Economics* 48, 754–788.

[48] Semykina, A. and J. M. Wooldridge (2013). Estimation of dynamic panel data models with sample selection. *Journal of Applied Econometrics 28* (1), 47–61.

[49] Semykina, A. and J. M. Wooldridge (2018). Binary response panel data models with sample selection and self selection. *Journal of Applied Econometrics* 33, 179–197.

[50] Shi, X., Shum, M. and W. Song (2018). Estimating semiparametric panel multinomial choice models using cyclic monotonicity. *Econometrica* 86, 737–761.

[51] Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*, Second Edition. MIT Press. London, UK.

[52] Xia, Y., Tong, H., Li, W. K. and L-X. Zhu (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B 64* (3), 363–410.

[53] Xu, K., Guo, W., Xiong, M., Zhu, L. and L. Jin (2016). An estimating equation approach to dimension redcution for longitudinal data. *Biometrika 103* (1), 189–203.

[54] Zhu, L., Miao, B. and H. Peng (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association 101* (474), 630–643.

Supplemental material for Nonseparable panel models with index structure and correlated random effects

Pavel Čížek Serhan Sadikoğlu

This supplement contains several sections. Estimation with more than two time periods is discussed in Appendix D along with the corresponding simulation results, the proofs of auxiliary lemmas can be found in Appendix E, the generalized jackknife procedure for the bias reduction is described in Appendix F, the estimation with discrete variables is in Appendix G, and the moment conditions used in the application of Section 5 for estimation with functionally related regressors such as quadratic and interaction terms are introduced in Appendix H. Finally, the identification assumption for the dynamic sample selection model is verified in Appendix I.

D Estimation with $T > 2$ time periods

Let us consider panel data with a fixed number $T > 2$ of time periods. The GMM estimation procedure was introduced in Section 3 using two time periods t and $t - \Delta$. Since these two time periods are arbitrary, it is possible to construct the proposed moment conditions for any pair of time periods $\mathfrak{T} = (t, t - \Delta)$, where $t = \Delta + 1, \ldots, T$ and $\Delta = 1, \ldots, T$, and use jointly all moment conditions constructed for these pairs of time periods. This is facilitated by the asymptotic distribution in Theorem 3 obtained for any set of time-period pairs $\mathfrak{T} = (t, t - \Delta)$. Such a procedure is suitable especially if the cross-sectional sample size n is relatively large and sufficient to obtain nonparametric estimates of ADG and OPDG in Theorem 1.

On the other hand, if the cross-sectional sample size n is relatively small, while the panel has a large number T of time periods, it might be preferable to pool the estimation across all available time periods and estimate $E(Y_{it}|X_{it}, X_{i(t-\Delta)})$ and its derivatives for a given Δ using observations $(Y_{it}, X_{it}, X_{i(t-\Delta)})$ for all $i = 1, \ldots, n$ and $t = \Delta + 1, \ldots, T$. For a given Δ, this pooling of data from all time periods results in one set of moment conditions irrespective of the number of time periods. Without further assumptions, this pooling does not lead to valid moment conditions in general model (4) since the regression function is allowed to change with the time period t. The pooling is however possible if certain stationarity properties are imposed to guarantee that matrices in Theorem 1 are time-invariant – $\Gamma_{1t} = \Gamma_1$ and $\Gamma_{2t} = \Gamma_2$ (B does not change over time). Since $\Gamma_{1,t} = \left\{\mathrm{E}\left[\varphi'_{tr}(X_{it}^\top B, \alpha_i)\right]\right\}_{r=1}^{R}$ and $\Gamma_{2,t} = \left\{\mathrm{E}\left[\varphi'_{tr}(X_{it}^\top B, \alpha_i)^\top \varphi'_{ts}(X_{it}^\top B, \alpha_i)\right]\right\}_{r,s=1}^{R}$, $\Gamma_{1t} = \Gamma_1$ and $\Gamma_{2t} = \Gamma_2$ means that

(a) the regression function $\phi_t = \phi$ should not change over time and model (4) becomes

$$Y_{it} = \phi(X_{it}^\top \beta_1, \ldots, X_{it}^\top \beta_R, \alpha_i, U_{it}) = \phi(X_{it}^\top B, \alpha_i, U_{it});$$

(b) the joint distributions $F_{\alpha, X_t, X_{t-\Delta}}$ of $(\alpha_i, X_{it}, X_{i(t-\Delta)})$ are identical for all $i \in \mathbb{N}$ and $\Delta < t \leq T$.

These specific types of stationarity conditions can be tested as discussed in Čížek and Lei (2018) and Ghanem (2017), and if (a) and (b) are acceptable, it is possible to pool and use data from all time periods for the same moment conditions.

Using the described pooling, the additional simulation results for data sets with more than two time periods are reported for the binary partially-linear single-index model (Table A.5), the

heteroskedastic censored-regression model (Table A.6), and the sample selection model (Table A.7). All considered sample sizes $(n, T) = (1000, 2), (250, 5), (100, 11)$ result after differencing in the same effective sample size of 1000 first-order differences ($\Delta = 1$). The proposed GMM estimator for $T > 2$ is also reported for various higher orders of differencing Δ: for $T = 2$, only $\Delta = 1$ is possible; for $T = 5$, $\Delta = 1, 2, 3$ is used; and for $T = 11$; $\Delta = 1, 2, 3, 4, 5$ is used.

Table A.5: The bias and RMSE of all estimators in the binary partially linear single-index model for the different sample sizes.

	β_{13}		β_{23}	
	Bias	RMSE	Bias	RMSE
		$n = 1000$, $T = 2$		
FE Logit	0.060	0.135		
GMM-OPDG				
$\Delta = 1$	-0.004	0.172	0.012	0.210
GMM-ADG-OPDG				
$\Delta = 1$	-0.006	0.174	0.012	0.210
SMS	-0.079	0.268		
		$n = 250$, $T = 5$		
FE Logit	0.054	0.104		
GMM-OPDG				
$\Delta = 1$	-0.015	0.155	0.021	0.232
$\Delta \leq 2$	-0.011	0.134	0.016	0.204
$\Delta \leq 3$	-0.011	0.132	0.015	0.199
GMM-ADG-OPDG				
$\Delta = 1$	-0.016	0.153	0.020	0.235
$\Delta \leq 2$	-0.012	0.135	0.018	0.207
$\Delta \leq 3$	-0.013	0.138	0.018	0.205
SMS	-0.013	0.140		
		$n = 100$, $T = 11$		
FE Logit	0.054	0.103		
GMM-OPDG				
$\Delta = 1$	-0.014	0.148	0.026	0.224
$\Delta \leq 2$	-0.012	0.127	0.029	0.194
$\Delta \leq 3$	-0.011	0.119	0.032	0.187
$\Delta \leq 4$	-0.011	0.115	0.031	0.179
$\Delta \leq 5$	-0.012	0.115	0.032	0.177
GMM-ADG-OPDG				
$\Delta = 1$	-0.016	0.147	0.027	0.224
$\Delta \leq 2$	-0.014	0.128	0.029	0.195
$\Delta \leq 3$	-0.013	0.119	0.032	0.187
$\Delta \leq 4$	-0.013	0.115	0.031	0.180
$\Delta \leq 5$	-0.015	0.117	0.032	0.178
SMS	-0.001	0.123		

Table A.6: The bias and RMSE of all estimators in the heteroskedastic censored regression model for the different sample sizes.

	β_{13}		β_{23}	
	Bias	RMSE	Bias	RMSE
		$n = 1000$, $T = 2$		
Pooled Tobit	1.125	1.137		
TLS	0.746	13.535		
GMM-OPDG				
$\Delta = 1$	-0.023	0.256	0.053	0.287
GMM-ADG-OPDG				
$\Delta = 1$	-0.023	0.255	0.053	0.291
SMS	-0.090	0.368		
		$n = 250$, $T = 5$		
Pooled Tobit	1.129	1.147		
TLS	-0.072	28.249		
GMM-OPDG				
$\Delta = 1$	-0.008	0.244	0.041	0.288
$\Delta \leq 2$	-0.007	0.215	0.031	0.255
$\Delta \leq 3$	-0.008	0.202	0.031	0.248
GMM-ADG-OPDG				
$\Delta = 1$	-0.011	0.241	0.040	0.287
$\Delta \leq 2$	-0.006	0.215	0.030	0.258
$\Delta \leq 3$	-0.009	0.206	0.032	0.253
SMS	-0.024	0.210		
		$n = 100$, $T = 11$		
Pooled Tobit	1.132	1.153		
TLS	0.887	42.477		
GMM-OPDG				
$\Delta = 1$	-0.032	0.247	0.036	0.277
$\Delta \leq 2$	-0.020	0.210	0.036	0.237
$\Delta \leq 3$	-0.018	0.197	0.037	0.222
$\Delta \leq 4$	-0.016	0.192	0.037	0.222
$\Delta \leq 5$	-0.015	0.190	0.039	0.220
GMM-ADG-OPDG				
$\Delta = 1$	-0.035	0.248	0.035	0.278
$\Delta \leq 2$	-0.021	0.207	0.035	0.237
$\Delta \leq 3$	-0.019	0.196	0.037	0.223
$\Delta \leq 4$	-0.018	0.192	0.037	0.223
$\Delta \leq 5$	-0.016	0.191	0.039	0.222
SMS	0.006	0.153		

Table A.7: The bias and RMSE of all estimators in the sample selection model for the different sample sizes.

	β_{12}		β_{13}		β_{23}	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
			$n = 1000$, $T = 2$			
FE Logit+KYR	0.009	0.123	0.002	0.121	0.001	0.115
GMM-OPDG						
$\Delta = 1$	0.022	0.142	0.039	0.242	0.003	0.102
GMM-ADG-OPDG						
$\Delta = 1$	0.008	0.109	0.015	0.157	0.003	0.102
			$n = 250$, $T = 5$			
FE Logit+KYR	0.004	0.079	0.001	0.083	0.001	0.096
GMM-OPDG						
$\Delta = 1$	0.021	0.123	0.23	0.196	0.009	0.092
$\Delta \leq 2$	0.015	0.111	0.029	0.174	0.009	0.087
$\Delta \leq 3$	0.016	0.111	0.033	0.176	0.010	0.091
GMM-ADG-OPDG						
$\Delta = 1$	0.009	0.104	0.010	0.135	0.008	0.085
$\Delta \leq 2$	0.007	0.094	0.009	0.124	0.008	0.083
$\Delta \leq 3$	0.009	0.098	0.016	0.136	0.010	0.092
			$n = 100$, $T = 11$			
FE Logit+KYR	0.005	0.065	0.003	0.066	0.007	0.077
GMM-OPDG						
$\Delta = 1$	0.018	0.136	0.031	0.202	0.004	0.089
$\Delta \leq 2$	0.010	0.106	0.025	0.149	0.006	0.078
$\Delta \leq 3$	0.009	0.100	0.024	0.140	0.006	0.077
$\Delta \leq 4$	0.009	0.101	0.027	0.142	0.007	0.077
$\Delta \leq 5$	0.009	0.101	0.029	0.143	0.007	0.078
GMM-ADG-OPDG						
$\Delta = 1$	0.010	0.110	0.015	0.129	0.002	0.082
$\Delta \leq 2$	0.007	0.097	0.014	0.111	0.004	0.072
$\Delta \leq 3$	0.007	0.091	0.013	0.106	0.004	0.072
$\Delta \leq 4$	0.007	0.091	0.015	0.107	0.005	0.073
$\Delta \leq 5$	0.007	0.092	0.015	0.108	0.006	0.074

E Proofs of auxiliary lemmas

In this appendix, we derive the properties of the local polynomial estimator (9) and the corresponding averages (10) and (11). As this estimation is performed for each response variable separately, we assume for simplicity of notation in this section that Y_{it} represents one particular (scalar) response variable, denoted in the main text as $Y_{c,it}$ for $c \in \{1, \dots, d_y\}$. The same notation, omitting the subscript indicating the response component considered, is then applied also to the conditional expectations $m_{\mathfrak{T}}(z)$, their derivative differences $\delta_{\mathfrak{T}}(z)$, and residuals $V_{i\mathfrak{T}}$. Given one particular response, we need to introduce notation, which is analogous to Čížek and Lei (2018), Masry (1996), and Li et al. (2003). We also introduce results of some theorems and arguments in Čížek and Lei (2018), Li et al. (2003), and Masry (1996) as the assumptions of those theorems are included in Assumptions 1–9. Thus, we first develop convenient notation, and then we prove the auxiliary lemmas required for the proofs of Theorem 2–4, which follow in Appendix C.

Since $m_{\mathfrak{T}}(\cdot)$ is $(p+1)$-times differentiable with uniformly bounded derivatives by Assumption 9, $m_{\mathfrak{T}}(z)$ can be locally approximated at a point z_0 by a polynomial of order p:

$$m_{\mathfrak{T}}(z) \approx \sum_{0 \leq |\underline{k}| \leq p} \frac{1}{\underline{k}!} D^{\underline{k}} m_{\mathfrak{T}}(v)|_{v=z_0} (z - z_0)^{\underline{k}},$$

where $\underline{k} = (k_1, \dots, k_{2d}) \in \mathbb{N}^{2d}$, $\underline{k}! = k_1! \times \cdots \times k_{2d}!$, $|\underline{k}| = \sum_{i=1}^{2d} k_i$, $z^{\underline{k}} = z_1^{k_1} \times \cdots \times z_{2d}^{k_{2d}}$,

$$\sum_{0 \leq |\underline{k}| \leq p} = \sum_{j=0}^{p} \sum_{k_1=0}^{j} \cdots \sum_{k_{2d}=0; k_1 + \cdots + k_{2d} = j}, \quad \text{and} \quad (D^{\underline{k}} m_{\mathfrak{T}})(z) = \frac{\partial^{\underline{k}} m_{\mathfrak{T}}(z)}{\partial z_1^{k_1} \dots \partial z_{2d}^{k_{2d}}}.$$

Further, define for $\underline{j} = (j_1, \dots, j_{2d}) \in \mathbb{N}^{2d}$ and $z \in \mathbb{R}^{2d}$

$$\begin{aligned} \bar{\tau}^e_{\mathfrak{T},\underline{j}}(z) &= \frac{1}{n} \sum_{i=1}^{n} \left(Y_{it} - m_{\mathfrak{T}}(Z_{i\mathfrak{T}})\right) \left(\frac{Z_{i\mathfrak{T}} - z}{h_n}\right)^{\underline{j}} K_h\left(Z_{i\mathfrak{T}} - z\right) \\ &= \frac{1}{n} \sum_{i=1}^{n} V_{i\mathfrak{T}} \left(\frac{Z_{i\mathfrak{T}} - z}{h_n}\right)^{\underline{j}} K_h\left(Z_{i\mathfrak{T}} - z\right), \end{aligned} \tag{S.1}$$

and

$$\bar{s}^e_{\mathfrak{T},\underline{j}}(z) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Z_{i\mathfrak{T}} - z}{h_n}\right)^{\underline{j}} K_h\left(Z_{i\mathfrak{T}} - z\right), \tag{S.2}$$

where $V_{i\mathfrak{T}} = Y_{it} - m_{\mathfrak{T}}(Z_{i\mathfrak{T}})$ and $K_h(u) = h_n^{-2d} K(u/h_n)$. We indicate the dependence of $\bar{\tau}^e_{\mathfrak{T},j}$, $\bar{s}^e_{\mathfrak{T},j}(z)$, and other averages on the sample size n by the bar above the letters for notational convenience. As we take a "short" T approach, the asymptotic expressions should be understood for $n \to +\infty$ with $T > 1$ being fixed in what follows.

As the moment conditions of the proposed GMM estimator are based on local polynomial estimation, we first explore the properties of local polynomial estimator by means of a Bahadur-type representation following Masry (1996) and Li et al. (2003). For that purpose, we adapt their conventional notation to our setting. We first express $\bar{\tau}^e_{\mathfrak{T},\underline{j}}$ in a matrix form in a lexicographical order. Let $N_l = (l+2d-1)!/(l!(2d-1)!)$ be the number of distinct $2d$-tuples with $|\underline{j}| \equiv j_1 + \cdots + j_{2d} = l$. In local polynomial estimation problem, N_l denotes the number of the distinct lth order partial derivatives of $m_{\mathfrak{T}}(z)$. We arrange these $2d$-tuples as a sequence in a lexicographical order. The highest priority is given to the last position so that $(0,\ldots,0,l)$ is the first element in the sequence and $(l,0,\ldots,1)$ is the last element. Let $g_l^{-1} \equiv g^{-1}_{|\underline{j}|}$ denote this one-to-one mapping. We arrange the $N_l = N_{|\underline{j}|}$ values of $\bar{\tau}^e_{\mathfrak{T},\underline{j}}(z)$ in a column vector $\bar{\tau}_{\mathfrak{T},l}(z) = (\bar{\tau}^e_{\mathfrak{T},g_l(k)}(z))_{k=1}^{N_l}$ in the lexicographical order. We further define the column vector $\bar{\tau}_{\mathfrak{T}}(z) = (\bar{\tau}^\top_{\mathfrak{T},0}(z), \bar{\tau}^\top_{\mathfrak{T},1}(z), \ldots, \bar{\tau}^\top_{\mathfrak{T},p}(z))^\top$, where $\bar{\tau}_{\mathfrak{T},l}(z)$ is an $N_l \times 1$ vector with elements $\bar{\tau}^e_{\mathfrak{T},\underline{j}}(z)$, $|\underline{j}| = l$, arranged according to the lexicographical order; $\bar{\tau}_{\mathfrak{T}}(z)$ has thus dimension $N \times 1$ with $N = \sum_{l=0}^p N_l$.

Furthermore, by arranging $\bar{s}^e_{\mathfrak{T},\underline{j}+\underline{k}}(z)$ in a matrix $\bar{S}_{\mathfrak{T},|\underline{j}|,|\underline{k}|}(z)$ in the lexicographical order with the (l_1,l_2)th element given by $[\bar{S}_{\mathfrak{T},|\underline{j}|,|\underline{k}|}(z)]_{l_1 l_2} = \bar{s}^e_{\mathfrak{T},g_{|\underline{j}|}(l_1)+g_{|\underline{k}|}(l_2)}(z)$, we define the $N \times N$ matrix $\bar{S}_{\mathfrak{T}}(z)$ and $N \times N_{p+1}$ matrix $\bar{B}_{\mathfrak{T}}(z)$ by

$$\bar{S}_{\mathfrak{T}}(z) = \begin{pmatrix} \bar{S}_{\mathfrak{T},0,0}(z) & \bar{S}_{\mathfrak{T},0,1}(z) & \ldots & \bar{S}_{\mathfrak{T},0,p}(z) \\ \bar{S}_{\mathfrak{T},1,0}(z) & \bar{S}_{\mathfrak{T},1,1}(z) & \ldots & \bar{S}_{\mathfrak{T},1,p}(z) \\ \vdots & \vdots & \ddots & \vdots \\ \bar{S}_{\mathfrak{T},p,0}(z) & \bar{S}_{\mathfrak{T},p,1}(z) & \ldots & \bar{S}_{\mathfrak{T},p,p}(z) \end{pmatrix} \quad \text{and} \quad \bar{B}_{\mathfrak{T}}(z) = \begin{pmatrix} \bar{S}_{\mathfrak{T},0,p+1}(z) \\ \bar{S}_{\mathfrak{T},1,p+1}(z) \\ \vdots \\ \bar{S}_{\mathfrak{T},p,p+1}(z) \end{pmatrix}.$$

Let $\mu_{\underline{j}} = \int_{\mathbb{R}^{2d}} u^{\underline{j}} K(u) du$ and $v_{s,\underline{j}} = \int_{\mathbb{R}^{2d}} u_s u^{\underline{j}} K(u) du$, where u_s is the sth element of vector u. Then we define $N_i \times N_j$ dimensional matrices $M_{i,j}$ and $Q_{s,i,j}$ to have their (l_1,l_2)th elements

given by $\mu_{g_i(l_1)+g_j(l_2)}$ and $v_{s,g_i(l_1)+g_j(l_2)}$, respectively, for $s = 1, \ldots, 2d$, and we further define

$$M = \begin{pmatrix} M_{0,0} & M_{0,1} & \ldots & M_{0,p} \\ M_{1,0} & M_{1,1} & \ldots & M_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ M_{p,0} & M_{p,1} & \ldots & M_{p,p} \end{pmatrix}, B = \begin{pmatrix} M_{0,p+1} \\ M_{1,p+1} \\ \vdots \\ M_{p,p+1} \end{pmatrix}, Q_s = \begin{pmatrix} Q_{s,0,0} & Q_{s,0,1} & \ldots & Q_{s,0,p} \\ Q_{s,1,0} & Q_{s,1,1} & \ldots & Q_{s,1,p} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{s,p,0} & Q_{s,p,1} & \ldots & Q_{s,p,p} \end{pmatrix}.$$

Let $f'_{\mathfrak{T},s}(z)$ denote the sth element of the first derivative $f'_{\mathfrak{T}}(z)$ of the density function $f_{\mathfrak{T}}(z)$ of $Z_{i\mathfrak{T}}$, $s = 1, \ldots, 2d$, and define $M^{f_{\mathfrak{T}}}(z) = Mf_{\mathfrak{T}}(z)$, $Q^{f_{\mathfrak{T}}}(z) = \sum_{s=1}^{2d} f'_{\mathfrak{T},s}(z)Q_s$, and $G^{f_{\mathfrak{T}}}(z) = [M^{f_{\mathfrak{T}}}(z)]^{-1}Q^{f_{\mathfrak{T}}}(z)[M^{f_{\mathfrak{T}}}(z)]^{-1}$.

By Masry (1996, equation (2.13)) and Li et al. (2003, equation (A.9)), as $n \to +\infty$

$$\hat{\beta}_{\mathfrak{T}}(z) - \beta_{\mathfrak{T}}(z) = \bar{S}_{\mathfrak{T}}^{-1}(z)\bar{\tau}_{\mathfrak{T}}(z) + h_n^{p+1}\bar{S}_{\mathfrak{T}}^{-1}\bar{B}_{\mathfrak{T}}(z)m_{\mathfrak{T}}^{(p+1)}(z) + O_p(h_n^{p+2}), \quad 0 \leq |k| \leq p, \text{ (S.3)}$$

where $\hat{\beta}_{\mathfrak{T}}(z) = (h_n^0\hat{b}_{0,\mathfrak{T}}^\top(z), \ldots, h_n^p\hat{b}_{p,\mathfrak{T}}^\top(z))^\top$ and $\hat{b}_{k,\mathfrak{T}}(z)$ are the estimates of parameter vectors $(b_{\underline{j},\mathfrak{T}}(z))_{|\underline{j}|=k}$ in objective function (6) and $m_{\mathfrak{T}}^{(p+1)}(z)$ is the N_{p+1} elements of derivatives $(D^{\underline{j}}m_{\mathfrak{T}})(z)/\underline{j}!$ for $|\underline{j}| = p+1$ arranged in the lexicographical order. Note that the term $O_p(h_n^{p+2})$ in (S.3) constitutes the terms that are bounded in probability by h_n^{p+2} uniformly in $z \in \mathcal{D}$ by Masry (1996, Theorem 2 and Corollary 3) since $[\ln n/nh_n^{2d+2}]^{1/2} \to \infty$ by Assumption 9.

Recall that local derivative estimator $\hat{\delta}_{\mathfrak{T}}(z) = L\hat{b}_{\mathfrak{T}}(z) = h_n^{-1}L\hat{\beta}_{\mathfrak{T}}(z)$ in equation (8) is defined as the difference of the first d and last d elements of $\hat{b}_{1,\mathfrak{T}}(z)$. Thus, the average derivative estimator and average outer product of gradients are then defined as in equations (10) and (11) by

$$\hat{\delta}_{\mathfrak{T}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\delta}_{\mathfrak{T}}(Z_{i\mathfrak{T}}) = \frac{1}{nh_n}\sum_{i=1}^{n}L\hat{\beta}_{\mathfrak{T}}(Z_{i\mathfrak{T}}).$$
$$\hat{\delta}_{\mathfrak{T}\mathfrak{T}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\delta}_{\mathfrak{T}}(Z_{i\mathfrak{T}})\hat{\delta}_{\mathfrak{T}}^\top(Z_{i\mathfrak{T}}) = \frac{1}{nh_n^2}\sum_{i=1}^{n}L\hat{\beta}_{\mathfrak{T}}(Z_{i\mathfrak{T}})\hat{\beta}_{\mathfrak{T}}^\top(Z_{i\mathfrak{T}})L^\top.$$

Using (S.3), we will study now the sample average of $\hat{\beta}_{\mathfrak{T}}(Z_{i\mathfrak{T}}) - \beta_{\mathfrak{T}}(Z_{i\mathfrak{T}})$:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}[\hat{\beta}_{\mathfrak{T}}(Z_{i\mathfrak{T}}) - \beta_{\mathfrak{T}}(Z_{i\mathfrak{T}})] \\
&= \frac{1}{n}\sum_{i=1}^{n}\bar{S}_{\mathfrak{T}}^{-1}(Z_{i\mathfrak{T}})\bar{\tau}_{\mathfrak{T}}(Z_{i\mathfrak{T}}) + \frac{h_n^{p+1}}{n}\sum_{i=1}^{n}\bar{S}_{\mathfrak{T}}^{-1}(Z_{i\mathfrak{T}})\bar{B}_{\mathfrak{T}}(Z_{i\mathfrak{T}})m_{\mathfrak{T}}^{(p+1)}(Z_{i\mathfrak{T}}) + O_p(h_n^{p+2}) \qquad \text{(S.4)} \\
&= \bar{A}_{\mathfrak{T}}^{11} + h_n^{p+1}\bar{A}_{\mathfrak{T}}^{12} + O_p(h_n^{p+2}),
\end{aligned}$$

where $\bar{A}_{\mathfrak{T}}^{11} = \frac{1}{n}\sum_{i=1}^{n}\bar{S}_{\mathfrak{T}}^{-1}(Z_{i\mathfrak{T}})\bar{\tau}_{\mathfrak{T}}(Z_{i\mathfrak{T}})$ and $\bar{A}_{\mathfrak{T}}^{12} = \frac{1}{n}\sum_{i=1}^{n}\bar{S}_{\mathfrak{T}}^{-1}(Z_{i\mathfrak{T}})\bar{B}_{\mathfrak{T}}(Z_{i\mathfrak{T}})m_{\mathfrak{T}}^{(p+1)}(Z_{i\mathfrak{T}})$. Similarly, we will also analyze the asymptotic behavior of $L\hat{\beta}_{\mathfrak{T}}(Z_{i\mathfrak{T}})\hat{\beta}_{\mathfrak{T}}^{\top}(Z_{i\mathfrak{T}})L^{\top} - L\beta_{\mathfrak{T}}(Z_{i\mathfrak{T}})\beta_{\mathfrak{T}}^{\top}(Z_{i\mathfrak{T}})L^{\top}$. This difference can be decomposed as

$$\begin{aligned}
&L\frac{1}{n}\sum_{i=1}^{n}[\hat{\beta}_{\mathfrak{T}}(Z_{i\mathfrak{T}})\hat{\beta}_{\mathfrak{T}}^{\top}(Z_{i\mathfrak{T}}) - \beta_{\mathfrak{T}}(Z_{i\mathfrak{T}})\beta_{\mathfrak{T}}^{\top}(Z_{i\mathfrak{T}})]L^{\top} \\
=&L\frac{1}{n}\sum_{i=1}^{n}[\beta_{\mathfrak{T}}(Z_{i\mathfrak{T}})\{\hat{\beta}_{\mathfrak{T}}^{\top}(Z_{i\mathfrak{T}}) - \beta_{\mathfrak{T}}^{\top}(Z_{i\mathfrak{T}})\}]L^{\top} \qquad \text{(S.5)} \\
+&L\frac{1}{n}\sum_{i=1}^{n}[\{\hat{\beta}_{\mathfrak{T}}(Z_{i\mathfrak{T}}) - \beta_{\mathfrak{T}}(Z_{i\mathfrak{T}})\}\beta_{\mathfrak{T}}^{\top}(Z_{i\mathfrak{T}})]L^{\top} \qquad \text{(S.6)} \\
+&L\frac{1}{n}\sum_{i=1}^{n}[\{\hat{\beta}_{\mathfrak{T}}(Z_{i\mathfrak{T}}) - \beta_{\mathfrak{T}}(Z_{i\mathfrak{T}})\}\{\hat{\beta}_{\mathfrak{T}}^{\top}(Z_{i\mathfrak{T}}) - \beta_{\mathfrak{T}}^{\top}(Z_{i\mathfrak{T}})\}]L^{\top}. \qquad \text{(S.7)}
\end{aligned}$$

Given the uniform consistency of $\hat{\beta}_{\mathfrak{T}}(Z_{i\mathfrak{T}})$ by Čížek and Lei (2018, Theorem 1), the last term (S.7) is asymptotically negligible with respect to (S.5) and (S.6). Since (S.5) is just the transpose of (S.6), we will only discuss the analysis of the latter term in what follows, the results for the first one can be derived analogously. Due to uniform boundedness of $L\beta_{\mathfrak{T}}(Z_{i\mathfrak{T}})/h_n = L_1 m_{\mathfrak{T}}^{'}(Z_{i\mathfrak{T}})$, see Assumption 9, substituting (S.4) into (S.6) results up to a term $O_p(h_n^{p+3})$ in

$$L\left\{\frac{h_n}{n}\sum_{i=1}^{n}\bar{S}_{\mathfrak{T}}^{-1}(Z_{i\mathfrak{T}})\bar{\tau}_{\mathfrak{T}}(Z_{i\mathfrak{T}})m_{\mathfrak{T}}^{'}(Z_{i\mathfrak{T}})^{\top} + \frac{h_n^{p+2}}{n}\sum_{i=1}^{n}\bar{S}_{\mathfrak{T}}^{-1}(Z_{i\mathfrak{T}})\bar{B}_{\mathfrak{T}}(Z_{i\mathfrak{T}})m_{\mathfrak{T}}^{(p+1)}(Z_{i\mathfrak{T}})m_{\mathfrak{T}}^{'}(Z_{i\mathfrak{T}})^{\top}\right\}L_1^{\top}.$$

Denoting the two sums in the curly brackets $\bar{A}_{\mathfrak{T}\mathfrak{T}}^{21}$ and $h_n^{p+2}\bar{A}_{\mathfrak{T}\mathfrak{T}}^{22}$, we first study the asymptotic properties of $\bar{A}_{\mathfrak{T}}^{12}$ and $\bar{A}_{\mathfrak{T}\mathfrak{T}}^{22}$; later, we discuss $\bar{A}_{\mathfrak{T}}^{11}$ and $\bar{A}_{\mathfrak{T}\mathfrak{T}}^{21}$.

Lemma 1. *Under Assumptions 1–9, $\bar{A}_{\mathfrak{T}}^{12} = A_{\mathfrak{T}}^{1} + O(h_n)$ and $\bar{A}_{\mathfrak{T}\mathfrak{T}}^{22} = A_{\mathfrak{T}\mathfrak{T}}^{2} + O(h_n)$ almost surely as $n \to \infty$, where $A_{\mathfrak{T}}^{1} = M^{-1}B\,\mathrm{E}[m_{\mathfrak{T}}^{(p+1)}(Z_{i\mathfrak{T}})]$ and $A_{\mathfrak{T}\mathfrak{T}}^{2} = M^{-1}B\,\mathrm{E}[m_{\mathfrak{T}}^{(p+1)}(Z_{i\mathfrak{T}})m_{\mathfrak{T}}^{'}(Z_{i\mathfrak{T}})^{\top}]$.*

Proof. The first claim was established in Čížek and Lei (2018, Lemma 1). For the second claim,

note that $\sup_{z\in\mathcal{D}}|\bar{S}_{\mathfrak{T}}(z) - M^{f_{\mathfrak{T}}}(z) - h_n Q^{f_{\mathfrak{T}}}(z)| = o(h_n^{3/2})$, $\sup_{z\in\mathcal{D}}|(\bar{S}_{\mathfrak{T}}(z))^{-1} - (M^{f_{\mathfrak{T}}}(z))^{-1}| = O(h_n)$, and $\sup_{z\in\mathcal{D}}|\bar{B}_{\mathfrak{T}}(z) - Bf_{\mathfrak{T}}(z)| = O(h_n)$ almost surely for $n \to \infty$ as discussed in the proof of Čížek and Lei (2018, Lemma 1). Using $M^{f_{\mathfrak{T}}}(z) = Mf_{\mathfrak{T}}(z)$, it follows that $\bar{A}^{22}_{\mathfrak{T}\mathfrak{T}} = M^{-1}B\frac{1}{n}\sum_{i=1}^n m^{(p+1)}_{\mathfrak{T}}(Z_{i\mathfrak{T}})m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top + O(h_n)$ almost surely for $n \to +\infty$ by Assumption 9.4.

By Assumption 9.4, the function $m^{(p+1)}_{\mathfrak{T}}(z)m'_{\mathfrak{T}}(z)^\top$ is bounded and uniformly continuous in z. Furthermore, 9.4 ensures that expectations $\mathrm{E}\,|m^{(p+1)}_{\mathfrak{T}}(Z_{i\mathfrak{T}})m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top|$ exist and are finite. Hence, the average $\frac{1}{n}\sum_{i=1}^n\{m^{(p+1)}_{\mathfrak{T}}(Z_{i\mathfrak{T}})m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top\}$ converges to $\mathrm{E}[m^{(p+1)}_{\mathfrak{T}}(Z_{i\mathfrak{T}})m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top]$ almost surely by Khintchine's law of large numbers (due to the distribution of $Z_{i\mathfrak{T}}$ being the same over time by Assumption 3). Therefore, we conclude that $\bar{A}^{22}_{\mathfrak{T}\mathfrak{T}} = M^{-1}B\,\mathrm{E}[m^{(p+1)}_{\mathfrak{T}}(Z_{i\mathfrak{T}})$ $m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top] + O(h_n) = A^2_{\mathfrak{T}} + O(h_n)$ almost surely as $n \to +\infty$. ■

We now study the terms $\bar{A}^{11}_{\mathfrak{T}}$ and $\bar{A}^{21}_{\mathfrak{T}\mathfrak{T}}$. Since it holds uniformly in $z \in \mathcal{D}$ that $\bar{S}^{-1}_{\mathfrak{T}}(z) = (M^{f_{\mathfrak{T}}}(z))^{-1} - h_n G^{f_{\mathfrak{T}}}(z) + o(h_n^{3/2})$ almost surely as $n \to \infty$ (see the proof of Lemma 1), we can decompose $\bar{A}^{11}_{\mathfrak{T}}$ and $\bar{A}^{21}_{\mathfrak{T}\mathfrak{T}}$ into

$$\begin{aligned}\bar{A}^{11}_{\mathfrak{T}} &= \frac{1}{n}\sum_{i=1}^n\Big[(M^{f_{\mathfrak{T}}}(Z_{i\mathfrak{T}}))^{-1}\bar{\tau}_{\mathfrak{T}}(Z_{i\mathfrak{T}}) - h_n G^{f_{\mathfrak{T}}}(Z_{i\mathfrak{T}})\bar{\tau}_{\mathfrak{T}}(Z_{i\mathfrak{T}}) + o(h_n^{3/2})\bar{\tau}_{\mathfrak{T}}(Z_{i\mathfrak{T}})\Big] \\ &= \bar{J}^{11}_{\mathfrak{T}} - h_n\bar{J}^{12}_{\mathfrak{T}} + o(h_n^{3/2}) = h_n\bar{J}^{1}_{\mathfrak{T}} + o(h_n^{3/2})\end{aligned} \tag{S.8}$$

and

$$\begin{aligned}\bar{A}^{21}_{\mathfrak{T}\mathfrak{T}} &= \frac{h_n}{n}\sum_{i=1}^n\Big[(M^{f_{\mathfrak{T}}}(Z_{i\mathfrak{T}}))^{-1}\bar{\tau}_{\mathfrak{T}}(Z_{i\mathfrak{T}})m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top - h_n G^{f_{\mathfrak{T}}}(Z_{i\mathfrak{T}})\bar{\tau}_{\mathfrak{T}}(Z_{i\mathfrak{T}})m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top + o(h_n^{3/2})\bar{\tau}_{\mathfrak{T}}(Z_{i\mathfrak{T}})m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top\Big] \\ &= h_n\bar{J}^{21}_{\mathfrak{T}\mathfrak{T}} - h_n^2\bar{J}^{22}_{\mathfrak{T}\mathfrak{T}} + o(h_n^{5/2}) = h_n^2\bar{J}^{2}_{\mathfrak{T}\mathfrak{T}} + o(h_n^{5/2})\end{aligned} \tag{S.9}$$

almost surely, where $\bar{J}^{11}_{\mathfrak{T}} = \frac{1}{n}\sum_{i=1}^n(M^{f_{\mathfrak{T}}}(Z_{i\mathfrak{T}}))^{-1}\bar{\tau}_{\mathfrak{T}}(Z_{i\mathfrak{T}})$, $\bar{J}^{12}_{\mathfrak{T}} = \frac{1}{n}\sum_{i=1}^n G^{f_{\mathfrak{T}}}(Z_{i\mathfrak{T}})\bar{\tau}_{\mathfrak{T}}(Z_{i\mathfrak{T}})$ and $\bar{J}^{21}_{\mathfrak{T}\mathfrak{T}} = \frac{1}{n}\sum_{i=1}^n(M^{f_{\mathfrak{T}}}(Z_{i\mathfrak{T}}))^{-1}\bar{\tau}_{\mathfrak{T}}(Z_{i\mathfrak{T}})m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top$, $\bar{J}^{22}_{\mathfrak{T}\mathfrak{T}} = \frac{1}{n}\sum_{i=1}^n G^{f_{\mathfrak{T}}}(Z_{i\mathfrak{T}})\bar{\tau}_{\mathfrak{T}}(Z_{i\mathfrak{T}})m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top$; $\bar{J}^{1}_{\mathfrak{T}} = \bar{J}^{11}_{\mathfrak{T}}/h_n - \bar{J}^{12}_{\mathfrak{T}}$ and $\bar{J}^{2}_{\mathfrak{T}\mathfrak{T}} = \bar{J}^{21}_{\mathfrak{T}\mathfrak{T}}/h_n - \bar{J}^{22}_{\mathfrak{T}\mathfrak{T}}$. Note that the second equalities in (S.8) and (S.9) follow from Theorem 5 in Masry (1996) as it proves (elementwise) that $\sup_{z\in\mathcal{D}}|\bar{\tau}_{\mathfrak{T}}(z)| = o(1)$ almost surely as $n \to +\infty$ under Assumption 9, and Khintchine's law of large numbers that applies by Assumption 9 again.

Let $\bar{J}^{11}_{\mathfrak{T},r} = e_r^\top\bar{J}^{11}_{\mathfrak{T}}$ denote the rth element of $\bar{J}^{11}_{\mathfrak{T}}$ and $\bar{J}^{21}_{\mathfrak{T}\mathfrak{T},r} = e_r^\top\bar{J}^{21}_{\mathfrak{T}\mathfrak{T}}$ denote the rth row of

$\bar{J}^{21}_{\mathfrak{T}\mathfrak{T}}$. Their asymptotic behavior is derived in the following lemmas.

Lemma 2. *Under Assumptions 1–9,* $\bar{J}^{11}_{\mathfrak{T},r} = O_p\left((nh_n^d)^{-1}\right)$ *and* $\sqrt{n}\bar{J}^{21}_{\mathfrak{T}\mathfrak{T},r}/h_n \to N(0, \Phi^{21}_{\mathfrak{T},r})$ *as* $n \to \infty$ *for* $r = 2, \ldots, 2d+1$*, where*

$$\Phi^{21}_{\mathfrak{T}\mathfrak{T},r} = \mathrm{E}\left[\sigma_{\mathfrak{T}\mathfrak{T}}(Z_{i\mathfrak{T}}) m^{''}_{\mathfrak{T}}(Z_{i\mathfrak{T}}) e_{r-1}^{\top} e_{r-1} m^{''}_{\mathfrak{T}}(Z_{i\mathfrak{T}})^{\top}\right],$$

and $m^{''}_{\mathfrak{T}}(z) = \partial m^{'}_{\mathfrak{T}}(z)/\partial z^{\top}$.

Proof. As the first claim was proved by Čížek and Lei (2018, Lemma 2), we focus on the second claim. Once we substitute $\bar{\tau}_{\mathfrak{T}}(Z_{i\mathfrak{T}})$ in $(M^{f_{\mathfrak{T}}}(Z_{i\mathfrak{T}}))^{-1}\bar{\tau}_{\mathfrak{T}}(Z_{i\mathfrak{T}}) m^{'}_{\mathfrak{T}}(Z_{i\mathfrak{T}})^{\top}$ from (S.1), the rth row $\bar{J}^{21}_{\mathfrak{T}\mathfrak{T},r}$ of matrix $\bar{J}^{21}_{\mathfrak{T}\mathfrak{T}}$ is an average

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{|\underline{k}|=0}^{p}(M^{f_{\mathfrak{T}}}(Z_{i\mathfrak{T}}))^{-1}_{r,\underline{k}}\frac{1}{n}\sum_{j=1}^{n} V_{j\mathfrak{T}}\left(\frac{Z_{j\mathfrak{T}} - Z_{i\mathfrak{T}}}{h_n}\right)^{\underline{k}} K_h(Z_{j\mathfrak{T}} - Z_{i\mathfrak{T}}) m^{'}_{\mathfrak{T}}(Z_{i\mathfrak{T}})^{\top}.$$

Considering this sum, it can be expressed for a given r and $\mathfrak{T}$ as

$$\begin{aligned}&\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{|\underline{k}|=0}^{p}(M^{f_{\mathfrak{T}}}(Z_{i\mathfrak{T}}))^{-1}_{r,\underline{k}} V_{j\mathfrak{T}}\left(\frac{Z_{j\mathfrak{T}} - Z_{i\mathfrak{T}}}{h_n}\right)^{\underline{k}} K_h(Z_{j\mathfrak{T}} - Z_{i\mathfrak{T}}) m^{'}_{\mathfrak{T}}(Z_{i\mathfrak{T}})^{\top}\\ +&\frac{1}{n^2}\sum_{|\underline{k}|=0}^{p}\sum_{i=1}^{n}(M^{f_{\mathfrak{T}}}(Z_{i\mathfrak{T}}))^{-1}_{r,\underline{k}} V_{i\mathfrak{T}}\left(\frac{Z_{i\mathfrak{T}} - Z_{i\mathfrak{T}}}{h_n}\right)^{\underline{k}} K_h(Z_{i\mathfrak{T}} - Z_{i\mathfrak{T}}) m^{'}_{\mathfrak{T}}(Z_{i\mathfrak{T}})^{\top},\end{aligned}$$

where the second term equals 0, whereas the first term forms the following vector of U-statistics:

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} H_r(V_{j\mathfrak{T}}, Z_{j\mathfrak{T}}, Z_{j\mathfrak{T}}; V_{i\mathfrak{T}}, Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}})[1 + O(h_n^2)] \tag{S.10}$$

with ($M^{f_{\mathfrak{T}}}(Z_{i\mathfrak{T}}) = M f_{\mathfrak{T}}(Z_{i\mathfrak{T}})$ and $m^{''}_{\mathfrak{T}}(z) = \partial m^{'}_{\mathfrak{T}}(z)/\partial z^{\top}$)

$$\begin{aligned}&H_r(V_{j\mathfrak{T}}, Z_{j\mathfrak{T}}, Z_{j\mathfrak{T}}; V_{i\mathfrak{T}}, Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}}) = \frac{1}{2}\sum_{|\underline{k}|=0}^{p}(M)^{-1}_{r,\underline{k}}\\ &\times\left[V_{j\mathfrak{T}}\left(\frac{Z_{j\mathfrak{T}} - Z_{i\mathfrak{T}}}{h_n}\right)^{\underline{k}}\frac{K_h(Z_{j\mathfrak{T}} - Z_{i\mathfrak{T}})}{f_{\mathfrak{T}}(Z_{i\mathfrak{T}})}\left\{m^{'}_{\mathfrak{T}}(Z_{j\mathfrak{T}})^{\top} - (Z_{j\mathfrak{T}} - Z_{i\mathfrak{T}})^{\top} m^{''}_{\mathfrak{T}}(Z_{j\mathfrak{T}})^{\top}\right\}\right.\\ &+\quad\left. V_{i\mathfrak{T}}\left(\frac{Z_{i\mathfrak{T}} - Z_{j\mathfrak{T}}}{h_n}\right)^{\underline{k}}\frac{K_h(Z_{i\mathfrak{T}} - Z_{j\mathfrak{T}})}{f_{\mathfrak{T}}(Z_{j\mathfrak{T}})}\left\{m^{'}_{\mathfrak{T}}(Z_{i\mathfrak{T}})^{\top} - (Z_{i\mathfrak{T}} - Z_{j\mathfrak{T}})^{\top} m^{''}_{\mathfrak{T}}(Z_{i\mathfrak{T}})^{\top}\right\}\right]\\ &= H^1_r(V_{j\mathfrak{T}}, Z_{j\mathfrak{T}}, Z_{j\mathfrak{T}}; V_{i\mathfrak{T}}, Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}}) - H^2_r(V_{j\mathfrak{T}}, Z_{j\mathfrak{T}}, Z_{j\mathfrak{T}}; V_{i\mathfrak{T}}, Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}});\end{aligned}$$

the term $O(h_n^2)$ represents the remainders of the Taylor expansions of $m_{\mathfrak{T}}^{'}(Z_{i\mathfrak{T}})$ at $Z_{j\mathfrak{T}}$ and at $Z_{i\mathfrak{T}}$ and is uniform due to Assumption 9 for $p \geq 1$, H_r^1 and H_r^2 represent terms with the first and second derivatives of $m_{\mathfrak{T}}$, respectively. Denoting the conditional expectation $\mathcal{H}_r(V_{i\mathfrak{T}}, Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}}) = \mathrm{E}[H_r(V_{j\mathfrak{T}}, Z_{j\mathfrak{T}}, Z_{j\mathfrak{T}}; V_{i\mathfrak{T}}, Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}})|V_{i\mathfrak{T}}, Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}}]$, Hoeffding's decomposition

$$R_n = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} H_r(V_{j\mathfrak{T}}, Z_{j\mathfrak{T}}, Z_{j\mathfrak{T}}; V_{i\mathfrak{T}}, Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}}) - \frac{1}{n}\sum_{i=1}^{n}\mathcal{H}_r(V_{i\mathfrak{T}}, Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}})$$

is asymptotically negligible: as in Kong and Xia (2014, proof of Theorem 1), Proposition 1 of Arcones (1995) implies there is some $c > 0$ such that it holds $P(\sqrt{n}R_n \geq h_n\epsilon) \leq 2\exp(-c\epsilon h_n n^{1/2})$ for any $\epsilon > 0$ due to Assumption 9, and by the Borel-Cantelli lemma, $|R_n| = o(n^{-1/2}h_n)$ and $|\sqrt{n}R_n/h_n| = o(1)$ almost surely as $n \to \infty$.

Let us now focus on $n^{-1}\sum_{i=1}^{n}\mathcal{H}_r(V_{i\mathfrak{T}}, Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}})$. Defining the conditional expectation $\mathcal{H}_r^1(V_{i\mathfrak{T}}, Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}}) = \mathrm{E}[H_r^1(V_{j\mathfrak{T}}, Z_{j\mathfrak{T}}, Z_{j\mathfrak{T}}; V_{i\mathfrak{T}}, Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}})|V_{i\mathfrak{T}}, Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}}]$, it follows that

$$\begin{aligned}
\mathcal{H}_r^1(V_{i\mathfrak{T}}, Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}}) &= \frac{1}{2}V_{i\mathfrak{T}}\sum_{|\underline{k}|=0}^{p}(M)_{r,\underline{k}}^{-1}\,\mathrm{E}\left[\left(\frac{Z_{i\mathfrak{T}} - Z_{j\mathfrak{T}}}{h_n}\right)^{\underline{k}}\frac{K_h(Z_{i\mathfrak{T}} - Z_{j\mathfrak{T}})}{f_{\mathfrak{T}}(Z_{j\mathfrak{T}})}|Z_{i\mathfrak{T}}\right]m_{\mathfrak{T}}^{'}(Z_{i\mathfrak{T}})^{\top} \\
&= \frac{1}{2}V_{i\mathfrak{T}}\cdot\sum_{|\underline{k}|=0}^{p}(M)_{r,\underline{k}}^{-1}\int v^{\underline{k}}K(v)dv\cdot m_{\mathfrak{T}}^{'}(Z_{i\mathfrak{T}})^{\top} \\
&= \frac{1}{2}V_{i\mathfrak{T}}\cdot\sum_{|\underline{k}|=0}^{p}(M)_{r,\underline{k}}^{-1}(M)_{\underline{k},1}\cdot m_{\mathfrak{T}}^{'}(Z_{i\mathfrak{T}})^{\top} \\
&= \frac{1}{2}V_{i\mathfrak{T}}\cdot I(r=1)\cdot m_{\mathfrak{T}}^{'}(Z_{i\mathfrak{T}})^{\top} = 0.
\end{aligned}$$

By the cross-sectional independence, let

$$\begin{aligned}
&\mathcal{H}_r^2(V_{i\mathfrak{T}}, Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}}) = \mathrm{E}[H_r^2(V_{j\mathfrak{T}}, Z_{j\mathfrak{T}}, Z_{j\mathfrak{T}}; V_{i\mathfrak{T}}, Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}})|V_{i\mathfrak{T}}, Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}}] \\
&= \frac{h_n}{2}V_{i\mathfrak{T}}\sum_{|\underline{k}|=0}^{p}(M)_{r,\underline{k}}^{-1}\,\mathrm{E}\left[\left(\frac{Z_{i\mathfrak{T}} - Z_{j\mathfrak{T}}}{h_n}\right)^{\underline{k}}\frac{K_h(Z_{i\mathfrak{T}} - Z_{j\mathfrak{T}})}{f_{\mathfrak{T}}(Z_{j\mathfrak{T}})}\left(\frac{Z_{i\mathfrak{T}} - Z_{j\mathfrak{T}}}{h_n}\right)^{\top}|Z_{i\mathfrak{T}}\right]m_{\mathfrak{T}}^{''}(Z_{i\mathfrak{T}})^{\top}.
\end{aligned}$$

Since Assumption 9.2 ensures the existence of kernel moments up to order $4p$, we obtain

$$\begin{aligned}
&\mathrm{E}\left[\left(\frac{Z_{i\mathfrak{T}}-Z_{j\mathfrak{T}}}{h_n}\right)^{\underline{k}}\frac{K_h(Z_{i\mathfrak{T}}-Z_{j\mathfrak{T}})}{f_{\mathfrak{T}}(Z_{j\mathfrak{T}})}\left(\frac{Z_{i\mathfrak{T}}-Z_{j\mathfrak{T}}}{h_n}\right)^{\top}|Z_{i\mathfrak{T}}\right]\\
&=\int\left(\frac{Z_{i\mathfrak{T}}-z_{jt}}{h_n}\right)^{\underline{k}}\left(\frac{Z_{i\mathfrak{T}}-z_{jt}}{h_n}\right)^{\top}K_h(Z_{i\mathfrak{T}}-z_{jt})dz_{jt}\\
&=\int u^{\underline{k}}u^{\top}K(u)du=(M_{\underline{k},2},\ldots,M_{\underline{k},2d+1}).
\end{aligned}$$

It follows for $i\in\mathcal{I}_{\mathfrak{T}}$ that

$$\begin{aligned}
\mathcal{H}_r(V_{i\mathfrak{T}},Z_{i\mathfrak{T}},Z_{i\mathfrak{T}}) &= -\frac{1}{2}h_n[1+O(h_n)]V_{i\mathfrak{T}}\sum_{|\underline{k}|=0}^{p}(M)^{-1}_{r,\underline{k}}(M_{\underline{k},2},\ldots,M_{\underline{k},2d+1})m''_{\mathfrak{T}}(Z_{i\mathfrak{T}})^{\top}\\
&= -\frac{1}{2}h_n[1+O(h_n)]V_{i\mathfrak{T}}(I(r=2),\ldots,I(r=2d+1))m''_{\mathfrak{T}}(Z_{i\mathfrak{T}})^{\top}\\
&= -\frac{1}{2}h_n[1+O(h_n)]V_{i\mathfrak{T}}e_{r-1}^{\top}m''_{\mathfrak{T}}(Z_{i\mathfrak{T}})^{\top}.
\end{aligned}$$

Combining all the derived results, it follows for $n\to+\infty$ that

$$\frac{\sqrt{n}}{h_n}\bar{J}^{21}_{\mathfrak{T},r} = \frac{2}{\sqrt{n}h_n}\sum_{i=1}^{n}\mathcal{H}_r(V_{i\mathfrak{T}},Z_{i\mathfrak{T}},Z_{i\mathfrak{T}})+o(1)+O_p(n^{-1}h_n^{-d-1}) \tag{S.11}$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[-V_{i\mathfrak{T}}e_{r-1}^{\top}m''_{\mathfrak{T}}(Z_{i\mathfrak{T}})^{\top}+V_{i\mathfrak{T}}O(h_n)\right] \tag{S.12}$$

$$+o(1)+O_p(n^{-1}h_n^{-d-1}) \tag{S.13}$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[-V_{i\mathfrak{T}}e_{r-1}^{\top}m''_{\mathfrak{T}}(Z_{i\mathfrak{T}})^{\top}\right]+O(h_n)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}V_{i\mathfrak{T}} \tag{S.14}$$

$$+o(1)+O_p((nh_n^{d+1})^{-1}). \tag{S.15}$$

Note that $O_p((nh_n^{d+1})^{-1})=o_p(1)$ by Assumption 9.1. Since Assumption 9.5 ensures that terms $V_{i\mathfrak{T}}$ in (S.14) are independent and identically distributed with zero means and finite second moments, the central limit theorem can be applied to the first and the second term in (S.14). Hence, it implies that the second term is $o_p(1)$ because of the term $O(h_n)$ in front of it. Thus, the first term in (S.14) governs the asymptotic distribution of $\sqrt{n}\bar{J}^{21}_{\mathfrak{T},r}/h_n$:

$$\left(\frac{\sqrt{n}}{h_n}\bar{J}^{21}_{\mathfrak{T},r}\right)^{\top} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[-V_{i\mathfrak{T}}m''_{\mathfrak{T}}(Z_{i\mathfrak{T}})e_{r-1}^{\top}\right]+o_p(1). \tag{S.16}$$

Therefore, it follows that $(\sqrt{n}\bar{J}^{21}_{\mathfrak{T},r}/h_n)^\top \to N(0,\Phi^{21}_{\mathfrak{T}\mathfrak{T},r})$ in distribution as $n\to\infty$ for $r=2,\ldots,2d+1$, where

$$\begin{aligned}\Phi^{21}_{\mathfrak{T}\mathfrak{T},r} &= \mathrm{Var}\left[V_{i\mathfrak{T}}m''_{\mathfrak{T}}(Z_{i\mathfrak{T}})e_{r-1}^\top\right]\\ &= \mathrm{E}\left[\sigma_{\mathfrak{T}\mathfrak{T}}(Z_{i\mathfrak{T}})m''_{\mathfrak{T}}(Z_{i\mathfrak{T}})e_{r-1}^\top e_{r-1}m''_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top\right].\end{aligned}$$

■

Lemma 3. *Under Assumptions 1–9, $\sqrt{n}\bar{J}^{12}_{\mathfrak{T},r}\to N(0,\Phi^{12}_{\mathfrak{T},r})$ and $\sqrt{n}\bar{J}^{22}_{\mathfrak{T},r}\to N(0,\Phi^{22}_{\mathfrak{T},r})$ in distribution as $n\to\infty$ for $r=2,\ldots,2d+1$, where*

$$\Phi^{12}_{\mathfrak{T},r} = \mathrm{E}\left[\sigma_{\mathfrak{T}\mathfrak{T}}(Z_{i\mathfrak{T}})G_{\mathfrak{T};r,1}(Z_{i\mathfrak{T}})G_{\mathfrak{T};r,1}(Z_{i\mathfrak{T}})\right]$$

and

$$\Phi^{22}_{\mathfrak{T},r} = \mathrm{E}\left[\sigma_{\mathfrak{T}\mathfrak{T}}(Z_{i\mathfrak{T}})G_{\mathfrak{T};r,1}(Z_{i\mathfrak{T}})G_{\mathfrak{T};r,1}(Z_{i\mathfrak{T}})m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top\right]$$

and the matrix $G_{\mathfrak{T}}(z)=G^{f_{\mathfrak{T}}}(z)M^{f_{\mathfrak{T}}}(z)=[M^{f_{\mathfrak{T}}}(z)]^{-1}Q^{f_{\mathfrak{T}}}(z)$.

Proof. The first claim was proved by Čížek and Lei (2018, Lemma 3). We focus on the second claim. Recall that $V_{i\mathfrak{T}}=Y_{it}-m_{\mathfrak{T}}(Z_{i\mathfrak{T}})$ andlet $G^{f_{\mathfrak{T}}}_{i,j}(z)$ denote the (i,j)-th element of $G^{f_{\mathfrak{T}}}(z)$. Substituting from (S.1), the rth row $\bar{J}^{22}_{\mathfrak{T}\mathfrak{T},r}$ of matrix $\bar{J}^{21}_{\mathfrak{T}\mathfrak{T}}$ is again an average

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n e_r^\top G^{f_{\mathfrak{T}}}(Z_{i\mathfrak{T}})\bar{\tau}_{\mathfrak{T}}(Z_{i\mathfrak{T}})m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top\\ =&\frac{1}{n}\sum_{i=1}^n\sum_{|\underline{k}|=0}^p G^{f_{\mathfrak{T}}}_{r,\underline{k}}(Z_{i\mathfrak{T}})\bar{\tau}_{\mathfrak{T},\underline{k}}(Z_{i\mathfrak{T}})m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top\\ =&\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n\sum_{|\underline{k}|=0}^p G^{f_{\mathfrak{T}}}_{r,\underline{k}}(Z_{i\mathfrak{T}})V_{j\mathfrak{T}}\left(\frac{Z_{j\mathfrak{T}}-Z_{i\mathfrak{T}}}{h_n}\right)^{\underline{k}}K_h(Z_{j\mathfrak{T}}-Z_{i\mathfrak{T}})m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top.\end{aligned}$$

Considering particular values of r, s, t and the sums inside of the curly brackets, we obtain

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{|\underline{k}|=0}^{p} V_{j\mathfrak{T}} G_{r,\underline{k}}^{f_{\mathfrak{T}}}(Z_{i\mathfrak{T}}) \left(\frac{Z_{j\mathfrak{T}}-Z_{i\mathfrak{T}}}{h_n}\right)^{\underline{k}} K_h(Z_{j\mathfrak{T}}-Z_{i\mathfrak{T}}) m_{\mathfrak{T}}^{'}(Z_{i\mathfrak{T}})^{\top} \\
=&\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{|\underline{k}|=0}^{p} V_{j\mathfrak{T}} G_{r,\underline{k}}^{f_{\mathfrak{T}}}(Z_{i\mathfrak{T}}) \left(\frac{Z_{j\mathfrak{T}}-Z_{i\mathfrak{T}}}{h_n}\right)^{\underline{k}} K_h(Z_{j\mathfrak{T}}-Z_{i\mathfrak{T}}) m_{\mathfrak{T}}^{'}(Z_{i\mathfrak{T}})^{\top} \\
&+\frac{1}{n^2}\sum_{|\underline{k}|=0}^{p}\sum_{i=1}^{n} V_{i\mathfrak{T}} G_{r,\underline{k}}^{f_{\mathfrak{T}}}(Z_{i\mathfrak{T}}) \left(\frac{Z_{i\mathfrak{T}}-Z_{i\mathfrak{T}}}{h_n}\right)^{\underline{k}} K_h(Z_{i\mathfrak{T}}-Z_{i\mathfrak{T}}) m_{\mathfrak{T}}^{'}(Z_{i\mathfrak{T}})^{\top},
\end{aligned} \tag{S.17}$$

where the second term equals 0. Denoting the symmetrized elements

$$\begin{aligned}
&H_r(V_{j\mathfrak{T}}, Z_{j\mathfrak{T}}, Z_{j\mathfrak{T}}; V_{i\mathfrak{T}}, Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}}) = \frac{1}{2}\sum_{|\underline{k}|=0}^{p} \\
&\times\left[V_{j\mathfrak{T}} G_{r,\underline{k}}^{f_{\mathfrak{T}}}(Z_{i\mathfrak{T}}) \left(\frac{Z_{j\mathfrak{T}}-Z_{i\mathfrak{T}}}{h_n}\right)^{\underline{k}} K_h(Z_{j\mathfrak{T}}-Z_{i\mathfrak{T}}) m_{\mathfrak{T}}^{'}(Z_{i\mathfrak{T}})^{\top}\right. \\
&\left.+V_{i\mathfrak{T}} G_{r,\underline{k}}^{f_{\mathfrak{T}}}(Z_{j\mathfrak{T}}) \left(\frac{Z_{i\mathfrak{T}}-Z_{j\mathfrak{T}}}{h_n}\right)^{\underline{k}} K_h(Z_{i\mathfrak{T}}-Z_{j\mathfrak{T}}) m_{\mathfrak{T}}^{'}(Z_{j\mathfrak{T}})^{\top}\right],
\end{aligned}$$

we can rewrite (S.17) as $n^{-2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{|\underline{k}|=0}^{p} H_r(V_{j\mathfrak{T}}, Z_{j\mathfrak{T}}, Z_{j\mathfrak{T}}; V_{i\mathfrak{T}}, Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}})$. By the cross-sectional independence, let

$$\begin{aligned}
&\mathcal{H}_r(V_{i\mathfrak{T}}, Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}}) = \mathrm{E}[H_r(V_{j\mathfrak{T}}, Z_{j\mathfrak{T}}, Z_{j\mathfrak{T}}; V_{i\mathfrak{T}}, Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}})|V_{i\mathfrak{T}}, Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}}] \\
&=\frac{1}{2}\sum_{|\underline{k}|=0}^{p} V_{i\mathfrak{T}}\,\mathrm{E}\left[G_{r,\underline{k}}^{f_{\mathfrak{T}}}(Z_{j\mathfrak{T}}) \left(\frac{Z_{i\mathfrak{T}}-Z_{j\mathfrak{T}}}{h_n}\right)^{\underline{k}} K_h(Z_{i\mathfrak{T}}-Z_{j\mathfrak{T}}) m_{\mathfrak{T}}^{'}(Z_{j\mathfrak{T}})^{\top}|Z_{i\mathfrak{T}}\right].
\end{aligned} \tag{S.18}$$

To study the above expectation, recall that Assumption 9.2 ensures the existence of the kernel moments up to order $4p$. Hence, by Taylor expansion around z_{jt} and using the standard change of variables argument

$$\begin{aligned}
&\mathrm{E}\left[G_{r,\underline{k}}^{f_{\mathfrak{T}}}(Z_{j\mathfrak{T}}) \left(\frac{Z_{i\mathfrak{T}}-Z_{j\mathfrak{T}}}{h_n}\right)^{\underline{k}} K_h(Z_{i\mathfrak{T}}-Z_{j\mathfrak{T}}) m_{\mathfrak{T}}^{'}(Z_{j\mathfrak{T}})^{\top}|Z_{i\mathfrak{T}}\right] \\
&=\int f_{\mathfrak{T}}(z_{jt}) G_{r,\underline{k}}^{f_{\mathfrak{T}}}(z_{jt}) \left(\frac{Z_{i\mathfrak{T}}-z_{jt}}{h_n}\right)^{\underline{k}} K_h(Z_{i\mathfrak{T}}-z_{jt}) m_{\mathfrak{T}}^{'}(z_{jt})^{\top} dz_{jt} \\
&=\int f_{\mathfrak{T}}(Z_{i\mathfrak{T}}-h_n u) G_{r,\underline{k}}^{f_{\mathfrak{T}}}(Z_{i\mathfrak{T}}-h_n u) u^{\underline{k}} K(u) m_{\mathfrak{T}}^{'}(Z_{i\mathfrak{T}}-h_n u)^{\top} du \\
&=f_{\mathfrak{T}}(Z_{i\mathfrak{T}}) G_{r,\underline{k}}^{f_{\mathfrak{T}}}(Z_{i\mathfrak{T}}) \mu_{\underline{k}} m_{\mathfrak{T}}^{'}(Z_{i\mathfrak{T}})^{\top} + O(h_n),
\end{aligned} \tag{S.19}$$

where the last equality follows from Assumption 9.3 that $f_{\mathfrak{T}}$ and $m'_{\mathfrak{T}}$, their derivatives, and $G^{f_{\mathfrak{T}}}_{r,\underline{j}}(z)$ are uniformly bounded in $z \in \mathcal{D}$. Hence, the term $O(h_n)$ is uniform on $\mathcal{D}$. Substituting (S.19) into (S.18) thus yields

$$\begin{aligned}\mathcal{H}_r(V_{i\mathfrak{T}}, Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}}) &= \frac{1}{2} V_{i\mathfrak{T}} \sum_{|\underline{k}|=0}^{p} [f_{\mathfrak{T}}(Z_{i\mathfrak{T}}) G^{f_{\mathfrak{T}}}_{r,\underline{k}}(Z_{i\mathfrak{T}}) \mu_{\underline{k}} m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top + O(h_n)] \\ &= \frac{1}{2} V_{i\mathfrak{T}} (G^{f_{\mathfrak{T}}}(Z_{i\mathfrak{T}}) M^{f_{\mathfrak{T}}}(Z_{i\mathfrak{T}}))_{r,1} m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top + \frac{1}{2} V_{i\mathfrak{T}} O(h_n) \\ &= \frac{1}{2} V_{i\mathfrak{T}} G_{\mathfrak{T};r,1}(Z_{i\mathfrak{T}}) m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top + \frac{1}{2} V_{i\mathfrak{T}} O(h_n)\end{aligned}$$

for $i \in \mathcal{I}_{\mathfrak{T}}$. Therefore, again by the Hoeffding's decomposition for U-statistics, it follows for $n \to +\infty$ that

$$\sqrt{n}\bar{J}^{22}_{\mathfrak{T},r} = \frac{2}{\sqrt{n}} \sum_{i=1}^{n} \mathcal{H}_r(V_{i\mathfrak{T}}, Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}}) + o_p(1) + O_p(n^{-1}h_n^{-d}) \tag{S.20}$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[V_{i\mathfrak{T}} G_{\mathfrak{T};r,1}(Z_{i\mathfrak{T}}) m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top + V_{i\mathfrak{T}} O(h_n) \right] \tag{S.21}$$

$$+ o_p(1) + O_p(n^{-1}h_n^{-d}) \tag{S.22}$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[V_{i\mathfrak{T}} G_{\mathfrak{T};r,1}(Z_{i\mathfrak{T}}) m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top \right] + \frac{O(h_n)}{\sqrt{n}} \sum_{i=1}^{n} V_{i\mathfrak{T}} \tag{S.23}$$

$$+ o_p(1) + O_p((nh_n^d)^{-1}). \tag{S.24}$$

The term $\sqrt{n}\bar{J}^{22}_{\mathfrak{T},r}$ is analyzed similarly to $\sqrt{n}/h_n \bar{J}^{21}_{\mathfrak{T},r}$ given in (S.11). Again by Assumption 9.1, $O_p((nh_n^d)^{-1}) = o_p(1)$, the central limit theorem is applicable to the terms in (S.23) by Assumption 9.5 and the second term in (S.23) is negligible because of the term $O(h_n)$. Therefore,

$$\left(\sqrt{n}\bar{J}^{22}_{\mathfrak{T},r}\right)^\top = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[V_{i\mathfrak{T}} G_{\mathfrak{T};r,1}(Z_{i\mathfrak{T}}) m'_{\mathfrak{T}}(Z_{i\mathfrak{T}}) \right] + o_p(1), \tag{S.25}$$

and consequently, $(\sqrt{n}\bar{J}^{22}_{\mathfrak{T},r})^\top \to N(0, \Phi_{\mathfrak{T}\mathfrak{T},r})$ in distribution as $n \to \infty$ for $r = 2, \ldots, 2d+1$, where

$$\begin{aligned}\Phi^2_{\mathfrak{T},r} &= \mathrm{Var}\left[V_{i\mathfrak{T}} G_{\mathfrak{T};r,1}(Z_{i\mathfrak{T}}) m'_{\mathfrak{T}}(Z_{i\mathfrak{T}}) \right] \\ &= \mathrm{E}\left[\sigma_{\mathfrak{T}\mathfrak{T}}(Z_{i\mathfrak{T}}) G_{\mathfrak{T};r,1}(Z_{i\mathfrak{T}}) G_{\mathfrak{T};r,1}(Z_{i\mathfrak{T}}) m'_{\mathfrak{T}}(Z_{i\mathfrak{T}}) m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top \right].\end{aligned}$$

■

Lemma 4. *Let* $G_{\mathfrak{T};\cdot 1}(z) = G_{\mathfrak{T}}(z)e_1$ *be the first column of* $G_{\mathfrak{T}}(z) = [M^{f_{\mathfrak{T}}}(z)]^{-1}Q^{f_{\mathfrak{T}}}(z)$ *and* $GM_{\mathfrak{T}}(z) = vec(-LG_{\mathfrak{T};\cdot 1}(z), -2L_1 m''_{\mathfrak{T}}(Z_{i\mathfrak{T}})L_1^\top - LG_{\mathfrak{T};\cdot 1}(Z_{i\mathfrak{T}})m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top L_1^\top - L_1 m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})G_{\mathfrak{T};\cdot 1}(Z_{i\mathfrak{T}})^\top L^\top)$. *Under Assumptions 1–9,*

$$\sqrt{n}vec(L\bar{J}^1_{\mathfrak{T}}, L[\bar{J}^2_{\mathfrak{T}\mathfrak{T}} + \bar{J}^{2\top}_{\mathfrak{T}\mathfrak{T}}]L^\top) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[V_{i\mathfrak{T}}GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})] \to N(0, \Phi_{\mathfrak{T}}) \tag{S.26}$$

in distribution as $n \to +\infty$, *where*

$$\Phi_{\mathfrak{T}} = \mathrm{E}\left[\sigma_{\mathfrak{T}\mathfrak{T}}(Z_{i\mathfrak{T}})GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top\right].$$

Proof. By equation Čížek and Lei (2018, (B.10)), Assumption 9.1, and equations (S.16) and (S.25) in the proof of Lemmas 1 and 3, respectively, we know that

$$\begin{aligned}\sqrt{n}L\bar{J}^1_{\mathfrak{T}} &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}L\left[V_{i\mathfrak{T}}G_{\mathfrak{T};\cdot 1}(Z_{i\mathfrak{T}})\right] + o_p(1) \\ &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[V_{i\mathfrak{T}}LG_{\mathfrak{T};\cdot 1}(Z_{i\mathfrak{T}})] + o_p(1),\end{aligned} \tag{S.27}$$

and again similarly up to a term negligible in probability,

$$\sqrt{n}L\bar{J}^2_{\mathfrak{T}\mathfrak{T}}L^\top = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[V_{i\mathfrak{T}}\left\{-2L_1 m''_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top L_1^\top - LG_{\mathfrak{T};\cdot,1}(Z_{i\mathfrak{T}})m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top L_1^\top\right\}\right]. \tag{S.28}$$

Consequently,

$$\sqrt{n}\mathrm{vec}(L\bar{J}^1_{\mathfrak{T}}, L[\bar{J}^2_{\mathfrak{T}\mathfrak{T}} + \bar{J}^{2\top}_{\mathfrak{T}\mathfrak{T}}]L^\top) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[V_{i\mathfrak{T}}GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})]. \tag{S.29}$$

Again by Assumption 9.5, we know that terms $\frac{1}{T-\mathfrak{T}}\sum_{t=\mathfrak{T}+1}^{T}V_{i\mathfrak{T}}GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})$ are independent and identically distributed with zero means. Furthermore, the same assumption ensures the

existence of their variance matrix with its $(r_1, r_2)th$ element given by

$$\begin{aligned}
&\mathrm{E}\left[\left(V_{i\mathfrak{T}}\{GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})\}_{r_1}\right)\left(V_{i\mathfrak{T}}[GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})\}_{r_2}\right)\right] \\
=&\mathrm{E}\left[\sigma_{\mathfrak{T}\mathfrak{T}}(Z_{i\mathfrak{T}}, Z_{i\mathfrak{T}})\{GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})\}_{r_1}\{GM_{\mathfrak{T}}(Z_{i\mathfrak{T}})\}_{r_2}\right] \\
=&(\Phi_{\mathfrak{T}})_{r_1,r_2}.
\end{aligned}$$

for $(r_1, r_2 = 1, \ldots, d^2 + d)$. Therefore, the multivariate central limit theorem can be used to conclude that $\sqrt{n}\mathrm{vec}(L\bar{J}^1_{\mathfrak{T}}, L[\bar{J}^2_{\mathfrak{T}\mathfrak{T}} + \bar{J}^{2\top}_{\mathfrak{T}\mathfrak{T}}]L^\top) \to N(0, \Phi_{\mathfrak{T}})$ in distribution as $n \to +\infty$. ■

Finally, applying Lemmas 1–4 to the original decompositions (S.4) and (S.5)–(S.7), we obtain for $n \to \infty$

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}[\hat{\beta}_{\mathfrak{T}}(Z_{i\mathfrak{T}}) - \beta_{\mathfrak{T}}(Z_{i\mathfrak{T}})] \\
&= h_n\bar{J}^1_{\mathfrak{T}} + \frac{o(h_n^{3/2})}{n}\sum_{i=1}^{n}\bar{\tau}_{\mathfrak{T}}(Z_{i\mathfrak{T}}) + h_n^{p+1}A^1_{\mathfrak{T}} + O_p(h_n^{p+2}) \\
&= h_n\bar{J}^1_{\mathfrak{T}} + o_p(h_n^{3/2}n^{-1/2}) + h_n^{p+1}A^1_{\mathfrak{T}} + O_p(h_n^{p+2})
\end{aligned} \tag{S.30}$$

and

$$\begin{aligned}
&\frac{1}{n}L\sum_{i=1}^{n}[\hat{\beta}_{\mathfrak{T}}(Z_{i\mathfrak{T}})\hat{\beta}^\top_{\mathfrak{T}}(Z_{i\mathfrak{T}}) - \beta_{\mathfrak{T}}(Z_{i\mathfrak{T}})\beta^\top_{\mathfrak{T}}(Z_{i\mathfrak{T}})]L^\top \\
&= \left\{h_n^2\left\{\bar{J}^2_{\mathfrak{T}\mathfrak{T}} + \bar{J}^{2\top}_{\mathfrak{T}\mathfrak{T}}\right\}\{1 + o_p(1)\} + \frac{o(h_n^{5/2})}{n}\sum_{i=1}^{n}\left[\bar{\tau}_{\mathfrak{T}}(Z_{i\mathfrak{T}})m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top + m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})\bar{\tau}_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top\right]\right. \\
&\left. + h_n^{p+2}\left\{A^2_{\mathfrak{T}\mathfrak{T}} + A^{2\top}_{\mathfrak{T}\mathfrak{T}}\right\} + O_p(h_n^{p+3})\right\}\{1 + o_p(1)\} \\
&= h_n^2\left\{\bar{J}^2_{\mathfrak{T}\mathfrak{T}} + \bar{J}^{2\top}_{\mathfrak{T}\mathfrak{T}}\right\}\{1 + o_p(1)\} + o_p(h_n^{5/2}n^{-1/2}) + h_n^{p+2}\left\{A^2_{\mathfrak{T}\mathfrak{T}} + A^{2\top}_{\mathfrak{T}\mathfrak{T}}\right\} + O_p(h_n^{p+3}),
\end{aligned} \tag{S.31}$$

where the last equalities follow by applying the arguments and proof of Lemma 4 on the terms $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\bar{\tau}_{\mathfrak{T}}(Z_{i\mathfrak{T}})$ and $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\bar{\tau}_{\mathfrak{T}}(Z_{i\mathfrak{T}})m'_{\mathfrak{T}}(Z_{i\mathfrak{T}})^\top$ and proving thus their asymptotic normality.

F Jackknife

To propose the generalized jackknife bias-correction procedure, let us first recall some asymptotic results for the average derivative estimator. The results for the outer product of gradients are analogous.

First, recall that we define the local derivative estimator $\hat{\delta}_{\mathfrak{T}}(z) = L\hat{b}_{\mathfrak{T}}(z) = h_n^{-1}L\hat{\beta}_{\mathfrak{T}}(z)$ in equation (8) as the difference of the first d and last d elements of $\hat{b}_{1,\mathfrak{T}}(z)$. Subsequently, we define the average derivative estimator and average outer product of gradients as in equations (10) and (11) by

$$\hat{\delta}_{\mathfrak{T}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\delta}_{\mathfrak{T}}(Z_{i\mathfrak{T}}) = \frac{1}{nh_n}\sum_{i=1}^{n}L\hat{\beta}_{\mathfrak{T}}(Z_{i\mathfrak{T}}).$$
$$\hat{\delta}_{\mathfrak{T}\mathfrak{S}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\delta}_{\mathfrak{T}}(Z_{i\mathfrak{T}})\hat{\delta}_{\mathfrak{T}}^{\top}(Z_{i\mathfrak{T}}) = \frac{1}{nh_n^2}\sum_{i=1}^{n}L\hat{\beta}_{\mathfrak{T}}(Z_{i\mathfrak{T}})\hat{\beta}_{\mathfrak{T}}^{\top}(Z_{i\mathfrak{T}})L^{\top}.$$

In these expressions, it holds almost surely by (A.10) for $n \to +\infty$ that

$$\hat{\beta}_{\mathfrak{T}}(z) - \beta_{\mathfrak{T}}(z) = \bar{S}_{\mathfrak{T}}^{-1}(z)\bar{\tau}_{\mathfrak{T}}(z) + h_n^{p+1}\bar{S}_{\mathfrak{T}}^{-1}\bar{B}_{\mathfrak{T}}(z)m_{\mathfrak{T}}^{(p+1)}(z) + O_p(h_n^{p+2}), \tag{S.32}$$

where $\hat{\beta}_{\mathfrak{T}}(z) = (h_n^0\hat{b}_{0,\mathfrak{T}}^{\top}(z), \ldots, h_n^p\hat{b}_{p,\mathfrak{T}}^{\top}(z))^{\top}$ are the coefficients and p is the order of the local polynomial regression. Assuming that $m_{\mathfrak{T}}(\cdot)$ has $p+p'$ derivatives at point z_0, $m_{\mathfrak{T}}(z)$ can be locally approximated by a polynomial of order $p+p'$ and

$$\hat{\beta}_{\mathfrak{T}}(z) - \beta_{\mathfrak{T}}(z) = \bar{S}_{\mathfrak{T}}^{-1}(z)\bar{\tau}_{\mathfrak{T}}(z) \tag{S.33}$$
$$+ h_n^{p+1}\bar{S}_{\mathfrak{T}}^{-1}\bar{B}_{\mathfrak{T}}^{1}(z)m_{\mathfrak{T}}^{(p+1)}(z) + \ldots + h_n^{p+p'}\bar{S}_{\mathfrak{T}}^{-1}\bar{B}_{\mathfrak{T}}^{p'}(z)m_{\mathfrak{T}}^{(p+p')}(z) + O_p(h_n^{p+p'+1}), \tag{S.34}$$

where for $o = 1, \ldots, p'$

$$\bar{B}_{\mathfrak{T}}^{o}(z) = \begin{pmatrix} \bar{S}_{\mathfrak{T},0,p+o}(z) \\ \bar{S}_{\mathfrak{T},1,p+o}(z) \\ \vdots \\ \bar{S}_{\mathfrak{T},p,p+o}(z) \end{pmatrix}.$$

Analogously to Lemma 1, we can show that $\bar{S}_{\mathfrak{T}}^{-1}\bar{B}_{\mathfrak{T}}^{o}(z)m_{\mathfrak{T}}^{(p+o)}(z) \to A_{\mathfrak{T}}^{o} = M^{-1}B^{o}\,\mathrm{E}[m_{\mathfrak{T}}^{(p+o)}(Z_{i\mathfrak{T}})]$, where for $o = 1, \ldots, p'$

$$B^{o} = \begin{pmatrix} M_{0,p+o} \\ M_{1,p+o} \\ \vdots \\ M_{p,p+o} \end{pmatrix}.$$

Additionally, the central limit theorem will guarantee, using similar arguments as in Lemmas 3 and 4, that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\bar{S}_{\mathfrak{T}}^{-1}\bar{B}_{\mathfrak{T}}^{o}(Z_{i\mathfrak{T}})m_{\mathfrak{T}}^{(p+o)}(Z_{i\mathfrak{T}}) - A_{\mathfrak{T}}^{o}] = O_p(1).$$

Hence, we can write

$$\hat{\beta}_{\mathfrak{T}}(z) - \beta_{\mathfrak{T}}(z) = \bar{S}_{\mathfrak{T}}^{-1}(z)\bar{\tau}_{\mathfrak{T}}(z) \tag{S.35}$$

$$+ h_n^{p+1}A_{\mathfrak{T}}^{1} + \ldots + h_n^{p+p'}A_{\mathfrak{T}}^{p'} + O_p(n^{-1/2}h_n^{p+1}) + O_p(h_n^{p+p'+1}). \tag{S.36}$$

Next, to define the bias-correction procedure, let us assume without loss of generality there is only a single scalar parameter $\beta_{\mathfrak{T}}(z)$ to estimate and consider estimators based on bandwidths $h_{1n}, \ldots, h_{p'n}$ that differ from h_n. Then for $q = 1, \ldots, p'$

$$\hat{\beta}_{q\mathfrak{T}}(z) - \beta_{\mathfrak{T}}(z) = \bar{S}_{q\mathfrak{T}}^{-1}(z)\bar{\tau}_{q\mathfrak{T}}(z) + h_{qn}^{p+1}A_{\mathfrak{T}}^{1} + \ldots + h_{qn}^{p+p'}A_{\mathfrak{T}}^{p'} + O_p(n^{-1/2}h_{qn}^{p+1}) + O_p(h_{qn}^{p+p'+1}),$$

where the middle terms can be written as $(h_{qn}^{p+1}, \ldots, h_{qn}^{p+p'})(A_{\mathfrak{T}}^{1}, \ldots, A_{\mathfrak{T}}^{p'})^{\top}$. Accumulating these terms together for $q = 1, \ldots, p'$ results in

$$\begin{pmatrix} h_{1n}^{p+1} & \cdots & h_{1n}^{p+p'} \\ \vdots & \ddots & \vdots \\ h_{p'n}^{p+1} & \cdots & h_{p'n}^{p+p'} \end{pmatrix} \begin{pmatrix} A_{\mathfrak{T}}^{1} \\ \vdots \\ A_{\mathfrak{T}}^{p'} \end{pmatrix}.$$

To eliminate these higher-order bias terms, we can compute a weighted average of estimates

$\hat{\beta}_{q\mathfrak{T}}(z)$,$q = 1, \ldots, p'$, with weights $w_1, \ldots, w_{p'}$ such that

$$\begin{pmatrix} h_{1n}^{p+1} & \cdots & h_{p'n}^{p+1} \\ \vdots & \ddots & \vdots \\ h_{1n}^{p+p'} & \cdots & h_{p'n}^{p+p'} \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_{p'} \end{pmatrix} = \begin{pmatrix} h_n^{p+1} \\ \vdots \\ h_n^{p+p'} \end{pmatrix},$$

that is, after dividing each equation by h_n^{p+q} and denoting ratios $c_q = h_{qn}/h_n$ for $q = 1, \ldots, p'$, such that

$$\mathcal{C}_{p'} \begin{pmatrix} w_1 \\ \vdots \\ w_{p'} \end{pmatrix} := \begin{pmatrix} c_1^{p+1} & \cdots & c_{p'}^{p+1} \\ \vdots & \ddots & \vdots \\ c_1^{p+p'} & \cdots & c_{p'}^{p+p'} \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_{p'} \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} =: \iota_{p'}. \tag{S.37}$$

If $\mathcal{C}_{p'}$ is non-singular with finite elements, defining weights by $w = (w_0, w_1, \ldots, w_{p'}) = (-1, (\mathcal{C}_{p'}^{-1}\iota_{p'})^\top)^\top$ guarantees that

$$w^\top \begin{pmatrix} h_n^{p+1} & \cdots & h_n^{p+p'} \\ h_{1n}^{p+1} & \cdots & h_{1n}^{p+p'} \\ \vdots & \ddots & \vdots \\ h_{p'n}^{p+1} & \cdots & h_{p'n}^{p+p'} \end{pmatrix} \begin{pmatrix} A_{\mathfrak{T}}^1 \\ \vdots \\ A_{\mathfrak{T}}^{p'} \end{pmatrix} = 0,$$

and after labelling $\hat{\beta}_{0\mathfrak{T}}(z) = \hat{\beta}_{\mathfrak{T}}(z)$ and $\tilde{w}_q = w_q / \sum_{q=0}^{p'} w_q$ that

$$\sum_{q=0}^{p'} \tilde{w}_q [\hat{\beta}_{q\mathfrak{T}}(z) - \beta_{\mathfrak{T}}(z)] = \sum_{q=0}^{p'} \tilde{w}_q \bar{S}_{q\mathfrak{T}}^{-1}(z) \bar{\tau}_{q\mathfrak{T}}(z) + O_p(n^{-1/2} h_n^{p+1}) + O_p(h_n^{p+p'+1}).$$

Hence, the higher order bias terms are eliminated and the final estimator is asymptotically normal without asymptotic bias as long as $\sqrt{n} h_n^{p+p'} \to 0$.

The bias-correction procedure of an estimator $\hat{\beta}_{0\mathfrak{T}}(z) = \hat{\beta}_{\mathfrak{T}}(z)$ with bandwidth h_n is thus based on choosing constants $c_1, \ldots, c_{p'}$ (all different from 1 and each other), computing corresponding estimates $\hat{\beta}_{q\mathfrak{T}}(z)$ with bandwidths $c_q h_n$, $q = 1, \ldots, p'$, and finally, computing the weighted average of $\hat{\beta}_{q\mathfrak{T}}(z), q = 0, \ldots, p'$, using weights $w = (w_0, w_1, \ldots, w_{p'}) = (-1, (\mathcal{C}_{p'}^{-1}\iota_{p'})^\top)^\top$ solving equation (S.37). Note that, due to the asymptotic linearity (A.21) derived in the proof of Theorem 4, the same procedure can be also applied to the coefficient vectors B.

Let us provide a practical example related to the examples used in Section 4, where we used $p' = 3$ and $c_1 = 1.3$, $c_2 = 1.6$, and $c_3 = 1.9$. In applications, one might prefer local quadratic regression with $p = 2$. Assuming three explanatory variables $d = 3$, the bandwidth h_n has to satisfy the condition $nh_n^{2d+3}/\ln n \to \infty$, which implies $h_n = O(n^{-1/(9+\epsilon)})$, $\epsilon > 0$. On the other hand, $nh_n^{2(p+p')} \to 0$ if the bias is to be asymptotically negligible or $nh_n^{2(p+p'+1)} \to 0$ if the highest-order bias can be present as in Theorem 2. This implies that the number of auxiliary estimates required for the bias correction is equal to $p' = 3$ or $p' = 2$. Similarly for four explanatory variables, $d = 4$, $h_n = O(n^{-1/(11+\varepsilon)})$, $\epsilon > 0$, and $p' = 4$ or $p' = 3$.

G Discrete explanatory variables

The presented results were obtained for continuously distributed random variables X_{it}. There are however several possibilities though to accommodate discrete variables. First, let us note that traditionally used dummy variables capturing time-specific effects do not have to be explicitly included in the model as the regression function (1) is time-period specific. Next, the conditional expectations $\mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}]$ can be estimated even in the presence of discrete explanatory variables, for example, by sample splitting or the nonparametric regression of Racine and Li (2004). More specifically, it is possible to partition the data by the values of the discrete variables $\tilde{D}$ included in the regression model, construct the moment conditions in Theorem 1 for each partition, and use the constructed moment conditions jointly for the GMM estimation discussed in Section 3.2. This procedure will identify and provide consistent estimates of the corresponding index coefficients for the continuously distributed variables while accounting for the effect of discrete control variables.

Suppose now that there are not only discrete control variables, but also discrete variables of interest, for which we want to identify their coefficients. To distinguish between continuous and discrete covariates, the data generating process is now assumed to be

$$Y_{it} = \phi(X_{it}^\top(\beta_1, \ldots, \beta_R) + \tilde{D}_{it}^\top(\eta_1, \ldots, \eta_R), \alpha_i, U_{it}), \tag{S.38}$$

where X_{it} denotes a $d \times 1$ vector of continuous random variables, the R linear combinations $B = (\beta_1, \ldots, \beta_R)$ can be identified and estimated as discussed in the previous paragraph, $\tilde{D}_{it}$ denotes a $l \times 1$ vector of discrete random variables, and the corresponding coefficients are $\eta_1, \ldots, \eta_R$.

Given the values of the coefficients B identified using Theorem 1, it is possible to construct additional moment conditions to identify the coefficients of discrete explanatory variables as proposed by Horowitz and Härdle (1996) and adapted to nonseparable panel-data models by Čížek and Lei (2018, Appendix D). The solution based on the identification Assumption 3 is to compare pairs of observations that have the same sum of discrete variables in two different time periods, but that have different values of discrete variables in the current time period. For example, consider the simplest case with two time periods t and $t - \Delta$ and a single discrete variable being a dummy with values 0 and 1. In this case, all possible observations of

$(\tilde{D}_{it} = \tilde{d}_t, \tilde{D}_{i(t-\Delta)} = \tilde{d}_{(t-\Delta)})$ are $(0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$. The aim is to compare pairs of observations such that $\tilde{d}_t + \tilde{d}_{(t-\mathfrak{T})}$ are equal, but $\tilde{d}_t$ are different, that is, to compare pairs of observations with $(\tilde{D}_{it} = 0, \tilde{D}_{i(t-\Delta)} = 1)$ and $(\tilde{D}_{it} = 1, \tilde{D}_{i(t-\Delta)} = 0)$.

Let us first introduce the notation. As in Horowitz and Härdle (1996), define $S_{\tilde{d}} \equiv \{\tilde{d}^{(s)} : s = 1, \ldots, N_{\tilde{d}}\}$ to be the support of the discrete random vector $\tilde{D}_{it}$ and $S_{\bar{d}} \equiv \{\bar{d}^{(s)} : s = 1, \ldots, N_{\bar{d}}\}$ to be the support of $\tilde{D}_{it} + \tilde{D}_{i(t-\Delta)}$. Let us also define the following index set $\mathcal{I}_s = \{(\iota_1, \iota_2) : \iota_1 > \iota_2$ and $\bar{d}^{(s)} = \tilde{d}^{(\iota_1)} + \tilde{d}^{(\iota_1')} = \tilde{d}^{(\iota_2)} + \tilde{d}^{(\iota_2')}$ for some $\iota_1', \iota_2' \in \{1, \ldots, N_{\tilde{d}}\}\}$ of pairs of values that can be used for comparing and differencing out the individual effects. Finally, we introduce the abbreviated notation for indices $XD^r_{it}(v) = v + X^\top_{it}(\beta_1, \ldots, \beta_{r-1}, \beta_{r+1}, \ldots, \beta_R) + \tilde{D}^\top_{it}(\eta_1, \ldots, \eta_{r-1}, \eta_{r+1}, \ldots, \eta_R)$, $XD^r_{i(t-\Delta)}(v) = v + X^\top_{it}(\beta_1, \ldots, \beta_{r-1}, \beta_{r+1}, \ldots, \beta_R) + \tilde{D}^\top_{it}(\eta_1, \ldots, \eta_{r-1}, \eta_{r+1}, \ldots, \eta_R)$, and $XD^r_{i\mathfrak{T}}(v, \bar{v}) = (XD^r_{it}(v), XD^r_{i(t-\Delta)}(\bar{v} - v))$; in all cases, v and $\bar{v}$ represent the values of the rth index $X^\top_{it}\beta_r$ and $X^\top_{it}\beta_r + X^\top_{i(t-\Delta)}\beta_r$, respectively. The corresponding symbols $xd^r_{it}(v, \bar{v})$, $xd^r_{i(t-\Delta)}(v, \bar{v})$, and $xd^r_{i\mathfrak{T}}(v, \bar{v})$ correspond to the realizations of the random variables $XD^r_{it}(v, \bar{v})$, $XD^r_{i(t-\Delta)}(v, \bar{v})$, and $XD^r_{\mathfrak{T}}(v, \bar{v})$, respectively.

To describe the principles, on which the identification is based, let for $(\iota_1, \iota_2) \in \mathcal{I}_s$, $s \in S_{\bar{d}}$ and $k = 1, 2$

$$
\begin{aligned}
&G^r_{\iota_k,s}(xd^r_{i\mathfrak{T}}(v, \bar{v})) = G^r_{\iota_k,s}(xd^r_{it}(v), xd^r_{i(t-\Delta)}(\bar{v} - v)) \\
&\equiv E[Y_{it} | X^\top_{it}(\beta_1, \ldots, \beta_R) + \tilde{D}^\top_{it}(\eta_1, \ldots, \eta_R) = xd_{it}(v), \tilde{D}_{it} = \tilde{d}^{(\iota_1)}, \\
&\qquad X^\top_{i(t-\Delta)}(\beta_1, \ldots, \beta_R) + \tilde{D}^\top_{i(t-\Delta)}(\eta_1, \ldots, \eta_R) = xd_{i(t-\Delta)}(\bar{v} - v), \tilde{D}_{it} + \tilde{D}_{i(t-\Delta)} = \bar{d}^{(s)}].
\end{aligned}
$$

To identify the coefficients of the rth linear combination, let us impose the following weak monotonicity condition on the rth index given the values of the remaining indices $XD^r_{i\mathfrak{T}} = XD^r_{i\mathfrak{T}}(0, 0)$. Assume that there are finite numbers $v^r_0(XD^r_{i\mathfrak{T}}), v^r_1(XD^r_{i\mathfrak{T}}), c^r_0(XD^r_{i\mathfrak{T}})$, and $c^r_1(XD^r_{i\mathfrak{T}})$ such that we have $v^r_0(XD^r_{i\mathfrak{T}}) < v^r_1(XD^r_{i\mathfrak{T}})$, $c^r_0(XD^r_{i\mathfrak{T}}) < c^r_1(XD^r_{i\mathfrak{T}})$, and for any given $\tilde{d} \in S_{\tilde{d}}$ and $\bar{d} \in S_{\bar{d}}$, $G^r_{\iota_k,s}(XD^r_{i\mathfrak{T}}(v + \tilde{d}^\top \eta_r, \bar{v})) < c^r_0(XD^r_{i\mathfrak{T}})$ if $v < v^r_0(XD^r_{i\mathfrak{T}})$, $G^r_{\iota_k,s}(XD^r_{i\mathfrak{T}}(v + \tilde{d}^\top \eta_r, \bar{v})) > c^r_1(XD^r_{i\mathfrak{T}})$ if $v > v^r_1(XD^r_{i\mathfrak{T}})$. Note that the estimates of $G^r_{\iota_k,s}(XD^r_{i\mathfrak{T}}(v + \tilde{d}^\top \eta_r, \bar{v}))$ for all values of $\tilde{d}$ are already known from the estimation steps neccessary to obtain $(\beta_1, \ldots, \beta_R)$ described at the start of this section, and it is thus possible to check whether the chosen values $v^r_0(XD^r_{i\mathfrak{T}}), v^r_1(XD^r_{i\mathfrak{T}}), c^r_0(XD^r_{i\mathfrak{T}}), c^r_1(XD^r_{i\mathfrak{T}})$ satisfy the stated conditions given the considered parameter space. Additionally, we also assume that the densities of $X^\top_{it}\beta_r$ and of $X^\top_{i(t-\Delta)}\beta_r$

conditional on $\tilde{D}_{it} = \tilde{d}$, $\tilde{D}_{it} + \tilde{D}_{i(t-\mathfrak{T})} = \bar{d}$, and $XD^r_{i\mathfrak{T}}(0)$ are bounded away from zero on an open interval containing $[v_0, v_1]$.

Using the introduced quantities $v^r_0(XD^r_{i\mathfrak{T}}), v^r_1(XD^r_{i\mathfrak{T}}), c^r_0(XD^r_{i\mathfrak{T}})$, and $c^r_1(XD^r_{i\mathfrak{T}})$, we can now define the following integral

$$\begin{aligned} J^r_{\iota_k,s}(\tilde{d},\bar{d}) \equiv \int_{2v^r_0(XD^r_{i\mathfrak{T}})}^{2v^r_1(XD^r_{i\mathfrak{T}})} \int_{v^r_0(XD^r_{i\mathfrak{T}})}^{v^r_1(XD^r_{i\mathfrak{T}})} \{ & c^r_0(XD^r_{i\mathfrak{T}}) I[G^r_{\iota_k,s}(XD^r_{i\mathfrak{T}}(v + \tilde{d}^\top \eta_r, \bar{v})) < c^r_0(XD^r_{i\mathfrak{T}})] \\ & + c^r_1(XD^r_{i\mathfrak{T}}) I[G^r_{\iota_k,s}(XD^r_{i\mathfrak{T}}(v + \tilde{d}^\top \eta, \bar{v})) > c^r_1(XD^r_{i\mathfrak{T}})] \\ & + G^r_{\iota_k,s}(XD^r_{i\mathfrak{T}}(v + \tilde{d}^\top \eta, \bar{v})) I[c^r_0(XD^r_{i\mathfrak{T}}) \leq G^r_{\iota_k,s}(XD^r_{i\mathfrak{T}}(v + \tilde{d}^\top \eta, \bar{v})) \leq c^r_1(XD^r_{i\mathfrak{T}})]\} dv d\bar{v}. \end{aligned}$$

Considering now all possible indices for $(\iota_1, \iota_2) \in \mathcal{I}_s$, $s \in S_{\bar{d}}$ and $k = 1, 2$, we also define the differences of the integrals $J^r_{\iota_1,s}(\tilde{d},\bar{d})$ and $J^r_{\iota_2,s}(\tilde{d},\bar{d})$ for all possible index combinations that, similarly to Theorem 1, eliminate the effect the individual-specific effects. In particular for any $s \in S_{\bar{d}}$, let $\Delta\tilde{D}_s = (\tilde{d}^{(\iota_1)} - \tilde{d}^{(\iota_2)})_{(\iota_1,\iota_2)\in\mathcal{I}_s}$ and $\Delta J^r_s = (J^r_{\iota_1,s}(\tilde{d}^{(\iota_1)}, \bar{d}^{(s)}) - J^r_{\iota_2,s}(\tilde{d}^{(\iota_2)}, \bar{d}^{(s)}))_{(\iota_1,\iota_2)\in\mathcal{I}_s}$, and additionally, let matrices $\Delta\tilde{D} = (\Delta\tilde{D}_1^\top, \ldots, \Delta\tilde{D}_{N_{\bar{d}}}^\top)^\top$ and $\Delta J^r = (\Delta J^r_1, \ldots, \Delta J^r_{N_{\bar{d}}})^\top$. Under the above mentioned assumptions, Čížek and Lei (2018, Lemma 1) showed that

$$2(v^r_1(XD^r_{i\mathfrak{T}}) - v^r_0(XD^r_{i\mathfrak{T}}))(c^r_1(XD^r_{i\mathfrak{T}}) - c^r_0(XD^r_{i\mathfrak{T}}))(\Delta\tilde{D}^\top\Delta\tilde{D})\eta_r = \Delta\tilde{D}^\top\Delta J^r, \tag{S.39}$$

which can be constructed for any $r = 1, \ldots, R$. If $\Delta\tilde{D}^\top\Delta\tilde{D}$ is a nonsingular matrix, it is thus possible to construct a system of equations

$$2(v^r_1(XD^r_{i\mathfrak{T}}) - v^r_0(XD^r_{i\mathfrak{T}}))(c^r_1(XD^r_{i\mathfrak{T}}) - c^r_0(XD^r_{i\mathfrak{T}}))\eta_r = (\Delta\tilde{D}^\top\Delta\tilde{D})^{-1}\Delta\tilde{D}^\top\Delta J^r,$$

$r = 1, \ldots, R$ to identify parameters $\eta_1, \ldots, \eta_R$ since $v^r_1(XD^r_{i\mathfrak{T}}) - v^r_0(XD^r_{i\mathfrak{T}}) > 0$ and $c^r_1(XD^r_{i\mathfrak{T}}) - c^r_0(XD^r_{i\mathfrak{T}}) > 0$.

Based on this result, we can obtain consistent estimates of $\eta_1, \ldots, \eta_R$ by replacing $G^r_{\iota_k,s}(xd^r_{i\mathfrak{T}}(v, \bar{v}))$ by its nonparametric estimate. If $\hat{B}_W$ is the average of the GMM-ADG-OPDG estimators of B proposed in this paper, where the average is taken across all possible values of $\tilde{d}$ and $\bar{d}$, $XD^r_{i\mathfrak{T}}(v, \bar{v})$ can be evaluated at $\hat{B}_W$ and any values of $\eta_1, \ldots, \eta_R$. Consequently, the estimates of $G^r_{\iota_k,s}(XD^r_{i\mathfrak{T}}(v, \bar{v}))$ can be estimated by nonparametric regression method discussed in Section

3. Denoting the estimator of $G^r_{\iota_k,s}(XD^r_{i\mathfrak{T}}(v,\bar{v}))$ by $\hat{G}^r_{\iota_k,s}(XD^r_{i\mathfrak{T}}(v,\bar{v}))$,

$$\begin{aligned}\hat{J}^r_{\iota_k,s}(\tilde{d},\bar{d}) \equiv \int_{2v_0^r(XD^r_{i\mathfrak{T}})}^{2v_1^r(XD^r_{i\mathfrak{T}})} \int_{v_0^r(XD^r_{i\mathfrak{T}})}^{v_1^r(XD^r_{i\mathfrak{T}})} \{c_0^r(XD^r_{i\mathfrak{T}})I[\hat{G}^r_{\iota_k,s}(XD^r_{i\mathfrak{T}}(v+\tilde{d}^\top\eta_r,\bar{v})) < c_0^r(XD^r_{i\mathfrak{T}})] \\ + c_1^r(XD^r_{i\mathfrak{T}})I[\hat{G}^r_{\iota_k,s}(XD^r_{i\mathfrak{T}}(v+\tilde{d}^\top\eta,\bar{v})) > c_1^r(XD^r_{i\mathfrak{T}})] \\ + \hat{G}^r_{\iota_k,s}(XD^r_{i\mathfrak{T}}(v+\tilde{d}^\top\eta,\bar{v}))I[c_0^r(XD^r_{i\mathfrak{T}}) \le \hat{G}^r_{\iota_k,s}(XD^r_{i\mathfrak{T}}(v+\tilde{d}^\top\eta,\bar{v})) \le c_1^r(XD^r_{i\mathfrak{T}})]\}dvd\bar{v},\end{aligned}$$

$\Delta\hat{J}^r_s = (\hat{J}^r_{\iota_1,s}(\tilde{d}^{(\iota_1)},\bar{d}^{(s)}) - \hat{J}^r_{\iota_2,s}(\tilde{d}^{(\iota_2)},\bar{d}^{(s)}))_{(\iota_1,\iota_2)\in\mathcal{I}_s}$, and $\Delta J^r = (\Delta J^r_1,\ldots,\Delta J^r_{N_{\bar{d}}})^\top$, the estimation of $\eta_1,\ldots,\eta_R$ based on (S.39) can be then performed by a GMM estimator based on the moment equations

$$2(v_1^r(XD^r_{i\mathfrak{T}}) - v_0^r(XD^r_{i\mathfrak{T}}))(c_1^r(XD^r_{i\mathfrak{T}}) - c_0^r(XD^r_{i\mathfrak{T}}))\eta_r - (\Delta\tilde{D}^\top\Delta\tilde{D})^{-1}\Delta\tilde{D}^\top\Delta\hat{J}^r = 0, \quad \text{(S.40)}$$

$r = 1,\ldots,R$.

H Application: quadratic functions

Using the application in Section 5, where the dependent variable possibly depends on the linear combination of explanatory variables and their squares or their interactions, let us demonstrate how the proposed GMM estimator can be adjusted to estimation of coefficients in a linear combination of non-linear transformations of explanatory variables such as quadratic or interaction terms. We will show on the example of our application, that the additional coefficients corresponding to the labor market experience and its square can be identified if we compute ADG and OPDG not only for the conditional expectation $\mathrm{E}[Y_{it}|X_{it}, X_{i(t-\Delta)}]$, but also for $\mathrm{E}[X_{it,3}Y_{it}|X_{it}, X_{i(t-\Delta)}]$ and $\mathrm{E}[X_{it,3}^2 Y_{it}|X_{it}, X_{i(t-\Delta)}]$, where $X_{it} = (X_{it,1}, X_{it,2}, X_{it,3})$ and $X_{it,3}$ represent the labor market experience.

Let us consider the model in Section 5 with a quadratic term:

$$\begin{aligned} Y_{it,2} &= \phi_{t2}(X_{it,1}\beta_{11} + X_{it,2}\beta_{12} + X_{it,3}\beta_{13} + X_{it,3}^2\beta_{14}, X_{it,1}\beta_{21} + X_{it,2}\beta_{22} + X_{it,3}\beta_{23} + X_{it,3}^2\beta_{24}, \alpha_i, U_{it}), \\ Y_{it,1} &= \phi_{t1}(X_{it,1}\beta_{11} + X_{it,2}\beta_{12} + X_{it,3}\beta_{13} + X_{it,3}^2\beta_{14}, \alpha_i, U_{it}), \end{aligned} \tag{S.41}$$

where $X_{it,1}$, $X_{it,2}$, and $X_{it,3}$ represent the lagged dependent variable, the number of children, and the labor market experience in Section 5. Recall that $B = ((\beta_{21}, \beta_{22}, \beta_{23}, \beta_{24})^\top, (\beta_{11}, \beta_{12}, \beta_{13}, \beta_{14})^\top)$ and the coefficients are normalized to $\beta_{11} = 0$, $\beta_{12} = 1$, $\beta_{21} = 1$, and $\beta_{22} = 0$. As follows from the proof of Theorem 1, we can write – for clarity of exposition separately for the selection decision $Y_{it,1}$ and for the outcome variable $Y_{it,2}$ and their ADG ($\delta_{\mathfrak{X}}^1$,$\delta_{\mathfrak{X}\mathfrak{X}}^1$) and OPDG ($\delta_{\mathfrak{X}}^2$,$\delta_{\mathfrak{X}\mathfrak{X}}^2$) –

$$\delta_{\mathfrak{X}}^1 = \sum_{r=1}^{1} \mathrm{E}\left\{\varphi_{t1,1}'(X_{it}^\top B, \alpha_i)(0, 1, (\beta_{13} + 2\beta_{14}X_{it,3}))^\top\right\} \tag{S.42}$$

$$\begin{aligned} \delta_{\mathfrak{X}}^2 &= \sum_{r=1}^{2} \mathrm{E}\left\{\varphi_{t2,r}'(X_{it}^\top B, \alpha_i)(\beta_{r1}, \beta_{r2}, \beta_{r3} + 2\beta_{r4}X_{it,3})^\top\right\} \\ &= \mathrm{E}\left\{\varphi_{t2,1}'(X_{it}^\top B, \alpha_i)(0, 1, (\beta_{13} + 2\beta_{14}X_{it,3})^\top\right\} + \mathrm{E}\left\{\varphi_{t2,2}'(X_{it}^\top B, \alpha_i)(1, 0, \beta_{23} + 2\beta_{24}X_{it,3})^\top\right\} \\ \delta_{\mathfrak{X}\mathfrak{X}}^1 &= \sum_{r,s=1}^{1} \mathrm{E}\left\{(0, 1, \beta_{13} + 2\beta_{14}X_{it,3})\left\{\varphi_{t1,r}'(X_{it}^\top B, \alpha_i)\varphi_{t1,s}'(X_{it}^\top B, \alpha_i)\right\}_{r,s=1}^{R}(0, 1, \beta_{13} + 2\beta_{14}X_{it,3})^\top\right\} \\ \delta_{\mathfrak{X}\mathfrak{X}}^2 &= \sum_{r,s=1}^{2} \mathrm{E}\left\{(\beta_{r1}, \beta_{r2}, \beta_{r3} + 2\beta_{r4}X_{it,3})[\varphi_{t2,r}'(X_{it}^\top B, \alpha_i)\varphi_{t2,s}'(X_{it}^\top B, \alpha_i)](\beta_{r1}, \beta_{r2}, (\beta_{r3} + 2\beta_{r4}X_{it,3})^\top\right\}. \end{aligned} \tag{S.43}$$

Considering for simplicity on the ADG moment conditions, let us first explain why these equations do not provide enough conditions for the identification of all coefficients. Given that there are 3 explanatory variables, it is clear that $\delta^1_{\mathfrak{T}}$ and $\delta^2_{\mathfrak{T}}$ provide only 6 moment conditions not only for 4 β-parameters $\beta_{13}, \beta_{14}, \beta_{23}, \beta_{24}$ (some of them can be of course identified using $\delta^1_{\mathfrak{T}\mathfrak{T}}$ and $\delta^2_{\mathfrak{T}\mathfrak{T}}$), but also 6 γ-parameters – $\gamma_{11} = \mathrm{E}\,\varphi'_{t1,1}(X_{it}^\top B, \alpha_i)$, $\gamma_{21} = \mathrm{E}\,\varphi'_{t2,1}(X_{it}^\top B, \alpha_i)$, $\gamma_{22} = \mathrm{E}\,\varphi'_{t2,2}(X_{it}^\top B, \alpha_i)$ – and the corresponding interactions with $X_{it,3}$ – $\gamma^X_{11} = \mathrm{E}\,X_{it,3}\varphi'_{t1,1}(X_{it}^\top B, \alpha_i)$, $\gamma^X_{21} = \mathrm{E}\,X_{it,3}\varphi'_{t2,1}(X_{it}^\top B, \alpha_i)$, $\gamma^X_{22} = \mathrm{E}\,X_{it,3}\varphi'_{t2,2}(X_{it}^\top B, \alpha_i)$. Thus, the ADG moment conditions cannot identify the β-coefficients unless there are some additional moment conditions identifying some of the γ-coefficients such as $\gamma^X_{11}, \gamma^X_{21}, \gamma^X_{22}$. We will show they can be identified using ADG for $\mathrm{E}[X_{it,3}Y_{it}|X_{it}, X_{i(t-\Delta)}]$ and $\mathrm{E}[X^2_{it,3}Y_{it}|X_{it}, X_{i(t-\Delta)}]$; the analogous claim will be made also for the OPDG moment conditions.

Rewriting the ADG moment conditions in (S.42)–(S.43) in a concise way analogously to Theorem 1, we see that

$$\delta^1_{\mathfrak{T}} = \Gamma^1_{1t}(\beta_{11}, \beta_{12}, \beta_{13}) + \Gamma^{X1}_{1t}(\beta_{11}, \beta_{12}, 2\beta_{14}) = \Gamma^1_{1t}(0, 1, \beta_{13}) + \Gamma^{X1}_{1t}(0, 1, 2\beta_{14}) \tag{S.44}$$

$$\begin{aligned}\delta^2_{\mathfrak{T}} &= \Gamma^2_{1t}\begin{pmatrix} \beta_{11} & \beta_{12} & \beta_{13} \\ \beta_{21} & \beta_{22} & \beta_{23} \end{pmatrix} + \Gamma^{X2}_{1t}\begin{pmatrix} \beta_{11} & \beta_{12} & 2\beta_{14} \\ \beta_{21} & \beta_{22} & 2\beta_{24} \end{pmatrix} \\ &= \Gamma^2_{1t}\begin{pmatrix} 0 & 1 & \beta_{13} \\ 1 & 0 & \beta_{23} \end{pmatrix} + \Gamma^{X2}_{1t}\begin{pmatrix} 0 & 1 & 2\beta_{14} \\ 1 & 0 & 2\beta_{24} \end{pmatrix},\end{aligned} \tag{S.45}$$

where $\Gamma^1_{1t} = \mathrm{E}\,\varphi'_{t1,1}(X_{it}^\top B, \alpha_i)$, $\Gamma^2_{1t} = (\mathrm{E}\,\varphi'_{t2,1}(X_{it}^\top B, \alpha_i), \mathrm{E}\,\varphi'_{t2,2}(X_{it}^\top B, \alpha_i))$ and
$\Gamma^{X1}_{1t} = \mathrm{E}[X_{it,3}\varphi'_{t1,1}(X_{it}^\top B, \alpha_i)]$, $\Gamma^{X2}_{1t} = (\mathrm{E}\,X_{it,3}\varphi'_{t2,1}(X_{it}^\top B, \alpha_i), \mathrm{E}\,X_{it,3}\varphi'_{t2,2}(X_{it}^\top B, \alpha_i))$.
Similarly, the OPDG moment conditions can be expressed as

$$\begin{aligned}\delta^1_{\mathfrak{T}\mathfrak{T}} &= (\beta_{11}, \beta_{12}, \beta_{13})^\top \Gamma^1_{2t}(\beta_{11}, \beta_{12}, \beta_{13}) + (\beta_{11}, \beta_{12}, \beta_{13})^\top \Gamma^{X1}_{2t}(\beta_{11}, \beta_{12}, 2\beta_{14}) \\ &\quad + (\beta_{11}, \beta_{12}, 2\beta_{14})^\top \Gamma^{X1}_{2t}(\beta_{11}, \beta_{12}, \beta_{13}) + (\beta_{11}, \beta_{12}, 2\beta_{14})^\top \Gamma^{XX1}_{2t}(\beta_{11}, \beta_{12}, 2\beta_{14}) \\ &= (0, 1, \beta_{13})^\top \Gamma^1_{2t}(0, 1, \beta_{13}) + (0, 1, \beta_{13})^\top \Gamma^{X1}_{2t}(0, 1, 2\beta_{14}) \\ &\quad + (0, 1, 2\beta_{14})^\top \Gamma^{X1}_{2t}(0, 1, \beta_{13}) + (0, 1, 2\beta_{14})^\top \Gamma^{XX1}_{2t}(0, 1, 2\beta_{14}),\end{aligned} \tag{S.46}$$

where $\Gamma^1_{2t} = \mathrm{E}[\varphi'_{t1,1}(X_{it}^\top B, \alpha_i)\varphi'_{t1,1}(X_{it}^\top B, \alpha_i)]$, $\Gamma^{X1}_{2t} = \mathrm{E}[X_{it,3}\varphi'_{t1,1}(X_{it}^\top B, \alpha_i)\varphi'_{t1,1}(X_{it}^\top B, \alpha_i)]$,

$\Gamma_{2t}^{XX1} = \mathrm{E}[X_{it,3}^2 \varphi'_{t1,1}(X_{it}^\top B, \alpha_i)\varphi'_{t1,1}(X_{it}^\top B, \alpha_i)]$, and finally,

$$\delta_{\mathfrak{T}\mathfrak{T}}^2 = \begin{pmatrix} 0 & 1 & \beta_{13} \\ 1 & 0 & \beta_{23} \end{pmatrix}^\top \Gamma_{2t}^2 \begin{pmatrix} 0 & 1 & \beta_{13} \\ 1 & 0 & \beta_{23} \end{pmatrix} + \begin{pmatrix} 0 & 0 & \beta_{13} \\ 0 & 0 & \beta_{23} \end{pmatrix}^\top \Gamma_{2t}^{X2} \begin{pmatrix} 0 & 0 & 2\beta_{14} \\ 0 & 0 & 2\beta_{24} \end{pmatrix} \tag{S.47}$$
$$+ \begin{pmatrix} 0 & 0 & 2\beta_{14} \\ 0 & 0 & 2\beta_{24} \end{pmatrix} \Gamma_{2t}^{X2} \begin{pmatrix} 0 & 0 & \beta_{13} \\ 0 & 0 & \beta_{23} \end{pmatrix} + \begin{pmatrix} 0 & 0 & 2\beta_{14} \\ 0 & 0 & 2\beta_{24} \end{pmatrix}^\top \Gamma_{2t}^{XX2} \begin{pmatrix} 0 & 0 & 2\beta_{14} \\ 0 & 0 & 2\beta_{24} \end{pmatrix},$$

where $\Gamma_{2t}^2 = \{\mathrm{E}[\varphi'_{t2,r}(X_{it}^\top B, \alpha_i)\varphi'_{t2,s}(X_{it}^\top B, \alpha_i)]\}_{r,s=1}^2$, $\Gamma_{2t}^{X2} = \{\mathrm{E}[X_{it,3}\varphi'_{t2,r}(X_{it}^\top B, \alpha_i)\varphi'_{t2,s}(X_{it}^\top B, \alpha_i)]\}_{r,s=1}^2$, and $\Gamma_{2t}^{XX2} = \{\mathrm{E}[X_{it,3}^2\varphi'_{t2,r}(X_{it}^\top B, \alpha_i)\varphi'_{t2,s}(X_{it}^\top B, \alpha_i)]\}_{r,s=1}^2$. Similarly to the ADG moment conditions containing Γ_{1t}^{X1}, Γ_{1t}^{X2}, the OPDG moment conditions thus also contain – next to the β-coefficients β_{13}, β_{23}, β_{14}, and β_{24} – many additional coefficients of matrices Γ_{2t}^{X1}, Γ_{2t}^{XX1}, Γ_{2t}^{X2}, and Γ_{2t}^{XX2}. Given the normalization of the β-coefficients, all these quantities can be identified analogously to Γ_{1t}^1 (which is equal to the second element of $\delta_{\mathfrak{T}}^1$ in (S.44)), Γ_{1t}^2 (which equals to the first two elements of $\delta_{\mathfrak{T}}^2$ in (S.45)), Γ_{2t}^1 (which equals to the second element on the diagonal of $\delta_{\mathfrak{T}\mathfrak{T}}^1$ in (S.46)), and Γ_{2t}^2 (which equals to the 2×2 submatrix of $\delta_{\mathfrak{T}\mathfrak{T}}^2$ in (S.47)). Since Γ_{1t}^{X1}, Γ_{1t}^{X2}, Γ_{2t}^{X1}, Γ_{2t}^{XX1}, Γ_{2t}^{X2}, and Γ_{2t}^{XX2} characterize the expectations of the products of $X_{it,3}$ or $X_{it,3}^2$ with the derivatives of the φ functions, they can directly obtained using the ADG and OPDG corresponding to expectations $\mathrm{E}[X_{it,3}Y_{it}|X_{it}, X_{i(t-\Delta)}]$ and $\mathrm{E}[X_{it,3}^2 Y_{it}|X_{it}, X_{i(t-\Delta)}]$, that is, using the following ADG moments ($c = 1, 2$)

$$\delta_{\mathfrak{T}}^{Xc} = \mathrm{E}\left\{ \frac{\partial}{\partial X_{it}^\top} \mathrm{E}[X_{it,3}Y_{it,c}|X_{it}, X_{i(t-\Delta)}] - \frac{\partial}{\partial X_{i(t-\Delta)}^\top} \mathrm{E}[X_{it,3}Y_{it,c}|X_{it}, X_{i(t-\Delta)}] \right\},$$

$$\delta_{\mathfrak{T}}^{XXc} = \mathrm{E}\left\{ \frac{\partial}{\partial X_{it}^\top} \mathrm{E}[X_{it,3}^2Y_{it,c}|X_{it}, X_{i(t-\Delta)}] - \frac{\partial}{\partial X_{i(t-\Delta)}^\top} \mathrm{E}[X_{it,3}^2Y_{it,c}|X_{it}, X_{i(t-\Delta)}] \right\},$$

and the following OPG moments ($c = 1, 2$)

$$\delta_{\mathfrak{T}\mathfrak{T}}^{Xc} = \mathrm{E}\left[\left\{ \frac{\partial}{\partial X_{it}^\top} \mathrm{E}[X_{it,3}Y_{it,c}|X_{it}, X_{i(t-\Delta)}] - \frac{\partial}{\partial X_{i(t-\Delta)}^\top} \mathrm{E}[X_{it,3}Y_{it,c}|X_{it}, X_{i(t-\Delta)}] \right\}^\top \right.$$
$$\left. \times \left\{ \frac{\partial}{\partial X_{it}^\top} \mathrm{E}[X_{it,3}Y_{it,c}|X_{it}, X_{i(t-\Delta)}] - \frac{\partial}{\partial X_{i(t-\Delta)}^\top} \mathrm{E}[X_{it,3}Y_{it,c}|X_{it}, X_{i(t-\Delta)}] \right\} \right],$$

$$\delta^{XXc}_{\mathfrak{X}\mathfrak{X}} = \mathrm{E}\left[\left\{\frac{\partial}{\partial X_{it}^{\top}}\mathrm{E}[X_{it,3}^2 Y_{it,c}|X_{it}, X_{i(t-\Delta)}] - \frac{\partial}{\partial X_{i(t-\Delta)}^{\top}}\mathrm{E}[X_{it,3}^2 Y_{it,c}|X_{it}, X_{i(t-\Delta)}]\right\}^{\top}\right.$$
$$\left.\times\left\{\frac{\partial}{\partial X_{it}^{\top}}\mathrm{E}[X_{it,3}^2 Y_{it,c}|X_{it}, X_{i(t-\Delta)}] - \frac{\partial}{\partial X_{i(t-\Delta)}^{\top}}\mathrm{E}[X_{it,3}^2 Y_{it,c}|X_{it}, X_{i(t-\Delta)}]\right\}\right];$$

in this particular application, the derivatives used for Γ_{1t}^{X1}, Γ_{1t}^{X2}, Γ_{2t}^{X1}, Γ_{2t}^{XX1}, Γ_{2t}^{X2}, and Γ_{2t}^{XX2} are the derivatives with respect to variables $X_{it,1}$ and $X_{it,2}$.

Once the all Γ's are identified, there are two ADG (S.44)–(S.45) and at least two OPDG conditions (S.46)–(S.47) corresponding to the derivatives with respect to $X_{it,3}$ available for the identification of four β-coefficients. In the applications, where the moment conditions (S.44)–(S.47) would not be sufficient to identify all β-coefficients, the moments based on $\delta^{Xc}_{\mathfrak{X}}$ and $\delta^{Xc}_{\mathfrak{X}\mathfrak{X}}$ can also provide further moment conditions for the identification of the β-coefficients if the derivatives with respect to other variables that do not have normalized coefficient are used; for example, with respect to $X_{it,3}$ in the present application.

I Example: Dynamic sample selection model

In this appendix, we verify Assumption 3(iii) for the dynamic sample selection models with the stationary initial condition for normally-distributed idiosyncratic shocks; the results can be directly generalized to other elliptical distributions. For the sake of simplicity we consider $T = 2$ and $\Delta = 1$ which could be extended to larger T in a straightforward manner as we demonstrate for $T = 3$ and $\Delta = 2$.

Let us consider the selection variable $Y_{1it} = \mathbb{1}\{X_{it}\beta_1 + \alpha_{1i} + U_{1it} > 0\}$ with $\alpha_{1i} = X_{i1} + X_{i2} + \xi_i$ and the outcome variable $Y_{2it} = Y_{2it-1}\beta_2 + \alpha_{2i} + U_{2it}$, where the initial condition is assumed to be $Y_{2i0} = \alpha_{2i}/(1-\beta_2) + U_{2i0}/\sqrt{1-\beta_2^2}$ to ensure the stationarity of the latent process of Y_{2it}. For simplicity, we assume that $(U_{1i1}, U_{1i2}, U_{2i0}, U_{2i1}, U_{2i2}, X_{i1}, X_{i2}, \alpha_{2i}, \xi_i)^T$ has a multivariate normal distribution with zero mean, unit variance, and zero covariance except for U_{1it} and U_{2it}, which can be correlated for $t = 1, 2$. Due to this assumption, $\alpha_{1i}, \alpha_{2i}, X_{i1}, X_{i2}, Y_{2i0}, Y_{2i1}$ are jointly normally distributed and $\alpha_{1i}, \alpha_{2i} \,|X_{i1}, X_{i2}, Y_{2i0}, Y_{2i1}$ has a conditional normal distribution with a variance that does not depend on X_{i1}, X_{i2}, Y_{2i0} and Y_{2i1}. Hence, we only need to consider the mean of the conditional distribution of $\alpha_{1i}, \alpha_{2i} \,|X_{i1}, X_{i2}, Y_{2i0}, Y_{2i1}$ to verify Assumption 3(iii). Note that the estimation is based on three time periods, which means Y_{2i0}, Y_{2i1}, and Y_{2i2} are observed and thus $Y_{1i0} = Y_{1i1} = Y_{1i2} = 1$ here.

First, let us define two matrices $\Sigma_{\alpha_1\alpha_2,XY}$ and Σ_{XY} as follows:

$$\Sigma_{\alpha_1\alpha_2,XY} = \begin{bmatrix} \operatorname{cov}(\alpha_{1i}, X_{i1}) & \operatorname{cov}(\alpha_{1i}, X_{i2}) & \operatorname{cov}(\alpha_{1i}, Y_{2i0}) & \operatorname{cov}(\alpha_{1i}, Y_{2i1}) \\ \operatorname{cov}(\alpha_{2i}, X_{i1}) & \operatorname{cov}(\alpha_{2i}, X_{i2}) & \operatorname{cov}(\alpha_{2i}, Y_{2i0}) & \operatorname{cov}(\alpha_{2i}, Y_{2i1}) \end{bmatrix},$$

$$\Sigma_{XY} = \begin{bmatrix} \operatorname{var}(X_{i1}) & \operatorname{cov}(X_{i1}, X_{i2}) & \operatorname{cov}(X_{i1}, Y_{2i0}) & \operatorname{cov}(X_{i1}, Y_{2i1}) \\ \operatorname{cov}(X_{i2}, X_{i1}) & \operatorname{var}(X_{i2}) & \operatorname{cov}(X_{i2}, Y_{2i0}) & \operatorname{cov}(X_{2i}, Y_{2i1}) \\ \operatorname{cov}(Y_{2i0}, X_{i1}) & \operatorname{cov}(Y_{2i0}, X_{i2}) & \operatorname{var}(Y_{2i0}) & \operatorname{cov}(Y_{2i0}, Y_{2i1}) \\ \operatorname{cov}(Y_{2i1}, X_{i1}) & \operatorname{cov}(Y_{2i1}, X_{i2}) & \operatorname{cov}(Y_{2i1}, Y_{2i0}) & \operatorname{var}(Y_{2i1}) \end{bmatrix}.$$

To determine the elements of the above matrices, we start with $\Sigma_{\alpha_1\alpha_2,XY}$. First, $\operatorname{cov}(\alpha_{1i}, X_{i1}) = \operatorname{cov}(X_{i1} + X_{i2} + \xi_i, X_{i1}) = 1$ and $\operatorname{cov}(\alpha_{1i}, X_{i2}) = \operatorname{cov}(X_{i1} + X_{i2} + \xi_i, X_{i2}) = 1$. Furthermore, $\operatorname{cov}(\alpha_{1i}, Y_{2i0}) = \operatorname{cov}(\alpha_{1i}, Y_{2i1}) = \operatorname{cov}(\alpha_{2i}, X_{i1}) = \operatorname{cov}(\alpha_{2i}, X_{i2}) = 0$. Finally, $\operatorname{cov}(\alpha_{2i}, Y_{2i0}) = \operatorname{cov}\left(\alpha_{2i}, \alpha_{2i}/(1-\beta_2) + U_{2i0}/\sqrt{1-\beta_2^2}\right) = 1/(1-\beta_2)$ and

$\text{cov}(\alpha_{2i}, Y_{2i1}) = \text{cov}(\alpha_{2i}, Y_{2i0}\beta_2 + \alpha_{2i} + U_{2i1}) = \beta_2/(1-\beta_2) + 1 = 1/(1-\beta_2)$.

Now consider the elements of Σ_{XY}. First note that $var(X_{i1}, X_{i1}) = \text{var}(X_{i2}, X_{i2}) = 1$, $\text{cov}(X_{i1}, X_{i2}) = 0$ and $\text{cov}(X_{ij}, Y_{2i0}) = \text{cov}(X_{ij}, Y_{2i1}) = 0$ for $j = 1, 2$. Next,

$$\text{var}(Y_{2i0}) = \frac{1}{(1-\beta_2)^2}\text{var}(\alpha_{2i}) + \frac{1}{1-\beta_2^2}\text{var}(U_{2i0}) = \frac{1}{(1-\beta_2)^2} + \frac{1}{1-\beta_2^2} = \frac{2}{(1-\beta_2)^2(1+\beta_2)},$$

$$\begin{aligned}\text{var}(Y_{2i1}) &= \beta_2^2\text{var}(Y_{2i0}) + \text{var}(\alpha_{2i}) + \text{var}(U_{2i1}) \\ &\quad + 2\text{cov}(Y_{2i0}\beta_2, \alpha_{2i}) + 2\text{cov}(Y_{2i0}\beta_2, U_{2i1}) + 2\text{cov}(U_{2i1}, \alpha_{2i}) \\ &= \frac{2\beta_2^2}{(1-\beta_2)^2(1+\beta_2)} + 2 + \frac{2\beta_2}{1-\beta_2} = \frac{2}{(1-\beta_2)^2(1+\beta_2)},\end{aligned}$$

and

$$\begin{aligned}\text{cov}(Y_{2i0}, Y_{2i1}) &= \text{cov}(Y_{2i0}, Y_{2i0}\beta_2 + \alpha_{2i} + U_{2i1}) \\ &= \beta_2\text{var}(Y_{i0}) + \text{cov}(Y_{i0}, \alpha_{2i}) \\ &= \frac{2\beta_2}{(1-\beta_2)^2(1+\beta_2)} + \frac{1}{1-\beta_2} \\ &= \frac{-\beta_2^2 + 2\beta_2 + 1}{(1-\beta_2)^2(1+\beta_2)}.\end{aligned}$$

Given matrices $\Sigma_{\alpha_1\alpha_2,XY}$ and Σ_{XY} with their elements derived above, the mean of the

conditional distribution of $\alpha_{1i}, \alpha_{2i} \,|X_{i1}, X_{i2}, Y_{2i0}, Y_{2i1}$ is given by

$$\Sigma_{\alpha_1\alpha_2,XY}\Sigma_{XY}^{-1}\begin{bmatrix} x_{i1} \\ x_{i2} \\ y_{2i0} \\ y_{2i1} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{1-\beta_2} & \frac{1}{1-\beta_2} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{2}{(1-\beta_2)^2(1+\beta_2)} & \frac{-\beta_2^2+2\beta_2+1}{(1-\beta_2)^2(1+\beta_2)} \\ 0 & 0 & \frac{-\beta_2^2+2\beta_2+1}{(1-\beta_2)^2(1+\beta_2)} & \frac{2}{(1-\beta_2)^2(1+\beta_2)} \end{bmatrix}^{-1} \begin{bmatrix} x_{i1} \\ x_{i2} \\ y_{2i0} \\ y_{2i1} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{1-\beta_2} & \frac{1}{1-\beta_2} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{2}{3-\beta_2} & \frac{\beta_2^2-2\beta_2-1}{3-\beta_2} \\ 0 & 0 & \frac{\beta_2^2-2\beta_2-1}{3-\beta_2} & \frac{2}{3-\beta_2} \end{bmatrix} \begin{bmatrix} x_{i1} \\ x_{i2} \\ y_{2i0} \\ y_{2i1} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & \frac{\beta_2^2-2\beta_2+1}{(3-\beta_2)(1-\beta_2)} & \frac{\beta_2^2-2\beta_2+1}{(3-\beta_2)(1-\beta_2)} \end{bmatrix} \begin{bmatrix} x_{i1} \\ x_{i2} \\ y_{2i0} \\ y_{2i1} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & \frac{1-\beta_2}{3-\beta_2} & \frac{1-\beta_2}{3-\beta_2} \end{bmatrix} \begin{bmatrix} x_{i1} \\ x_{i2} \\ y_{2i0} \\ y_{2i1} \end{bmatrix} = \begin{bmatrix} x_{i1} + x_{i2} \\ \frac{1-\beta_2}{3-\beta_2}\left(y_{2i0} + y_{2i1}\right) \end{bmatrix}.$$

This verifies Assumption 3(iii) since the conditional distribution of the individual specific effects depends on of the covariates from the two time periods only by means of their sums.

To verify Assumption 3(iii) for a higher number of time periods such as $T = 3$ and $\Delta = 2$, one can follow the same derivations as in $T = 2$ and $\Delta = 1$ and easily obtain the the mean of the conditional distribution of $\alpha_{1i}, \alpha_{2i} \,|X_{i1}, X_{i3}, Y_{2i0}, Y_{2i2}$ as

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & \frac{\beta_2^3-\beta_2^2-\beta_2+1}{\left(3+\beta_2+\beta_2^2-\beta_2^3\right)(1-\beta_2)} & \frac{\beta_2^3-\beta_2^2-\beta_2+1}{\left(3+\beta_2+\beta_2^2-\beta_2^3\right)(1-\beta_2)} \end{bmatrix} \begin{bmatrix} x_{i1} \\ x_{i3} \\ y_{i0} \\ y_{i2} \end{bmatrix} = \begin{bmatrix} x_{i1} + x_{i3} \\ \frac{\beta_2^3-\beta_2^2-\beta_2+1}{\left(3+\beta_2+\beta_2^2-\beta_2^3\right)(1-\beta_2)}\left(y_{i0} + y_{i2}\right) \end{bmatrix}.$$

The result relies on the fact that the estimation for given T and Δ requires observations of the dependent variable Y_{2iT} and its lags $Y_{2i(T-1)}$, $Y_{2i(T-\Delta)}$; hence, $Y_{1iT} = Y_{1i(T-1)} = Y_{1i(T-\Delta)} = 1$.