



Bias reduction in nonlinear and dynamic panels in the presence of cross-section dependence

Cavit Pakel

Department of Economics, Bilkent University, 06800 Ankara, Turkey

ARTICLE INFO

Article history:

Received 6 February 2016

Received in revised form 17 March 2019

Accepted 6 May 2019

Available online 6 August 2019

JEL classification:

C13

C23

Keywords:

Nonlinear dynamic panels

Incidental parameter bias

Integrated likelihood method

Profile likelihood method

Female labour force participation

ABSTRACT

Fixed effects estimation of nonlinear dynamic panel models is subject to the incidental parameter issue, leading to a biased asymptotic distribution. While this problem has been studied extensively in the literature, a general analysis allowing for both serial and cross-sectional dependence is missing. In this paper we investigate the large- N , T theory of the profile and integrated likelihood estimators, allowing for dependence across both dimensions. We show that under stronger dependence types the asymptotic bias disappears, but a $O_p(1/T)$ small-sample bias remains. We provide bias correction and inference methods, and also obtain primitive conditions for asymptotic normality under various dependence settings.

© 2019 Elsevier B.V. All rights reserved.

1. Introduction

Estimation of panel data models with both common and individual-specific parameters generally suffers from an asymptotic bias under large- N large- T asymptotics. To illustrate, let x_{it} ($i = 1, \dots, N$ and $t = 1, \dots, T$) be the data and $L_{iT}(\theta, \lambda_i) = L(\theta, \lambda_i; x_{i1}, \dots, x_{iT})$ be the marginal likelihood function, where λ_i is an individual-specific parameter and θ is the common parameter. Let $\ell_{iT}(\theta, \lambda_i) = \ln L_{iT}(\theta, \lambda_i)$ and θ_0 be the true value of θ . Assuming independence across strata, the profile likelihood estimator of θ_0 is given by

$$\hat{\theta} = \arg \max_{\theta} \sum_{i=1}^N \ell_{iT}(\theta, \hat{\lambda}_i(\theta)), \quad (1)$$

where $\hat{\lambda}_i(\theta) = \arg \max_{\lambda_i} \ell_{iT}(\theta, \lambda_i)$. Intuitively, the estimation error in each $\hat{\lambda}_i(\theta)$ vanishes as $T \rightarrow \infty$. However, at the same time the number of individual-specific parameters approaches infinity as $N \rightarrow \infty$, and the accumulated error grows. This latter effect dominates, leading to a wrongly centred likelihood function in (1), and bias in $\hat{\theta}$. This is an instance of the well-known incidental parameter issue (Neyman and Scott (1948); see Lancaster (2000) for a survey). This problem manifests itself in two ways: as a small- T bias and as a non-vanishing bias in the asymptotic distribution. To illustrate, generally

$$\mathbb{E}[\hat{\theta} - \theta_0] = \frac{A_{\theta_0}}{T} + o\left(\frac{1}{T}\right), \quad (2)$$

E-mail address: cavit.pakel@bilkent.edu.tr.

where the $O(T^{-1})$ term A_{θ_0}/T is known as the first-order bias. This small- T bias is not necessarily a problem in itself, as long as T is large enough. The main issue, however, is that under the standard \sqrt{NT} rate of convergence,

$$\sqrt{NT}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(\beta, \Omega), \quad (3)$$

where β is an asymptotic bias term and Ω is the asymptotic variance. As such the bias remains even as $T \rightarrow \infty$. The typical solution is to theoretically derive the bias of $\hat{\theta}$ (or of the likelihood function or of the score), and then correct this bias analytically or by the split-panel jackknife method; see, e.g., Hahn and Kuersteiner (2002, 2011), Woutersen (2002), Sartori (2003), Hahn and Newey (2004), Carro (2007), Bester and Hansen (2009), Fernández-Val (2009), Dhaene and Jochmans (2015), Arellano and Hahn (2016) and Fernández-Val and Weidner (2016). Arellano and Hahn (2007) provide a survey.

Arellano and Bonhomme (2009) take a different route and consider the integrated likelihood method, which is based on the maximisation of

$$\sum_{i=1}^N \ell_{iT}^l(\theta) = \sum_{i=1}^N \frac{1}{T} \ln \int_{\lambda_i} L_{iT}(\theta, \lambda_i) \pi_i(\lambda_i|\theta) d\lambda_i,$$

where $\pi_i(\lambda_i|\theta)$ is some weight function. This method encompasses some well-known estimation approaches, including fixed effects, random effects and Bayesian.¹ The issues in (2) and (3) beset this method, as well. As an alternative solution method, Arellano and Bonhomme (2009) propose using *robust* weight functions, designed to remove the bias of the score of $\ell_{iT}^l(\theta)$, and prove that this yields bias-corrected estimators.

Cross-section dependence is typically assumed away in the bias-correction literature, except for a few studies that focus on particular models rather than a generic likelihood estimation problem (e.g. Phillips and Sul (2007) and Bai (2009)). However, many interesting datasets can potentially be subject to some type of cross-section dependence. Bai (2009), for example, suggests different scenarios for macro, micro and financial data that would cause factor-type dependence. Spatial and network dependence are other possibilities. An important question is whether cross-section dependence leads to changes in (2) and (3), in which case the available correction methods will not be appropriate anymore. Furthermore, cross-section dependence also poses technical problems since the joint likelihood function will have a more complicated form than $\sum_{i=1}^N \ell_{iT}^l(\theta)$ or $\sum_{i=1}^N \ell_{iT}(\theta, \hat{\lambda}_i(\theta))$. In the case of multivariate normality, for example, there will be $O(N^2)$ additional parameters from the covariance matrix, posing both computational and statistical difficulties when N is large. Nevertheless, as long as one is interested in estimating θ and λ_i only, under certain conditions it would be sufficient to use $\sum_{i=1}^N \ell_{iT}^l(\theta)$ and $\sum_{i=1}^N \ell_{iT}(\theta, \hat{\lambda}_i(\theta))$ as approximations to the joint likelihood function.

This paper extends the bias-correction literature to dependence across both panel dimensions. Using ideas from the composite likelihood literature (Cox and Reid (2004)), we base our analysis on $\sum_{i=1}^N \ell_{iT}^l(\theta)$ and $\sum_{i=1}^N \ell_{iT}(\theta, \hat{\lambda}_i(\theta))$, which we call the *composite integrated likelihood* and *composite profile likelihood* functions, respectively. Under large- N large- T asymptotics, we analyse the bias of the resulting integrated likelihood and profile likelihood estimators, allowing for serial and cross-section dependence.

Our main results and contributions are as follows: we derive the first-order bias of the integrated and profile likelihood estimators, and find that the asymptotic distributions of both estimators become bias-free unless dependence is weak enough to yield \sqrt{NT} convergence. In other words, the incidental parameter bias ceases to be an asymptotic bias, and instead remains a $O_p(T^{-1})$ small-sample bias. As for this small-sample bias, its composition remains the same unless dependence is strong enough to yield \sqrt{T} convergence. In that case, the first-order bias contains two extra terms which arise due to the reduced convergence rate. The key factor that determines the rate of convergence is the strength of cross-section dependence. We therefore study in substantial detail three general dependence settings that cover a variety of possibilities. We also propose bias-corrected estimators using the split-panel jackknife method, and prove that the resulting estimators are asymptotically normal and correctly centred. The paper also develops a valid inference method. Importantly, neither bias correction nor inference requires knowledge of the underlying dependence structure. This is an important advantage as this information is rarely available. Finally, we consider the empirical application of female labour force participation decision, using PSID data.

The bias correction literature almost exclusively focuses on microeconomic applications. However, the considered framework is general enough to accommodate other interesting problems. One example is GARCH-type volatility modelling in short financial panels, which we investigate in the Supplementary Appendix. Typically, GARCH models require several hundreds of observations, due to the persistent nature of financial data and complex nonlinearities of volatility models. This rules out volatility modelling on inherently short data, such as monthly hedge fund returns. We propose a bias-corrected version of the GARCH model for panel data, and show that this approach works well with as few as 150 observations. This is a substantial improvement, potentially making GARCH tools available to short financial series for the first time.

The rest is organised as follows: Section 2 introduces the notation and relevant concepts. The assumptions are listed and discussed in Section 3. Consistency and asymptotic expansions are obtained in Section 4. Section 5 considers the analysis of various cross-section dependence types. Sections 6 and 7 deal with bias correction and inference. Section 8 provides the simulation analysis, followed by the empirical illustration in Section 9. Section 10 concludes. The main proofs are in the Mathematical Appendix. The Supplementary Appendix contains all remaining material and the GARCH application.

¹ Severini (1999, 2007) also shows that the modified profile likelihood (Barndorff-Nielsen (1983)) and adjusted profile likelihood (Cox and Reid (1987)) methods have integrated likelihood representations.

2. Main concepts and notation

Let x_{it} be the data, indexed by individuals and time ($i = 1, \dots, N$ and $t = 1, \dots, T$, respectively), and $\ell(\theta, \lambda_i; x_{it})$ be the correctly specified marginal log-likelihood function. Let θ be a P -dimensional vector-valued parameter and λ_i be a scalar individual-specific parameter. The corresponding true parameter values are given by θ_0 and λ_{i0} . The variable x_{it} can be defined flexibly. For example, [Arellano and Hahn \(2016\)](#) consider a conditional likelihood setting where $\ell(\theta, \lambda_i; x_{it}) = \ell(\theta, \lambda_i; y_{it} | y_{i,t-1}, \dots, y_{i,t-q})$, implying $x_{it} = (y_{it}, y_{i,t-1}, \dots, y_{i,t-q})$. Under cross-section independence and weak serial dependence, the estimator

$$(\hat{\theta}, \hat{\lambda}_1, \dots, \hat{\lambda}_N) = \arg \max_{\theta, \lambda_1, \dots, \lambda_N} \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \ell(\theta, \lambda_i; x_{it}) \quad (4)$$

would be fixed- N large- T consistent for $(\theta_0, \lambda_{10}, \dots, \lambda_{N0})$ as in [Hahn and Kuersteiner \(2011\)](#) (referred to as HK11 henceforth).

Under cross-section dependence, neither the objective function in (4) nor $\sum_{i=1}^N \ell_{it}^l(\theta)$ corresponds to the joint likelihood function anymore, as they neglect dependence. The obvious option is to use the full joint likelihood function. However, estimation of such models for panels with large cross-sections is likely to suffer from statistical and computational issues. Instead, we use the composite likelihood approach, which is based on approximating the joint density by using combinations of lower dimensional marginal densities ([Lindsay \(1988\)](#) and [Cox and Reid \(2004\)](#)). Under cross-section dependence, this approximation will obviously lead to a misspecified joint likelihood function. This is the only type of misspecification allowed in this study.

The composite likelihood function can be constructed using combinations of, e.g., bivariate or trivariate marginal likelihoods. This approach would be suitable for applications which are concerned with modelling dependence, as this preserves some of the dependence structure. While we are interested in the behaviour of panel estimators under dependence, we are not interested in estimating the dependence structure itself. For that reason, we consider the most basic composite likelihood function which is the equally weighted average of all univariate marginal likelihoods. This yields the objective function in (4) or, for integrated likelihood estimation, $\sum_{i=1}^N \ell_{it}^l(\theta)$.²

We next introduce the different likelihood functions used in the theoretical analysis. In what follows $\ell_{it}(\theta, \lambda_i) = \ell(\theta, \lambda_i; x_{it})$, $\ell_{iT}(\theta, \lambda_i) = T^{-1} \sum_{t=1}^T \ell_{it}(\theta, \lambda_i)$ and $\ell_{NT}(\theta, \lambda_1, \dots, \lambda_N) = N^{-1} \sum_{i=1}^N \ell_{iT}(\theta, \lambda_i)$. The familiar *profile likelihood* function is given by

$$\ell_{iT}(\theta, \hat{\lambda}_i(\theta)) \quad \text{where} \quad \hat{\lambda}_i(\theta) = \arg \max_{\lambda_i} \ell_{iT}(\theta, \lambda_i).$$

Next, the benchmark least favourable *target likelihood* function is given by

$$\ell_{iT}(\theta, \bar{\lambda}_{iT}(\theta)), \quad \text{where} \quad \bar{\lambda}_{iT}(\theta) = \arg \max_{\lambda_i} \frac{1}{T} \sum_{t=1}^T \mathbb{E}[\ell_{it}(\theta, \lambda_i)].$$

Note that here and elsewhere all moments are with respect to the true underlying density evaluated at $(\theta_0, \lambda_{10}, \dots, \lambda_{N0})$. The target likelihood is maximised at θ_0 and the expected information at θ_0 is the same as the partial expected information (hence the ‘least favourable’ property). The curve $(\theta, \bar{\lambda}_{iT}(\theta))$ is such that, by definition, $\bar{\lambda}_{iT}(\theta_0) = \lambda_{i0}$. Crucially, the target likelihood function guarantees orthogonality of the score for θ , to the plane spanned by the scores for $(\lambda_1, \dots, \lambda_N)$, which leads to unbiased estimation of θ_0 (see [Severini and Wong \(1992\)](#), [Severini \(2000, Section 4.6\)](#) and [Pace and Salvan \(2006\)](#)). Notice that this is an infeasible benchmark as it is based on $\mathbb{E}[\ell_{it}(\theta, \lambda_i)]$.

Finally, the *integrated likelihood* function is given by

$$\ell_{iT}^l(\theta) = \frac{1}{T} \ln \int_{\lambda_i} \exp[T \ell_{iT}(\theta, \lambda_i)] \pi_i(\lambda_i | \theta) d\lambda_i. \quad (5)$$

The integrated likelihood estimator offers the possibility to obtain a first-order unbiased estimator by choosing the weight/prior function, $\pi_i(\lambda_i | \theta)$, carefully. While this prior function can be chosen based on subjective judgement, our approach to this problem is frequentist. In particular, we follow [Arellano and Bonhomme \(2009\)](#) (referred to as AB09 henceforth), who derive the analytical specification of such a bias-correcting *robust* prior under time-series and cross-section independence.

² This approach is obviously subject to efficiency loss and can in some cases lead to lower convergence rates – these issues are beyond our scope and not investigated here. A subtle technical advantage of this method is worth noting: the infinite dimensional parameter $(\lambda_1, \dots, \lambda_N)$ is usually identified via a suitable finite-dimensional model, e.g. $\hat{\lambda}_i(\theta) = \arg \max_{\lambda_i} T^{-1} \sum_{t=1}^T \ell(\theta, \lambda_i; x_{it})$. This makes it possible to obtain individual expansions for each $\hat{\lambda}_i$, a key step in the asymptotic analysis of $\hat{\theta} - \theta_0$. If, however, the joint likelihood function contains incidental parameters which cannot be identified via finite-dimensional models, this standard approach fails. The composite likelihood method is immune to such possibilities, as it neglects cross-section dependence.

The composite integrated likelihood and composite profile likelihood estimators are then given by

$$\hat{\theta}_{IL} = \arg \max_{\theta} \sum_{i=1}^N \ell_{iT}^l(\theta) \quad \text{and} \quad \hat{\theta} = \arg \max_{\theta} \sum_{i=1}^N \ell_{iT}(\theta, \hat{\lambda}_i(\theta)),$$

respectively. To avoid cumbersome definitions, henceforth we will refer to these as the integrated and profile likelihood estimators.

In the rest of the paper we use the following notation: When a function is evaluated at $(\theta, \bar{\lambda}_{iT}(\theta))$, the argument will usually omit $\bar{\lambda}_{iT}(\theta)$; e.g., $\ell_{iT}(\theta) = \ell_{iT}(\theta, \bar{\lambda}_{iT}(\theta))$, $\ell_{NT}(\theta) = \ell_{NT}(\theta, \bar{\lambda}_{1T}(\theta), \dots, \bar{\lambda}_{NT}(\theta))$ etc. For partial derivatives with respect to λ , we use the following convention:

$$\ell_{iT}^{\lambda}(\theta, \lambda) = \frac{\partial \ell_{iT}(\theta, \lambda)}{\partial \lambda}, \quad \ell_{iT}^{\lambda\lambda}(\theta, \bar{\lambda}_{iT}(\theta)) = \left. \frac{\partial^2 \ell_{iT}(\theta, \lambda)}{\partial \lambda^2} \right|_{\lambda=\bar{\lambda}_{iT}(\theta)} \quad \text{etc.}$$

Total derivatives with respect to θ are denoted by d_{θ} , $d_{\theta\theta}$, $d_{\theta\theta\theta}$, ... For example,

$$d_{\theta} \ell_{iT}^{\lambda\lambda}(\theta_0) = \left\{ \frac{d}{d\theta} \left[\left. \frac{\partial^2 \ell_{iT}(\theta, \lambda)}{\partial \lambda^2} \right|_{\lambda=\bar{\lambda}_{iT}(\theta)} \right] \right\} \Big|_{\theta=\theta_0}$$

and similarly for other derivatives. We also define $V_{iT}^{\lambda\lambda}(\theta, \lambda_i) = \ell_{iT}^{\lambda\lambda}(\theta, \lambda_i) - \mathbb{E}[\ell_{iT}^{\lambda\lambda}(\theta, \lambda_i)]$. Furthermore, $S = d_{\theta} \ell_{NT}(\theta_0)$, $H = d_{\theta\theta} \ell_{NT}(\theta_0)$ and $v = \mathbb{E}[H]$. S is also referred to as the projected score, given by $S = \ell_{NT}^{\theta}(\theta_0) - N^{-1} \sum_{i=1}^N \{\mathbb{E}[\ell_{iT}^{\lambda\lambda}(\theta_0)]\}^{-1} \mathbb{E}[\ell_{iT}^{\lambda\theta}(\theta_0)] \ell_{iT}^{\lambda}(\theta_0)$, where the superscript θ denotes partial differentiation with respect to θ . Finally, letting $p \in \{1, \dots, P\}$,

$$Z_p = \mathbb{E} \left\{ \left[d_{\theta\theta} \frac{d\ell_{NT}(\theta)}{d\theta_p} \right] \Big|_{\theta=\theta_0} \right\}, \quad \text{and} \quad M = \begin{bmatrix} S'v^{-1}Z_1v^{-1}S \\ \vdots \\ S'v^{-1}Z_Pv^{-1}S \end{bmatrix}.$$

3. Assumptions

Assumption 3.1. (i) $N, T \rightarrow \infty$ jointly and, for $0 < c < \infty$, $\lim_{N,T \rightarrow \infty} N/T = c$; (ii) the parameter space for (θ, λ_i) is given by $\Psi = \Theta \times \Lambda$ for all i where Θ and Λ are compact subsets of \mathbb{R}^P and \mathbb{R} , respectively; (iii) $\ell_{iT}(\theta, \lambda_i)$ has a unique maximum at $\hat{\lambda}_i(\theta)$; (iv) $\inf_{i,T} \inf_{(\theta, \lambda) \in \Psi} |\ell_{iT}^{\lambda\lambda}(\theta, \lambda)| > 0$ as $N, T \rightarrow \infty$, $\mathbb{E}[-d_{\theta\theta} \ell_{iT}(\theta_0)]$ is uniformly positive definite as $N, T \rightarrow \infty$ and $\lim_{N,T \rightarrow \infty} \mathbb{E}[-d_{\theta\theta} \ell_{NT}(\theta_0)]$ exists and is positive definite.

Assumption 3.2. The following identification conditions hold: (i) For any $\eta > 0$, $\varepsilon = \inf_{i,T} \inf_{\theta} \{\mathbb{E}[\ell_{iT}(\theta, \bar{\lambda}_{iT}(\theta))]\} - \sup_{\{\lambda: ||\lambda - \bar{\lambda}_{iT}(\theta)|| > \eta\}} \mathbb{E}[\ell_{iT}(\theta, \lambda)] > 0$, and (ii) for any $\tilde{\eta} > 0$, $\tilde{\varepsilon} = \inf_{i,T} \{\mathbb{E}[\ell_{iT}(\theta_0, \bar{\lambda}_{iT}(\theta_0))]\} - \sup_{\{\theta: ||\theta - \theta_0|| > \tilde{\eta}\}} \mathbb{E}[\ell_{iT}(\theta, \bar{\lambda}_{iT}(\theta))] > 0$.

Assumption 3.3. Let $k = (k_0, k_1, \dots, k_P)$ be a vector of $P + 1$ non-negative integers and let $|k| = \sum_{p=1}^P k_p$. Let, also, $\mathcal{Z}_{it}^k(\theta, \lambda) = \partial^{|k|} \ell_{it}(\theta, \lambda) / (\partial \lambda^{k_0} \partial \theta_1^{k_1} \partial \theta_2^{k_2} \dots \partial \theta_P^{k_P})$. Then, for any k such that $|k| \leq 7$, $|\mathcal{Z}_{it}^k(\theta', \lambda') - \mathcal{Z}_{it}^k(\theta'', \lambda'')| \leq d(x_{it}) ||(\theta', \lambda') - (\theta'', \lambda'')||$ for all $\theta', \theta'' \in \Theta$ and $\lambda', \lambda'' \in \Lambda$, where $d(\cdot)$ is a measurable function of x_{it} and $d(x_{it})$ is such that $\sup_{(\theta, \lambda) \in \Psi} ||\mathcal{Z}_{it}^k(\theta, \lambda)|| \leq d(x_{it})$ and $\sup_{1 \leq i \leq N} \sup_{1 \leq t \leq T} \mathbb{E}[|d(x_{it})|^{(10+10q)/(1-10v)+\epsilon}] < \infty$ for some $q \geq (P+1)/2 + 1$, $\epsilon > 0$ and $0 < v < 1/10$, as $N, T \rightarrow \infty$.

Assumption 3.4. Define the σ -fields, $\mathcal{G}_{i,-\infty}^t = \sigma(x_{it}, x_{i,t-1}, \dots)$ and $\mathcal{H}_{i,t}^{\infty} = \sigma(x_{it}, x_{i,t+1}, \dots)$, and let $\alpha_i(m) = \sup_t \sup_{G \in \mathcal{G}_{i,-\infty}^t, H \in \mathcal{H}_{i,t+m}^{\infty}} |P(G \cap H) - P(G)P(H)|$. Then, for all $i = 1, \dots, N$, $\{x_{it}\}_{t=1}^T$ is an α -mixing sequence in the sense that $\lim_{m \rightarrow \infty} \alpha_i(m) = 0$ and the mixing coefficients are such that $\sup_i \sum_{m=1}^{\infty} \alpha_i(m)^{1/r_1-1/r_2} < \infty$ where $r_1 = (10 + 10q)/(1 - 10v) + \epsilon/2$, $r_2 = r_1 + \epsilon/2$, and q and v are defined as before.

Assumption 3.5. Define $\Pi_i^k(\theta, \lambda) = \partial^{|k|} \pi_i(\lambda|\theta) / (\partial \lambda^{k_0} \partial \theta_1^{k_1} \partial \theta_2^{k_2} \dots \partial \theta_P^{k_P})$ and let k be as in Assumption 3.3. Then, $\sup_i \sup_{(\theta, \lambda) \in \Psi} \Pi_i^k(\theta, \lambda) < \infty$ for any k such that $|k| \leq 5$ as $N, T \rightarrow \infty$. In addition, $\inf_i \inf_{(\theta, \lambda) \in \Psi} \pi_i(\lambda|\theta) > 0$.

Assumption 3.6. $N^{-1} \sum_{i=1}^N [\ell_{iT}^{\lambda}(\theta_0)]^2$ and $N^{-1} \sum_{i=1}^N \ell_{iT}^{\lambda}(\theta_0)[d_{\theta} \ell_{iT}^{\lambda}(\theta_0)]$ are $O_p(T^{-1})$ as $N, T \rightarrow \infty$.

Assumption 3.7. There is some $0 \leq \rho_c \leq 1$ for $c \in \{1, 2, 3\}$, such that as $N, T \rightarrow \infty$, (i) $\sqrt{N^{\rho_1} T} d_{\theta} \ell_{NT}(\theta_0) = O_p(1)$; (ii) $\sqrt{N^{\rho_2} T} \{d_{\theta\theta} \ell_{NT}(\theta_0) - \mathbb{E}[d_{\theta\theta} \ell_{NT}(\theta_0)]\} = O_p(1)$; and (iii) $\sqrt{N^{\rho_3} T} \{d_{\theta\theta\theta} \ell_{NT}(\theta_0) - \mathbb{E}[d_{\theta\theta\theta} \ell_{NT}(\theta_0)]\} = O_p(1)$.

Assumption 3.8. Let $\tilde{\theta}$ be such that $\tilde{\theta} \xrightarrow{P} \theta_0$. The integrated likelihood function (5) is such that $N^{-1} \sum_{i=1}^N d_{\theta\theta\theta} \ell_{iT}^l(\theta) \Big|_{\theta=\tilde{\theta}} = O(1)$ as $N, T \rightarrow \infty$.

A substantial portion of the above assumptions is similar to those of HK11, who discuss the validity of their assumptions for the probit, logit and tobit models. The main differences arise due to the presence of cross-section dependence and our focus on the integrated likelihood estimator. Assumptions 3.5 and 3.8, for instance, would not be required for the analysis of the profile likelihood estimator.

Assumption 3.1 formalises the large- N large- T asymptotics and contains other standard conditions. The uniqueness of $\hat{\lambda}_i(\theta)$ is required for the existence of the Laplace approximation to the integrated likelihood function.³ The concavity conditions of Assumption 3.1(iv) are similar to e.g. Assumption 4.1(v) of Fernández-Val and Weidner (2016) or Conditions 5–7 of HK11. Assumption 3.2 yields identification of θ_0 and $\bar{\lambda}_{iT}(\theta)$. This ensures that we obtain $|\hat{\lambda}_i(\theta) - \bar{\lambda}_{iT}(\theta)| \xrightarrow{p} 0$, uniformly over both i and θ . This stronger convergence result is required in proving the consistency of $\hat{\theta}_{IL}$. Assumption 3.3 is almost identical to Condition 4 of HK11, except that we require the same assumptions for higher order derivatives, due to our focus on the integrated likelihood function. Assumption 3.4 establishes the mixing characteristics of data and is slightly weaker than HK11's Condition 3(iii). Smoothness assumptions on the prior/weight function are contained in Assumption 3.5, which are required for the existence of asymptotic expansions.

The two terms in Assumption 3.6 determine the rate of the incidental parameter bias. They also appear in HK11's first-order bias term. Under cross-section independence these terms are $O_p(T^{-1})$ without further assumptions. Once cross-section dependence is allowed, obtaining the same rate requires more structure on the nature of dependence. Assumption 3.7 is a similarly high-level assumption, where the particular values of ρ_1 , ρ_2 and ρ_3 are closely related to the nature of cross-section dependence. Usually, $\rho_1 = \rho_2 = \rho_3 = 1$ in the case of cross-section independence. On the other hand, $\rho_c = 0$ for all c implies that dependence is so strong that cross-section variation does not make any information contribution. In Section 5 we will provide primitive conditions for these two assumptions, under various dependence settings.

Assumption 3.8 is a high-level smoothness condition, which ensures that the remainder in the third-order expansion of $N^{-1} \sum_{i=1}^N [d_\theta \ell_{iT}^l(\theta)]|_{\theta=\hat{\theta}_{IL}}$ about θ_0 is well-behaved. One possibility for proving this condition via earlier assumptions is to use the fourth-order derivative of the Laplace expansion for $\ell_{iT}^l(\theta)$, in order to analyse the behaviour of $d_{\theta\theta\theta\theta} \ell_{iT}^l(\theta)$ in terms of the familiar likelihood derivatives. For this approach to work, the fourth-order derivative of the remainder of the Laplace expansion should be $o(1)$. The difficulty with this approach is that the Laplace expansion itself consists of multiple linear expansions; as such, the analysis of the remainder and its derivatives becomes complicated very quickly. A better option would be to verify this assumption on a case-by-case basis, by investigating $d_{\theta\theta\theta\theta} \ell_{iT}^l(\theta)$ directly for the model of interest.

4. Consistency and asymptotic expansions

4.1. Integrated likelihood

Theorem 4.1. Under Assumptions 3.1–3.5, $\max_{1 \leq i \leq N} \sup_{\theta \in \Theta} |\hat{\lambda}_i(\theta) - \bar{\lambda}_{iT}(\theta)| \xrightarrow{p} 0$ and $\|\hat{\theta}_{IL} - \theta_0\| \xrightarrow{p} 0$ as $N, T \rightarrow \infty$.

The proof of the first result follows standard steps. For the second result, we utilise a Laplace expansion to prove the asymptotic equivalence of the integrated and profile likelihood functions. This and a uniform convergence result for the profile likelihood function allow us to obtain a uniform convergence result for the integrated likelihood function. This uniform convergence result, coupled with the identification condition of Assumption 3.2(ii), yields consistency of $\hat{\theta}_{IL}$.

In obtaining the asymptotic expansion for $\hat{\theta}_{IL} - \theta_0$ we follow the same steps as AB09: First, we obtain an asymptotic expansion for the difference between the integrated and target likelihoods (Theorem 4.2). This captures the $O_p(T^{-1})$ incidental parameter bias of the integrated likelihood function and reveals the specification of the bias-correcting prior. Building up on this we then obtain an asymptotic expansion for $\hat{\theta}_{IL} - \theta_0$ (Theorem 4.3).

Theorem 4.2. Let Assumptions 3.1, 3.2(i), 3.3–3.5 hold. Then,

$$\ell_{iT}^l(\theta) - \ell_{iT}(\theta) = \frac{1}{2T} \ln \left(\frac{2\pi}{T} \right) + \mathcal{B}_{iT}(\theta) + o_p \left(\frac{1}{T^{6/5}} \right), \quad (6)$$

uniformly over i and θ as $N, T \rightarrow \infty$, where

$$\mathcal{B}_{iT}(\theta) = \frac{1}{2} \frac{(\ell_{iT}^{\lambda}(\theta))^2}{\mathbb{E}[-\ell_{iT}^{\lambda\lambda}(\theta)]} - \frac{1}{2T} \ln \{\mathbb{E}[-\ell_{iT}^{\lambda\lambda}(\theta)]\} + \frac{1}{T} \ln \pi_i(\bar{\lambda}_{iT}(\theta)|\theta).$$

Notice that Theorem 4.2 implies that

$$\begin{aligned} \mathbb{E}[\ell_{iT}^l(\theta) - \ell_{iT}(\theta)] &= \frac{1}{2T} \ln \left(\frac{2\pi}{T} \right) + \frac{1}{2} \frac{\mathbb{E}[(\ell_{iT}^{\lambda}(\theta))^2]}{\mathbb{E}[-\ell_{iT}^{\lambda\lambda}(\theta)]} - \frac{1}{2T} \ln \{\mathbb{E}[-\ell_{iT}^{\lambda\lambda}(\theta)]\} \\ &\quad + \frac{1}{T} \ln \pi_i(\bar{\lambda}_{iT}(\theta)|\theta) + o \left(\frac{1}{T^{6/5}} \right). \end{aligned} \quad (7)$$

³ Multiple maxima can be dealt with by dividing Λ into intervals, with one maximum in each interval.

This is the same result as in Theorem 1 of AB09, up to a negligible remainder. As such, the prior specification which yields a first-order unbiased score is also identical to theirs (see Section 6). Unlike in AB09, Theorem 4.2 is obtained under serial dependence. We also formally prove that the remainder term's rate is valid uniformly over i as $N, T \rightarrow \infty$, which is crucial for investigating the asymptotic behaviour of $\sum_{i=1}^N \ell_{iT}^l(\theta)$.

Theorem 4.2 is based on the Laplace expansion, which we formally prove for our particular case in Lemma S1.20 in the Supplementary Appendix, closely following the approach of Kass et al. (1990). We also obtain the asymptotic rates of the remainder of this expansion and its first three derivatives with respect to θ , as these are used in proving later results. To the best of our knowledge these technical results on the remainder term are novel, and could be of use to studies utilising the Laplace expansion.

Theorem 4.3. Under Assumptions 3.1–3.8,

$$(\hat{\theta}_{IL} - \theta_0) = A_{NT} + B_{NT} + C_{NT} + D_{NT} + o_p\left(\frac{1}{T^{6/5}}\right), \quad (8)$$

as $N, T \rightarrow \infty$, where

$$\begin{aligned} A_{NT} &= -v^{-1}S = O_p\left(\frac{1}{\sqrt{N^{\rho_1}T}}\right), \\ B_{NT} &= -v^{-1}\frac{1}{N}\sum_{i=1}^N d_\theta \left\{ \frac{\ln \pi_i(\tilde{\lambda}_{iT}(\theta)|\theta)}{T} - \frac{\ln \mathbb{E}[-\ell_{iT}^{\lambda\lambda}(\theta)]}{2T} + \frac{[\ell_{iT}^\lambda(\theta)]^2}{2\mathbb{E}[-\ell_{iT}^{\lambda\lambda}(\theta)]} \right\} \bigg|_{\theta=\theta_0} = O_p\left(\frac{1}{T}\right), \\ C_{NT} &= v^{-1}(H - v)v^{-1}S = O_p\left(\frac{1}{N^{(\rho_1+\rho_2)/2}T}\right), \\ D_{NT} &= -\frac{1}{2}v^{-1}M = O_p\left(\frac{1}{N^{\rho_1}T}\right). \end{aligned}$$

4.2. Profile likelihood

As mentioned previously, some of the Assumptions 3.1–3.8 are too strong or unnecessary for the analysis of $\hat{\theta}$. In particular, Assumptions 3.5 and 3.8 are not needed. Moreover, we will also use the following modified versions of Assumptions 3.1–3.3.

Assumption 4.1. Same as Assumption 3.1(i), (ii) and (iv).

Assumption 4.2. $\varepsilon = \inf_{i,T} \{\mathbb{E}[\ell_{iT}(\theta_0, \lambda_{i0})] - \sup_{\{(\theta, \lambda): \|(\theta, \lambda) - (\theta_0, \lambda_{i0})\| > \eta\}} \mathbb{E}[\ell_{iT}(\theta, \lambda)]\} > 0$ for any $\eta > 0$.

Assumption 4.3. Same as Assumption 3.3, except that $|k| \leq 5$.

As the Laplace expansion is only required for the integrated likelihood estimator, Assumption 4.1 retains Assumption 3.1 except for part (iii). Assumption 4.2 corresponds to Condition 1 of HK11, which is not surprising since they also consider the profile likelihood estimator. Finally, the smoothness condition on the likelihood function is now imposed up to the fifth-order derivative as in HK11.

Theorem 4.4. Under Assumptions 4.1–4.3 and 3.4, $\max_{1 \leq i \leq N} |\hat{\lambda}_i(\hat{\theta}) - \lambda_{i0}| \xrightarrow{P} 0$ and $\|\hat{\theta} - \theta_0\| \xrightarrow{P} 0$ as $N, T \rightarrow \infty$.

Theorem 4.5. Under Assumptions 4.1–4.3, 3.4, 3.6 and 3.7

$$(\hat{\theta} - \theta_0) = A_{NT} + B'_{NT} + C_{NT} + D_{NT} + o_p\left(\frac{1}{T^{6/5}}\right), \quad (9)$$

as $N, T \rightarrow \infty$ where A_{NT} , C_{NT} and D_{NT} are the same as in Theorem 4.3, and

$$B'_{NT} = -v^{-1}\frac{1}{N}\sum_{i=1}^N d_\theta \left\{ \frac{[\ell_{iT}^\lambda(\theta)]^2}{2\mathbb{E}[-\ell_{iT}^{\lambda\lambda}(\theta)]} \right\} \bigg|_{\theta=\theta_0} = O_p\left(\frac{1}{T}\right).$$

The consistency result of Theorem 4.4 is standard and follows from the same arguments as in Theorems 3 and 4 of HK11. Theorem 4.5 reveals that the asymptotic expansions for $\hat{\theta}$ and $\hat{\theta}_{IL}$ are nearly identical, except for B_{NT} and B'_{NT} . The difference arises from the presence of the prior function in the integrated likelihood.

4.3. Remarks

Theorems 4.3 and **4.5** characterise the first-order bias of the integrated and profile likelihood estimators. B_{NT} and B'_{NT} are the standard incidental parameter bias terms. C_{NT} and D_{NT} , on the other hand, are novel. These two terms capture the effect of cross-section dependence on $\hat{\theta}_{IL}$ and $\hat{\theta}$. Under cross-section independence where $\rho_1 = \rho_2 = 1$, these terms would be subsumed in the remainder term. However, under dependence they can have an order as large as $O_p(T^{-1})$.⁴

Assuming that a central limit theorem (CLT) applies to A_{NT} , the rate of convergence of the two estimators will be equal to $\sqrt{N\rho_1 T}$ by **Assumption 3.7**. This implies that the incidental parameter bias ceases to be an asymptotic bias, unless dependence is weak enough to yield $\rho_1 = 1$. This is because, under any scenario where $\rho_1 < 1$ the terms $\sqrt{N\rho_1 T}(B_{NT} + C_{NT} + D_{NT})$ and $\sqrt{N\rho_1 T}(B'_{NT} + C_{NT} + D_{NT})$ both become $o_p(1)$ – consequently, the asymptotic distributions will be centred at zero. When dependence is weak enough to yield $\rho_1 = 1$, on the other hand, we return back to the same result as under independence, where $p \lim_{N,T \rightarrow \infty} \sqrt{NT}B_{NT}$ and $p \lim_{N,T \rightarrow \infty} \sqrt{NT}B'_{NT}$ become the $O(1)$ non-vanishing bias terms in the asymptotic distribution. The main message of these observations is that if one is only worried about bias in the asymptotic distribution of $\hat{\theta}_{IL}$ or $\hat{\theta}$, one need not take any action unless dependence is absent or weak enough to yield \sqrt{NT} convergence.⁵

Concerning the small-sample bias of $\hat{\theta}_{IL}$ and $\hat{\theta}$ (given by the leading $O_p(T^{-1})$ portions of $\hat{\theta}_{IL} - \theta_0$ and $\hat{\theta} - \theta_0$) **Theorems 4.3** and **4.5** imply that as long as $\rho_1 > 0$, these bias terms will consist only of B_{NT} and B'_{NT} . In the extreme case of strong dependence with $\rho_1 = \rho_2 = 0$, on the other hand, the $O_p(T^{-1})$ bias will be given by $B_{NT} + C_{NT} + D_{NT}$ and $B'_{NT} + C_{NT} + D_{NT}$. Therefore, the presence of dependence can lead to a change in the composition of the $O_p(T^{-1})$ small-sample bias. Standard bias correction approaches in the literature are designed to deal with the terms B_{NT} and B'_{NT} (e.g. AB09 and HK11), but not with C_{NT} and D_{NT} . As such, when correcting the $O_p(T^{-1})$ small-sample bias is of interest and strong dependence is a likely scenario, novel correction methods will be required that can deal with all bias terms. This task will be considered in Section 6.

These remarks are of course tentative as they rely on the existence of some CLT. Moreover, the asymptotic expansions are based on the high-level **Assumptions 3.6** and **3.7**. In Section 5 we will consider three relevant dependence scenarios and obtain **Assumptions 3.6** and **3.7**, as well as appropriate CLTs, under primitive conditions.

5. Particular types of cross-section dependence

In this part we focus on three dependence settings that cover a wide variety of scenarios. For each case we provide (i) a theorem which provides primitive conditions for **Assumptions 3.6–3.7**, and determines the rates of B_{NT} , B'_{NT} , C_{NT} , D_{NT} and of their expectations; (ii) a theorem that obtains an appropriate CLT; and (iii) a corollary that provides the asymptotic distribution of the integrated and profile likelihood estimators. In what follows, we let $\bar{v}_\infty = \lim_{N,T \rightarrow \infty} \mathbb{E}[d_{\theta\theta} \ell_{NT}(\theta_0, \bar{\lambda}_{1T}(\theta_0), \dots, \bar{\lambda}_{NT}(\theta_0))]$ and

$$S_{it,T} = \ell_{it}^\theta(\theta_0, \bar{\lambda}_{iT}(\theta_0)) - \frac{\mathbb{E}[\ell_{iT}^{\lambda\theta}(\theta_0, \bar{\lambda}_{iT}(\theta_0))]}{\mathbb{E}[\ell_{iT}^{\lambda\lambda}(\theta_0, \bar{\lambda}_{iT}(\theta_0))]} \ell_{it}^\lambda(\theta_0, \bar{\lambda}_{iT}(\theta_0)).$$

Then, $d_\theta \ell_{NT}(\theta_0) = (NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T S_{it,T}$. Note that although $S_{it,T}$ depends on T , this is due to expectation terms. Hence, T does not affect the random behaviour of $S_{it,T}$.

5.1. Weak dependence

We first consider weak dependence in the form of mixing random fields (Conley, 1999; Kelejian and Prucha, 2007; Bester et al., 2011). In this setting, although any two observations are assumed to be dependent, the magnitude of dependence approaches zero as the distance between them increases. This type of dependence is relevant for, in particular, networks. For example, Goldsmith-Pinkham and Imbens (2013) mention that if individuals exhibit homophily in some covariates, then the network can be characterised by a weak dependence setting based on mixing assumptions, very much like the case considered here. See also Chandrasekhar and Lewis (2016).

Our spatial setting is similar to Jenish and Prucha (2009) and Bester et al. (2011). Formally, the location indices $j = (i, t)$ are assumed to be located on an integer lattice $D \subseteq \mathbb{Z}^2$. The distance between two locations j' and j'' is measured by $d(j', j'') = \max_{l \in \{1, 2\}} |j'_l - j''_l|$, where j'_l is the l th component of j' (and likewise for j''_l). The distance between subsets of D is given by $d(D', D'') = \inf\{d(j', j'') : j' \in D' \text{ and } j'' \in D''\}$, for any $D', D'' \subseteq D$. Let $|A|$ be the cardinality of a finite set A . We use the following definition for the α -mixing coefficient for random fields.

⁴ An earlier study of the bias of nonlinear estimators in a non-panel setting is by Rilstone et al. (1996). The result in (8) is similar to their Lemma 3.1. In particular, letting θ be scalar and noting that in their case S is given by $T^{-1} \sum_{t=1}^T (d\ell_t(\theta)/d\theta)|_{\theta=\theta_0}$, we observe that C_{NT} and D_{NT} correspond (but are not exactly equivalent) to their $-QVa_{-1/2}$ and $-(1/2)QH_2[a_{-1/2} \otimes a_{-1/2}]$. They do not have a counterpart for B_{NT} , as an incidental parameter is not present. Note that their notation differs from ours.

⁵ It must be underlined that the absence of asymptotic bias is not a consequence of neglecting cross-section dependence in estimation. Indeed, the profile and integrated likelihood estimators are still asymptotically biased under weak dependence (see Corollary 5.1).

Definition 5.1. Let D_{NT} be a finite subset of D . For $D' \subseteq D_{NT}$ and $D'' \subseteq D_{NT}$, define the respective σ -algebras generated by the random field Y_j as $\mathcal{A}_{NT} = \sigma(Y_j : j \in D')$ and $\mathcal{B}_{NT} = \sigma(Y_j : j \in D'')$. Then, the α -mixing coefficient for random fields is given by

$$\tilde{\alpha}_{k,l,N,T}(m) = \sup_{\mathbb{S}_{NT}} |P(A \cap B) - P(A)P(B)|,$$

where $\mathbb{S}_{NT} = \{(A, B) : A \in \mathcal{A}_{NT}, B \in \mathcal{B}_{NT}, |D'| \leq k, |D''| \leq l, d(D', D'') \geq m\}$.

This definition differs from the standard time-series version in that it depends on the cardinalities of D' and D'' , as well. Intuitively, given a fixed distance between two sets, the dependence between larger sets will be at least as high as the dependence between smaller sets, due to accumulation of dependence. Hence, $\tilde{\alpha}_{k,l,N,T}(m)$ is non-decreasing in k and l . Note that since the location set D_{NT} depends on N and T , it makes sense to define the mixing coefficient dependent on these as well. We will not worry about this in the remainder since the focus will be on $\tilde{\alpha}_{k,l}(m) = \sup_{N,T} \tilde{\alpha}_{k,l,N,T}(m)$. Then, the underlying random field is α -mixing if $\lim_{m \rightarrow \infty} \tilde{\alpha}_{k,l}(m) = 0$.

The behaviour of the individual terms in [Theorems 4.3](#) and [4.5](#) is investigated in [Theorem 5.1](#), which also obtains the conditions of [Assumptions 3.6](#) and [3.7](#) (as well as the values of ρ_1, ρ_2, ρ_3) under primitive conditions. This is followed by [Theorem 5.2](#) which provides a CLT for this dependence setting.

Theorem 5.1. Let [Assumptions 3.1\(i\), 3.1\(iv\), 4.3](#) and [3.5](#) hold. If for some $\delta > 0$ we have $\sum_{m=1}^{\infty} \alpha_i(m)^{\delta/(2+\delta)} < \infty$ uniformly over i and $\sum_{m=1}^{\infty} m \tilde{\alpha}_{1,1}(m)^{\delta/(2+\delta)} < \infty$, then the conditions of [Assumption 3.7](#) hold with $\rho_1 = \rho_2 = \rho_3 = 1$. Moreover, $C_{NT} = O_p(N^{-1}T^{-1})$, $D_{NT} = O_p(N^{-1}T^{-1})$, $\mathbb{E}[B_{NT}] = O(T^{-1})$, $\mathbb{E}[B'_{NT}] = O(T^{-1})$, $\mathbb{E}[C_{NT}] = O(N^{-1}T^{-1})$ and $\mathbb{E}[D_{NT}] = O(N^{-1}T^{-1})$ as $N, T \rightarrow \infty$.

If, in addition, $\sum_{m=1}^{\infty} m^3 \tilde{\alpha}_{2,2}(m)^{\delta/(4+\delta)} < \infty$ and $\sum_{m=1}^{\infty} m^3 \tilde{\alpha}_{1,3}(m)^{\delta/(4+\delta)} < \infty$ for some $\delta > 0$, then the conditions of [Assumption 3.6](#) hold, and also B_{NT} and B'_{NT} are both $O_p(T^{-1})$ as $N, T \rightarrow \infty$.

Theorem 5.2. Suppose the following assumptions hold: (i) $S_{it,T}$, $t = 1, \dots, T$, is a $(k \times 1)$ α -mixing sequence with mixing coefficients $\alpha_i(m)$ of size $-(2 + \delta)/\delta$ for some $\delta > 0$, which holds uniformly over i ; (ii) $\sum_{m=1}^{\infty} m^3 \tilde{\alpha}_{2,2}(m)^{\delta/(4+\delta)} < \infty$ and $\sum_{m=1}^{\infty} m^3 \tilde{\alpha}_{1,3}(m)^{\delta/(4+\delta)} < \infty$; (iii) $\max_{i,t,T} \mathbb{E}[\|S_{it,T}\|^{4+\varepsilon}] < \infty$ for some $\varepsilon > 0$; (iv) $\min_{\kappa,N,T} e_{\kappa,N,T} > 0$, where $e_{\kappa,N,T}$, $\kappa = 1, \dots, k$, are the eigenvalues of $W_{NT} = \text{Var}\left((NT)^{-1/2} \sum_{i=1}^N \sum_{t=1}^T S_{it,T}\right)$; (v) $\bar{W}_{\infty} = \lim_{N,T \rightarrow \infty} W_{NT}$ exists. Then,

$$\sqrt{NT} \frac{1}{NT} \sum_{t=1}^T \sum_{i=1}^N S_{it,T} \xrightarrow{d} \mathcal{N}(0, \bar{W}_{\infty}), \quad \text{as } N, T \rightarrow \infty.$$

In [Sections 5.2](#) and [5.3](#) we will provide similar theorems for strong- and cluster-dependence scenarios, with similar types of assumptions. We therefore briefly discuss the assumptions of [Theorems 5.1](#) and [5.2](#): The only novel conditions in [Theorem 5.1](#) are the summability conditions on the mixing coefficients, which determine the panel dependence structure. [Assumption 4.3](#) is sufficient here to obtain the desired results for both the integrated and profile likelihood estimators. Also, as a quick look to the theorem's proof reveals, [Assumption 3.5](#) is only needed to obtain the rates of B_{NT} and $\mathbb{E}[B_{NT}]$. [Theorem 5.2](#) is based on applying a suitable mixing CLT ([Theorem 5.20](#) of [White \(2001\)](#)) on a properly normalised version of the sequence $N^{-1/2} \sum_{i=1}^N S_{it,T}$, $t = 1, \dots, T$. As such, this result is based on typical conditions for CLTs: (i)–(ii) determine the dependence structure of the panel, (iii) is a standard moment assumption which holds under [Assumption 4.3](#), whereas (iv)–(v) impose positive definiteness of the variance of the score and the existence of its limit, which are also standard.⁶ Notice that assumptions (ii)–(iii) of [Theorem 5.2](#) are already implied by the assumptions of [Theorem 5.1](#).

The next corollary summarises the implications of [Theorems 4.3, 4.5, 5.1](#) and [5.2](#).

Corollary 5.1. Let $\hat{\theta}_1 = \hat{\theta}_{IL}$, $\hat{\theta}_2 = \hat{\theta}$, $B_1 = p \lim_{N,T \rightarrow \infty} \sqrt{NT} B_{NT}$ and $B_2 = p \lim_{N,T \rightarrow \infty} \sqrt{NT} B'_{NT}$. Suppose the conditions of [Theorems 4.3, 4.5, 5.1](#) and [5.2](#) hold. Let \bar{W}_{∞} be as defined in [Theorem 5.2](#). Then for $\iota = 1, 2$,

$$\sqrt{NT}(\hat{\theta}_{\iota} - \theta_0) \xrightarrow{d} \mathcal{N}(B_{\iota}, \bar{v}_{\infty}^{-1} \bar{W}_{\infty} \bar{v}_{\infty}^{-1}) \quad \text{as } N, T \rightarrow \infty.$$

Hence, under the considered weak dependence setting $\hat{\theta}_{IL}$ and $\hat{\theta}$ converge at rate \sqrt{NT} , and are subject to the non-vanishing asymptotic bias B_{ι} . This mimics the well-known behaviour of these estimators under cross-section independence.

5.2. Strong dependence

The next setting allows dependence to decrease with distance across time but not across cross-section. To illustrate, consider the normalised cross-sectional sum of covariances for some random variable x_{it} at some point in time:

⁶ The size condition of assumption (i) is required by [Theorem 5.20](#) of [White \(2001\)](#). Note that this cannot be replaced by a summability condition on the mixing coefficient, since size conditions are slightly stronger than summability conditions. See [Definition 3.45](#) of [White \(2001\)](#).

$N^{-2} \sum_{i=1}^N \sum_{j=1}^N \text{Cov}(x_{it}, x_{jt}) = O(N^{-\alpha})$, where $\alpha \in [0, 1]$, assuming that all covariances are finite. If data are cross-sectionally mixing or independent, then under standard conditions $\alpha = 1$. The other extreme where $\alpha = 0$ is what we consider as strong cross-section dependence.⁷ The magnitude of dependence between any two observations across time is assumed to depend on the time distance only, $|t - s|$, for all i, j, s, t . To illustrate, let $r_{A,t}$ and $r_{B,t}$ be the daily equity returns for two firms A and B, respectively, on day t . Then, for example, $\text{Cov}(r_{A,t}, r_{B,t+m})$ and $\text{Cov}(r_{B,t}, r_{B,t+m})$ are assumed to be of the same order of magnitude. This setting would make sense for financial variables, especially during financial downturns when such variables exhibit increased correlation. To formalise these ideas, let

$$\alpha_{i,j}(m) = \sup_t \alpha(\mathcal{F}_{i,-\infty}^t, \mathcal{F}_{j,t+m}^\infty), \quad \alpha_{ij,kl}(m) = \sup_t \alpha(\mathcal{F}_{ij,-\infty}^t, \mathcal{F}_{kl,t+m}^\infty),$$

$$\alpha_{i,jkl}(m) = \sup_t \alpha(\mathcal{F}_{i,-\infty}^t, \mathcal{F}_{jkl,t+m}^\infty) \quad \text{and} \quad \alpha_{ijk,l}(m) = \sup_t \alpha(\mathcal{F}_{ijk,-\infty}^t, \mathcal{F}_{l,t+m}^\infty),$$

where $i, j, k, l = 1, \dots, N$, $\mathcal{F}_{i,-\infty}^t = \sigma(\dots, x_{i,t-1}, x_{it})$, $\mathcal{F}_{i,t+m}^\infty = \sigma(x_{i,t+m}, x_{i,t+m+1}, \dots)$, $\mathcal{F}_{ij,-\infty}^t = \sigma(\dots, x_{i,t-1}, x_{j,t-1}, x_{it}, x_{jt})$, $\mathcal{F}_{ij,t+m}^\infty = \sigma(x_{i,t+m}, x_{j,t+m}, x_{i,t+m+1}, x_{j,t+m+1}, \dots)$ and similarly for $\mathcal{F}_{ijk,-\infty}^t$ and $\mathcal{F}_{ijk,t+m}^\infty$. Intuitively, $\alpha_{i,j}(m)$ measures the dependence between m -period apart observations on individuals i and j . On the other hand, $\alpha_{ijk,l}(m)$ measures the dependence between m -period apart observations that belong to (i) the sigma-algebra generated by individuals i, j, k together and (ii) the sigma-algebra generated by individual l . Such coefficients are useful for controlling the behaviour of terms such as $\text{Cov}(x_{it}x_{jt}x_{kt}, x_{l,t+m})$. Notice that, by definition, $\alpha_{ii,j} = \alpha_{i,j}$.

While assumptions on these mixing coefficients determine the magnitudes of serial and non-contemporaneous cross-section dependence, the strength of contemporaneous dependence is left unspecified. Hence, unlike in the mixing random field setting, accumulation of dependence across strata is not ruled out. This is the source of strong dependence.

Theorem 5.3. Let Assumptions 3.1(i), 3.1(iv), 4.3 and 3.5 hold. If for some $\delta > 0$ we have $\sum_{m=1}^\infty \alpha_{i,j}(m)^{\delta/(2+\delta)} < \infty$ uniformly over $i, j = 1, \dots, N$, then, the conditions of Assumption 3.7 hold with $\rho_1 = \rho_2 = \rho_3 = 0$. Moreover, $C_{NT} = O_p(T^{-1})$, $D_{NT} = O_p(T^{-1})$, $\mathbb{E}[B_{NT}] = O(T^{-1})$, $\mathbb{E}[B'_{NT}] = O(T^{-1})$, $\mathbb{E}[C_{NT}] = O(T^{-1})$ and $\mathbb{E}[D_{NT}] = O(T^{-1})$ as $N, T \rightarrow \infty$.

If, in addition, $\sum_{m=1}^\infty m^2 \alpha_{i,jkl}(m)^{\delta/(4+\delta)} < \infty$, $\sum_{m=1}^\infty m^2 \alpha_{ij,kl}(m)^{\delta/(4+\delta)} < \infty$ and $\sum_{m=1}^\infty m^2 \alpha_{ijk,l}(m)^{\delta/(4+\delta)} < \infty$ uniformly over $i, j, k, l = 1, \dots, N$, for some $\delta > 0$, then the conditions of Assumption 3.6 hold, and also B_{NT} and B'_{NT} are both $O_p(T^{-1})$ as $N, T \rightarrow \infty$.

Theorem 5.4. Suppose the following assumptions hold: (i) $S_{it,T}$, $t = 1, \dots, T$, is a $(k \times 1)$ α -mixing sequence with mixing coefficients $\alpha_i(m)$ of size $-(2+\delta)/\delta$ for some $\delta > 0$, which holds uniformly over i ; (ii) $\sum_{m=1}^\infty \alpha_{i,j}(m)^{\delta/(2+\delta)} < \infty$ uniformly over $i, j = 1, \dots, N$; (iii) $\max_{i,t,T} \mathbb{E}[|S_{it,T}|^{2+\delta}] < \infty$; (iv) $\min_{\kappa,N,T} e_{\kappa,N,T} > 0$, where $e_{\kappa,N,T}$, $\kappa = 1, \dots, k$, are the eigenvalues of $W_{NT} = \text{Var}\left(N^{-1}T^{-1/2} \sum_{i=1}^N \sum_{t=1}^T S_{it,T}\right)$; (v) $\bar{W}_\infty = \lim_{N,T \rightarrow \infty} W_{NT}$ exists. Then,

$$\sqrt{T} \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T S_{it,T} \xrightarrow{d} \mathcal{N}(0, \bar{W}_\infty), \quad \text{as } N, T \rightarrow \infty.$$

Similar to Theorem 5.2, the proof of this result is based on applying Theorem 5.20 of White (2001) to the sequence $N^{-1} \sum_{i=1}^N S_{it,T}$, $t = 1, \dots, T$. However, due to the stronger type of dependence the convergence rate drops to \sqrt{T} . As before, Assumptions (ii) and (iii) are already contained in the Assumptions of Theorem 5.3. The structure of the assumptions for Theorems 5.3 and 5.4 is similar to those of Theorems 5.1 and 5.2, and the comments made there apply here, as well.

Corollary 5.2. Let $\hat{\theta}_1 = \hat{\theta}_{1L}$, $\hat{\theta}_2 = \hat{\theta}$. Suppose the conditions of Theorems 4.3, 4.5, 5.3 and 5.4 hold. Let \bar{W}_∞ be as defined in Theorem 5.4. Then for $\iota = 1, 2$,

$$\sqrt{T}(\hat{\theta}_\iota - \theta_0) \xrightarrow{d} \mathcal{N}(0, \bar{v}_\infty^{-1} \bar{W}_\infty \bar{v}_\infty^{-1}), \quad \text{as } N, T \rightarrow \infty.$$

This confirms that under strong dependence both $\hat{\theta}_{1L}$ and $\hat{\theta}$ obtain the rate \sqrt{T} . Consequently, although B_{NT} , B'_{NT} , C_{NT} and D_{NT} are all $O_p(T^{-1})$, the asymptotic distribution is free of any non-vanishing bias.

5.3. Clusters

Finally, we consider clustered samples; see, e.g., Wooldridge (2003), Bertrand et al. (2004), Hansen (2007) and Cameron and Miller (2011). Let the data consist of $G_N = O(N^\alpha)$ groups/clusters where $0 \leq \alpha \leq 1$. We define \mathcal{G}_g as the index set for group g , where $g = 1, \dots, G_N$. Let, furthermore, $L_{g,N}$ be the number of members of group g . For simplicity, we let

⁷ This is similar to strong dependence as defined by Bailey et al. (2016) in a factor model framework. Unfortunately, a factor model approach is not possible here. This is because we focus on a generic likelihood function and it is impossible to deduce the dependence structure of the likelihood function from the factor structure of the data without imposing a particular model (e.g. probit) on data.

$L_{g,N} = L_N$ for all g . Hence, $L_N = N/G_N$. By following the standard convention, we assume dependence within clusters and independence between clusters. In particular, within-cluster dependence is assumed to be as in Section 5.2.

Theorem 5.5. Let Assumptions 3.1(i), 3.1(iv), 4.3 and 3.5 hold. Suppose that for some $\delta > 0$, $\sum_{m=1}^{\infty} \alpha_i(m)^{\delta/(2+\delta)} < \infty$ uniformly over i . Suppose that the data can be grouped into G_N clusters of size L_N , where $G_N = O(N^\alpha)$. Assume also: (i) x_{is} and x_{jt} are independent for all s, t when they are in different clusters; (ii) the dependence structure between x_{is} and x_{jt} is determined by the Assumptions of Theorem 5.3 whenever they are in the same cluster. Then, the conditions of Assumption 3.7 hold with $\rho_1 = \rho_2 = \rho_3 = \alpha$. Moreover, $C_{NT} = O_p(N^{-\alpha}T^{-1})$, $D_{NT} = O_p(N^{-\alpha}T^{-1})$, $\mathbb{E}[B_{NT}] = O(T^{-1})$, $\mathbb{E}[B'_{NT}] = O(T^{-1})$, $\mathbb{E}[C_{NT}] = O(N^{-\alpha}T^{-1})$ and $\mathbb{E}[D_{NT}] = O(N^{-\alpha}T^{-1})$ as $N, T \rightarrow \infty$. Finally, the conditions of Assumption 3.6 hold, and also B_{NT} and B'_{NT} are both $O_p(T^{-1})$ as $N, T \rightarrow \infty$.

Not surprisingly, the orders of the extra bias terms depend on α . Since $L_N = O(N^{1-\alpha})$, $\alpha = 1$ implies that as $N \rightarrow \infty$ the group size remains bounded while the number of groups tends to infinity. Consequently, cross-section dependence vanishes asymptotically (similar to weak dependence). When $\alpha = 0$, the number of groups remains bounded while group size inflates with N , leading to accumulation of cross-section dependence (similar to strong dependence). Any other α will yield something in between.

Theorem 5.6. Suppose that the following assumptions hold: (i) $S_{it,T}$, $t = 1, \dots, T$, is a $(k \times 1)$ α -mixing sequence. (ii) Data have a cluster structure across the cross-section, with G_N clusters of size L_N where $L_N G_N = N$. Let \mathcal{G}_g be the index set containing the cross-section indices for group g , $g = 1, \dots, G_N$. The cluster setting is such that $S_{is,T}$ and $S_{jt,T}$ are independent whenever $i \in \mathcal{G}_g$, $j \in \mathcal{G}_h$ and $h \neq g$. Moreover, for i, j, k and l that belong to the index set of the same cluster, $\sum_{m=1}^{\infty} m^2 \alpha_{ijkl}(m)^{\delta/(4+\delta)} < \infty$, $\sum_{m=1}^{\infty} m^2 \alpha_{ij,kl}(m)^{\delta/(4+\delta)} < \infty$ and $\sum_{m=1}^{\infty} m^2 \alpha_{ijk,l}(m)^{\delta/(4+\delta)} < \infty$ uniformly over $i, j, k, l = 1, \dots, N$ for some $\delta > 0$. (iii) $\max_{i,t,T} \mathbb{E}[|S_{it,T}|^{4+\varepsilon}] < \infty$ for some $\varepsilon > 0$. (iv) $\min_{\kappa,N,T} e_{\kappa,N,T} > 0$, where $e_{\kappa,N,T}$, $\kappa = 1, \dots, k$, are the eigenvalues of $W_{NT} = \text{Var}\left(\sqrt{G_N T}(NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T S_{it,T}\right)$. (v) $\bar{W}_\infty = \lim_{N,T \rightarrow \infty} W_{NT}$ exists. Then,

$$\sqrt{G_N T} \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T S_{it,T} \xrightarrow{d} \mathcal{N}(0, \bar{W}_\infty), \quad \text{as } N, T \rightarrow \infty.$$

This time we exploit the independence between clusters, rather than weak serial dependence. Therefore, the CLT is based on an application of the Lindeberg–Feller CLT. As before, the structure of the assumptions of Theorems 5.5 and 5.6 is similar to that of Theorems 5.1 and 5.2.

Corollary 5.3. Let $\hat{\theta}_1 = \hat{\theta}_{1L}$, $\hat{\theta}_2 = \hat{\theta}$. Suppose the conditions of Theorems 4.3, 4.5, 5.5 and 5.6 hold. Let \bar{W}_∞ be as defined in Theorem 5.6. Let $0 < \alpha < 1$. Then for $\iota = 1, 2$,

$$\sqrt{N^\alpha T}(\hat{\theta}_\iota - \theta_0) \xrightarrow{d} \mathcal{N}(0, \bar{v}_\infty^{-1} \bar{W}_\infty \bar{v}_\infty^{-1}), \quad \text{as } N, T \rightarrow \infty.$$

Hence, not surprisingly, the incidental parameter bias again ceases to be a non-vanishing asymptotic bias, thanks to rate of convergence dropping below \sqrt{NT} .

6. Bias correction

We now consider bias correction for $\hat{\theta}_{1L}$ and $\hat{\theta}$. Our focus will be on both the bias of the asymptotic distribution and the $O_p(T^{-1})$ small-sample bias. While the terms B_{NT} and B'_{NT} have already been dealt with in the literature, C_{NT} and D_{NT} are novel. This requires a new correction method which can deal with all bias terms. We propose the split-panel jackknife (SPJ) method of Dhaene and Jochmans (2015) (referred to as DJ15 henceforth).

6.1. Profile likelihood estimator

Let $\hat{\theta}$, $\hat{\theta}_{T_1}$ and $\hat{\theta}_{T_2}$ be the profile likelihood estimators based on the full sample, first half-sample and second half-sample, respectively.⁸ The SPJ-corrected profile likelihood estimator is then given by $\hat{\theta}^j = 2\hat{\theta} - (\hat{\theta}_{T_1} + \hat{\theta}_{T_2})/2$.

Theorem 6.1. Suppose x_{it} , $i = 1, \dots, N$ and $t = 1, \dots, T$, is stationary across t for all N and T . Define $\tilde{S}_{NT} = (NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T S_{it}$ and let the following conditions hold: (1) Assumptions 4.1–4.3 and 3.4 hold. (2) $\bar{B}_\infty = \lim_{N,T \rightarrow \infty} T \mathbb{E}[B'_{NT}]$ exists. (3) One of the following options (i)–(iii) holds: (i) The Assumptions of Theorems 5.1 and 5.2 hold, and $\bar{W}_\infty = \lim_{N,T \rightarrow \infty} \text{Var}(\sqrt{NT} \tilde{S}_{NT})$. (ii) The Assumptions of Theorems 5.3 and 5.4 hold, and $\bar{W}_\infty = \lim_{N,T \rightarrow \infty} \text{Var}(\sqrt{T} \tilde{S}_{NT})$.

⁸ Without loss of generality, we let T be even. The case of odd T can be dealt with by considering the two possible half-panel partitions and then taking the average of the resulting two estimators, as in DJ15.

Moreover, $\bar{C}_\infty = \lim_{N,T \rightarrow \infty} T\mathbb{E}[C_{NT}]$ and $\bar{D}_\infty = \lim_{N,T \rightarrow \infty} T\mathbb{E}[D_{NT}]$ exist. (iii) The Assumptions of Theorems 5.5 and 5.6 hold, and $\bar{W}_\infty = \lim_{N,T \rightarrow \infty} \text{Var}(\sqrt{N^\alpha T} S_{NT})$. Moreover, $G_N = N^\alpha$ where $0 < \alpha < 1$.

Then, as $N, T \rightarrow \infty$ we have

$$\sqrt{N^\rho T}(\hat{\theta}^j - \theta_0) \xrightarrow{d} \mathcal{N}(0, \bar{v}_\infty^{-1} \bar{W}_\infty \bar{v}_\infty^{-1}) \quad \text{and} \quad \hat{\theta}^j - \theta_0 = A_{NT} + \beta_{NT} + o_p(T^{-1}),$$

where $\rho = 1$ in case (i), $\rho = 0$ in case (ii), and $\rho = \alpha$ in case (iii). Also, β_{NT} is identically equal to 0 in cases (i) and (iii), whereas $\beta_{NT} = O_p(T^{-1})$ and $\mathbb{E}[\beta_{NT}] = 0$ in case (ii).

The stationarity assumption is a standard requirement for SPJ estimation. Notice that under stationarity $S_{it,T}$ becomes S_{it} as it does not depend on T anymore. Theorem 6.1 confirms that the asymptotic distribution of $\hat{\theta}^j$ is unbiased under all three dependence settings considered here. The asymptotic expansion for $\hat{\theta}^j - \theta_0$ provides further information about the small-sample behaviour of $\hat{\theta}^j$, which we will further discuss in Section 6.2.

6.2. Integrated likelihood estimator

Our result in Eq. (7) is identical to Theorem 1 of AB09 (up to a negligible remainder). As such, the specification of the bias-correcting robust prior $\pi_i^R(\lambda_i|\theta)$ remains the same. Sample versions of two robust prior specifications are provided by AB09 in their Eqs. (12) and (14):

$$\hat{\pi}_i^R(\lambda_i|\theta) \propto \widehat{\mathbb{E}}[-\ell_{it}^{\lambda\lambda}(\theta, \lambda_i)](\widehat{\mathbb{E}}\{\ell_{it}^{\lambda}(\theta, \lambda_i)\}^2)^{-1/2}, \quad (10)$$

$$\hat{\pi}_i^R(\lambda_i|\theta) \propto \{\widehat{\mathbb{E}}[-\ell_{it}^{\lambda\lambda}(\theta, \lambda_i)]\}^{1/2} \exp\left(-\frac{T}{2}\{\widehat{\mathbb{E}}[-\ell_{it}^{\lambda\lambda}(\theta, \lambda_i)]\}^{-1}\widehat{\mathbb{E}}\{\ell_{it}^{\lambda}(\theta, \lambda_i)\}^2\right), \quad (11)$$

where $\widehat{\mathbb{E}}[-\ell_{it}^{\lambda\lambda}(\theta, \lambda_i)]$ and $\widehat{\mathbb{E}}\{\ell_{it}^{\lambda}(\theta, \lambda_i)\}^2$ are consistent estimators of $\mathbb{E}[-\ell_{it}^{\lambda\lambda}(\theta, \lambda_i)]$ and $\mathbb{E}\{\ell_{it}^{\lambda}(\theta, \lambda_i)\}^2$, respectively.⁹ Now, let

$$\begin{aligned} \Pi_{NT} &= \frac{1}{N} \sum_{i=1}^N d_\theta \left[\frac{\ln \pi_i(\bar{\lambda}_{it}(\theta)|\theta)}{T} \right] \Big|_{\theta=\theta_0}, \quad \hat{\Pi}_{NT} = \frac{1}{N} \sum_{i=1}^N d_\theta \left[\frac{\ln \hat{\pi}_i^R(\hat{\lambda}_i(\theta)|\theta)}{T} \right] \Big|_{\theta=\hat{\theta}} \\ B_{NT}'' &= \frac{1}{N} \sum_{i=1}^N d_\theta \left[-\frac{\ln \mathbb{E}[-\ell_{it}^{\lambda\lambda}(\theta, \bar{\lambda}_{it}(\theta))]}{2T} + \frac{[\ell_{it}^{\lambda}(\theta, \bar{\lambda}_{it}(\theta))]^2}{2\mathbb{E}[-\ell_{it}^{\lambda\lambda}(\theta, \bar{\lambda}_{it}(\theta))]} \right] \Big|_{\theta=\theta_0}, \end{aligned}$$

and notice that, for $\mathcal{B}_{it}(\theta)$ as defined in Theorem 4.2, $N^{-1} \sum_{i=1}^N d_\theta \mathcal{B}_{it}(\theta) = \Pi_{NT} + B_{NT}''$. Then, by AB09's definition of robust priors, Π_{NT} based on a robust prior will yield $\Pi_{NT} + \mathbb{E}[B_{NT}''] = o_p(T^{-1})$ as $N, T \rightarrow \infty$. Similarly, if $\hat{\pi}_i^R(\lambda_i|\theta)$ is one of (10) or (11), then $\hat{\Pi}_{NT} + \mathbb{E}[B_{NT}''] = o_p(T^{-1})$ – see Sections 2.3 and 3.1 of AB09.

Let $\hat{\theta}_{IL}$, $\hat{\theta}_{IL,T_1}$ and $\hat{\theta}_{IL,T_2}$ be the integrated likelihood estimators (with the weight function $\hat{\pi}_i^R(\lambda_i|\theta)$) based on the full sample, first half-sample and second-half sample, respectively. Then, the SPJ-corrected robust integrated likelihood estimator is given by $\hat{\theta}_{IL}^j = 2\hat{\theta}_{IL} - (\hat{\theta}_{IL,T_1} + \hat{\theta}_{IL,T_2})/2$.

Theorem 6.2. Let the following conditions hold: (1) Assumptions 3.1–3.5, and 3.8 hold. (2) There exists some $\hat{\pi}_i^R(\lambda_i|\theta)$, such that $\hat{\Pi}_{NT}$ based on $\hat{\pi}_i^R(\lambda_i|\theta)$ yields $\hat{\Pi}_{NT} + \mathbb{E}[B_{NT}''] = o_p(T^{-1})$. (3) One of the cases (i)–(iii) in condition (3) of Theorem 6.1 holds.

Then, the integrated likelihood estimator based on $\hat{\pi}_i^R(\lambda_i|\theta)$ yields

$$\sqrt{N^\rho T}(\hat{\theta}_{IL}^j - \theta_0) \xrightarrow{d} \mathcal{N}(0, \bar{v}_\infty^{-1} \bar{W}_\infty \bar{v}_\infty^{-1}) \quad \text{and} \quad \hat{\theta}_{IL}^j - \theta_0 = A_{NT} + \beta_{NT}^I + o_p(T^{-1}),$$

as $N, T \rightarrow \infty$, where $\rho = 1$ in case (i), $\rho = 0$ in case (ii), and $\rho = \alpha$ in case (iii). Also, β_{NT}^I is $O_p(T^{-1})$ and $\mathbb{E}[\beta_{NT}^I] = \mathbb{E}[C_{NT} + D_{NT}]$ in case (ii), whereas β_{NT}^I is identically equal to 0 in the remaining cases.

Suppose that, in addition, x_{it} , $i = 1, \dots, N$ and $t = 1, \dots, T$, is stationary across t for all N and T . Then in all three cases, as $N, T \rightarrow \infty$ we have

$$\sqrt{N^\rho T}(\hat{\theta}_{IL}^j - \theta_0) \xrightarrow{d} \mathcal{N}(0, \bar{v}_\infty^{-1} \bar{W}_\infty \bar{v}_\infty^{-1}) \quad \text{and} \quad \hat{\theta}_{IL}^j - \theta_0 = A_{NT} + \beta_{NT} + o_p(T^{-1}),$$

where ρ and β_{NT} are as defined in Theorem 6.1 for cases (i)–(iii).

⁹ $\mathbb{E}[-\ell_{it}^{\lambda\lambda}(\theta, \bar{\lambda}_{it}(\theta))]|_{\theta=\theta_0}$ is usually estimated by $\ell_{it}^{\lambda\lambda}(\theta, \hat{\lambda}_i(\theta))|_{\theta=\hat{\theta}}$. As for $\mathbb{E}\{[\ell_{it}^{\lambda}(\theta, \bar{\lambda}_{it}(\theta))]^2\}|_{\theta=\theta_0}$, the classical approach is to use a HAC-type estimator (e.g. Newey and West (1987)). As an alternative, HK11 suggest truncation-based estimation of spectral expectations, which is known to yield better results in analytical bias correction. One reason for this is that, unlike HAC-type estimators, this approach is not restricted to yield positive definite estimates. Notice that calculation of (10) requires a positive value for $\widehat{\mathbb{E}}\{\ell_{it}^{\lambda}(\theta, \lambda_i)\}^2$. So, for this specification HAC estimation is the only option.

Theorems 6.1 and **6.2** prove that all three bias-corrected estimators $(\hat{\theta}^j, \hat{\theta}_{IL}^j, \hat{\theta}_{LL}^j)$ have correctly centred asymptotic distributions, regardless of the dependence setting. Furthermore, the asymptotic expansions confirm that under weak and cluster dependence all methods are free of any $O_p(T^{-1})$ terms. Under strong dependence, a $O_p(T^{-1})$ term remains in the asymptotic expansions of all three estimators. However, for the SPJ-corrected estimators $\hat{\theta}^j$ and $\hat{\theta}_{LL}^j$ this term is zero-mean and so it does not make any contribution to $\mathbb{E}[\hat{\theta}^j - \theta_0]$ and $\mathbb{E}[\hat{\theta}_{LL}^j - \theta_0]$. This is not the case for $\hat{\theta}_{IL}^j$ where the $O_p(T^{-1})$ term makes a $O(T^{-1})$ contribution in expectation. In other words, a second-step SPJ correction is required for correct centering of $\hat{\theta}_{IL}$ in small samples under strong dependence. This suggests that, under strong dependence, $\hat{\theta}^j$ and $\hat{\theta}_{LL}^j$ may be more reliable in small samples. For sake of completeness, we note that this discussion on bias hinges directly on the relative rates of N, T . While the assumption that N, T diverge at the same rate is standard, a change in the relative rates would, naturally, lead to different outcomes. For example, if $N/T^2 = O(1)$ then a higher-order analysis would be called for.

7. Inference

The asymptotic variance matrix for both $\hat{\theta}_{IL}$ and $\hat{\theta}$ has the sandwich form $\bar{v}_\infty^{-1} \bar{W}_\infty \bar{v}_\infty^{-1}$. The inverse Hessian term \bar{v}_∞^{-1} can be estimated by the inverse of the corresponding sample Hessian. The variance of the score $\bar{W}_\infty = \lim_{N,T \rightarrow \infty} \text{Var} \left(\sqrt{N^\rho T} (NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T S_{it,T} \right)$, on the other hand, can be estimated by a suitable HAC variance estimator. At first sight, ρ seems to be a key input here, which requires the econometrician to have accurate knowledge about the dependence structure. However, as **Theorem 7.1** shows, the correct value of ρ is not necessary for obtaining valid confidence intervals, and one can simply use $\rho = 0$. See **Theorem 2.3** of **Gonçalves (2011)** for a similar result.

Theorem 7.1. Suppose the assumptions of **Theorems 5.2** or **5.4** or **5.6** hold. Let $\hat{\theta}_*$ be an estimator of θ_0 such that

$$\sqrt{N^\rho T}(\hat{\theta}_* - \theta_0) = -v^{-1} \frac{\sqrt{N^\rho T}}{NT} \sum_{i=1}^N \sum_{t=1}^T S_{it,T} + o_p(1),$$

with $0 \leq \rho \leq 1$, where the exact value of ρ depends on which one of **Theorems 5.2**, **5.4** and **5.6** is in effect. Let R be some $(Q \times P)$ matrix of rank Q , and r be some $(Q \times 1)$ vector. Let also \hat{v}_{NT} and \hat{W}_{NT} be such that $\hat{v}_{NT} - v = o_p(1)$ and $\hat{W}_{NT} - \text{Var} \left(\sqrt{T} (NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T S_{it,T} \right) = o_p(1)$ as $N, T \rightarrow \infty$. Then, under $H_0 : R\theta_0 = r$

$$T(R\hat{\theta}_* - r)'(R\hat{v}_{NT}^{-1}\hat{W}_{NT}\hat{v}_{NT}^{-1}R')^{-1}(R\hat{\theta}_* - r) \xrightarrow{d} \chi_Q^2.$$

Hence, a valid confidence interval can be constructed given a consistent estimator \hat{W}_{NT} of the asymptotic variance. Using the same idea as in **Driscoll and Kraay (1998)** and **Gonçalves (2011)**, this can be obtained using a HAC variance estimator, by observing the following: since all the settings considered in **Section 5** assume mixing-type serial dependence, the sequence $S_{t,NT} = N^{-1} \sum_{i=1}^N S_{it,T}$ ($t = 1, \dots, T$) is also α -mixing for all N, T , with the same mixing coefficient as for the underlying data x_{it} , independent of the nature of cross-section dependence. As such, $T^{-1} \sum_{t=1}^T S_{t,NT}$ is the sample average of a mixing sequence, which makes it possible to obtain a suitable HAC estimator for $\text{Var} \left(T^{-1/2} \sum_{t=1}^T S_{t,NT} \right)$. Unfortunately, standard estimators, such as **Newey and West (1987)**, are not directly applicable here since the estimation problem involves an incidental parameter. Below we provide a modified version of the Newey–West HAC estimator which takes estimation of the incidental parameter into account. Let $\hat{S}_{t,NT} = N^{-1} \sum_{i=1}^N S_{it,T}(\hat{\theta}_*, \hat{\lambda}_i(\hat{\theta}_*))$ and $S_{t,NT} = N^{-1} \sum_{i=1}^N S_{it,T}(\theta_0, \bar{\lambda}_{iT}(\theta_0))$. Let also $\hat{\Omega}_{j,NT} = T^{-1} \sum_{t=j+1}^T \hat{S}_{t,NT} \hat{S}_{t-j,NT}'$.

Theorem 7.2. Suppose **Assumptions 3.1** and **3.3** hold. In addition suppose that **Assumption 3.4** also holds with the additional condition that the mixing coefficient $\alpha_i(m)$ is of size $(4 + 2r)/r$, uniformly over i , for some $r > 0$. Let $w(j, m) = 1 - [j/(m + 1)]$ where $m = 1, 2, \dots$ and $j = 1, \dots, m$. Let also $m = m(T)$ be such that $\lim_{T \rightarrow \infty} m(T) = \infty$ and $\lim_{T \rightarrow \infty} m(T)T^{-1/4} = 0$. Finally, suppose that there exists some $\hat{\theta}_*$ such that $\hat{\theta}_* - \theta_0 = O_p(T^{-1/2})$, and that $\hat{\lambda}_i(\theta)$ satisfies $\max_{1 \leq i \leq N} |\hat{\lambda}_i(\theta_0) - \bar{\lambda}_{iT}(\theta_0)| = o_p(T^{-2/5})$. Let $\hat{W}_{NT} = \hat{\Omega}_{0,NT} + \sum_{j=1}^{m(T)} w(j, m(T)) [\hat{\Omega}_{j,NT} + \hat{\Omega}_{j,NT}']$. Then,

$$\hat{W}_{NT} - \frac{1}{T} \sum_{s=1}^T \sum_{t=1}^T \mathbb{E} [S_{s,NT} S_{t,NT}'] \xrightarrow{p} 0, \quad \text{as } N, T \rightarrow \infty.$$

This is essentially the same estimator as in **Newey and West (1987)**, except that it also applies to the incidental parameter setting. The only new condition here is the mixing size of $(4 + 2r)/r$.¹⁰ As for the required convergence rates, by **Theorems 6.1** and **6.2**, all estimators under consideration are $O_p(1/\sqrt{N^\rho T})$ where $0 \leq \rho \leq 1$. In addition, that

¹⁰ This is slightly stronger than the original condition in **Newey and West (1987)**. This is because we use the more recent version of Lemma 6.19 of **White (2001)**.

$\max_{1 \leq i \leq N} |\hat{\lambda}_i(\theta_0) - \bar{\lambda}_{iT}(\theta_0)| = o_p(T^{-2/5})$ holds in our setting is shown in the Proof of Lemma S1.17 in the Supplementary Appendix. Finally, [Assumption 3.3](#) is strong enough to imply bound conditions similar to those of Theorem 2 in [Newey and West \(1987\)](#).

Notice that the strong dependence setting with $\rho = 0$ contains the weak and cluster dependence settings (and virtually many other dependence settings) as special cases. Therefore, \hat{W}_{NT} as defined in [Theorem 7.2](#) is a valid, though not necessarily optimal, variance estimator for all the three dependence settings. Knowledge of the true value of ρ would certainly lead to a better estimator but this quantity is rarely, if at all, known to the econometrician. The estimator of [Theorem 7.2](#) can therefore be considered as a robust alternative in the absence of such information.

8. Simulation analysis

8.1. The setting

We compare the performances of four estimators: the profile likelihood estimator without correction (ML), SPJ applied to ML (SPJ) as in [Theorem 6.1](#), the integrated likelihood estimator based on the robust prior (IntL) as in the first part of [Theorem 6.2](#), and SPJ applied to IntL (SPJ_IntL) as in the second part of [Theorem 6.2](#).

We consider four dependence settings for the error term ε_{it} : weak, cluster, strong-factor and strong. In what follows, the first two of these will be referred to as ‘weaker’ types of dependence, whereas the last two will be called ‘stronger’ types of dependence. Throughout, $\gamma = 0.5$ and η_{it} are i.i.d. draws from $\mathcal{N}(0, 1)$. Then: Under *weak dependence*

$$\varepsilon_{it} = \frac{\tilde{\varepsilon}_{it}}{\sqrt{\sum_{|i-j| \leq 2} \gamma^{2|i-j|}}} \quad \text{and} \quad \tilde{\varepsilon}_{it} = \sum_{|i-j| \leq 2} \gamma^{|i-j|} \eta_{jt}.$$

Next, under *strong-factor dependence*

$$\varepsilon_{it} = 0.9\phi_t + \sqrt{1 - 0.9^2}\xi_{it},$$

where ϕ_t and ξ_{it} are i.i.d. draws from $\mathcal{N}(0, 1)$, implying a contemporaneous correlation of 81% between the error terms. Under *strong dependence* and *cluster dependence*, we have

$$\varepsilon_{it} = \frac{\tilde{\varepsilon}_{it}}{\sqrt{1 + \gamma^2 \sum_{j=1}^N A_{ij}}} \quad \text{and} \quad \tilde{\varepsilon}_{it} = \gamma \sum_{j=1}^N \eta_{jt} A_{ij} + \eta_{it},$$

where A_{ij} ($i, j = 1, \dots, N$) is an adjacency matrix, with $A_{ii} = 0$. In the case of *strong dependence* $A_{ij} = 1 \forall i \neq j$, whereas for *cluster dependence* $A_{ij} = 1$ if i and j belong to the same cluster, and $A_{ij} = 0$, otherwise. The number of groups for cluster dependence is given by $G_N = \sqrt{N}$ whereas $L_N = \sqrt{N}$ is the number of group members. Notice that *strong dependence* is a special case of *cluster dependence* with only one cluster.

In Sections 8.2 and 8.3 we consider the dynamic AR(1) and dynamic probit models. For each dependence setting, we report the estimation bias, root-mean squared error (RMSE), ratio of the estimated standard errors to the simulation standard deviation (SE/SD) and the 95% confidence interval coverage rates (CI). To measure the effect of dependence on bias, we also report the estimation bias under cross-section independence. Variance and confidence interval estimations are based on [Theorems 7.1](#) and [7.2](#). For the SPJ and SPJ_IntL estimators, we use the same convention as DJ15 and obtain variance estimates by taking the average of the estimated variances for each half-sample.

8.2. Dynamic AR(1) model

We consider the dynamic autoregressive panel model of order one,

$$y_{it} = \lambda_{i0} + \theta_0 y_{i,t-1} + \varepsilon_{it}.$$

The simulation analysis is based on 500 replications of this model where $T \in \{10, 24, 50, 100\}$, $N = 100$ and $\theta_0 = 0.5$. For each replication λ_i are i.i.d. draws from $\mathcal{N}(0, 1)$.

Simulation results are presented in [Tables 1](#) and [2](#). Results in columns 2–6 of [Table 1](#) reveal that there is a clear difference in bias between weaker and stronger dependence settings, especially for ML. Importantly, all correction methods are successful in removing a large portion of bias (and in many cases almost all of it). Interestingly, at $T = 10$ the integrated likelihood estimator yields better results compared to any jackknife-based approach. At this sample size, the relative performance of SPJ_IntL against IntL worsens dramatically as the magnitude of dependence increases. However, notice that for jackknife-based estimators, which utilise half-samples, this case corresponds to 5 time observations. Moreover, under stronger dependence types, the time-series dimension is the only source of information, which makes the situation even worse for jackknife-based estimators. Nevertheless, SPJ and SPJ_IntL pick up quickly as T increases.

In terms of RMSE (columns 7–10 of [Table 1](#)), IntL is always the best performer though the margin of difference narrows when dependence is of weaker type and/or T is large. Again, this is due to utilisation of $T/2$ time observations by SPJ-based methods, which leads to higher variance. Turning to the 95% confidence interval coverage rates (columns 15–18

Table 1

Simulation results for the Dynamic AR(1) panel model of Section 8.2, under various dependence settings. The considered dependence settings vary in the generation of the error term ε_{it} . Let $\gamma = 0.5$ and η_{it} be i.i.d. draws from $\mathcal{N}(0, 1)$. Then, the dependence settings are as follows. (1) Weak dependence (Wk): $\varepsilon_{it} = \frac{\sum_{|i-j| \leq 2} \gamma^{|i-j|} \eta_{jt}}{\sqrt{\sum_{|i-j| \leq 2} \gamma^{2|i-j|}}}$. (2) Cluster dependence (Cls): $\varepsilon_{it} = \frac{\gamma \sum_{j=1}^N \eta_{jt} A_{ij} + \eta_{it}}{\sqrt{1 + \gamma^2 \sum_{j=1}^N A_{ij}}}$, where $A_{ij} = 1$ if i and j belong to the same cluster ($i, j = 1, \dots, N$), and $A_{ii} = 0$ for all i . There are 10 clusters of equal size 10. (3) Strong dependence (Str): ε_{it} has the same structure as in cluster dependence but $A_{ij} = 1 \forall i \neq j$. (4) Strong-Factor dependence (Fact): $\varepsilon_{it} = 0.9\phi_t + \sqrt{1 - 0.9^2} \xi_{it}$ where ϕ_t and ξ_{it} are i.i.d. draws from $\mathcal{N}(0, 1)$. Column 2 reports the bias under cross section independence (Ind), 3–6 report the bias under the dependence settings, 7–10 report the root-mean squared error (RMSE), 11–14 report the ratio of estimated standard errors to the simulation standard deviation of estimated parameters (SE/SD), and 15–18 report the coverage rates of 95% confidence intervals (CI). All results are based on 500 replications.

$y_{it} = \lambda_{i0} + \theta_0 y_{i,t-1} + \varepsilon_{it}; \theta_0 = 0.5, N = 100$																	
	Bias					RMSE				SE/SD				CI			
	Ind	Wk	Cls	Str	Fact	Wk	Cls	Str	Fact	Wk	Cls	Str	Fact	Wk	Cls	Str	Fact
T = 10																	
ML	-.163	-.165	-.167	-.227	-.210	.172	.191	.378	.314	0.705	0.630	0.554	0.595	0.042	0.284	0.556	0.532
IntL	-.008	-.008	-.010	-.025	-.026	.058	.105	.339	.269	0.742	0.672	1.627	1.000	0.780	0.714	0.718	0.752
SPJ	.048	.050	.050	-.171	-.076	.095	.157	.406	.338	1.459	1.440	1.013	1.207	0.910	0.818	0.671	0.687
SPJ_IntL	-.003	-.006	-.016	-.160	-.146	.082	.155	.456	.378	0.816	0.946	1.320	0.898	0.714	0.668	0.661	0.614
T = 24																	
ML	-.065	-.069	-.071	-.093	-.090	.075	.088	.206	.166	0.824	0.847	0.736	0.828	0.242	0.606	0.736	0.752
IntL	-.001	-.005	-.006	-.018	-.019	.032	.055	.198	.149	0.823	0.772	0.643	0.725	0.858	0.850	0.712	0.746
SPJ	.010	.007	.006	-.003	.003	.038	.066	.229	.177	0.685	0.636	0.645	0.669	0.788	0.758	0.687	0.699
SPJ_IntL	.001	-.003	-.005	-.035	-.028	.037	.065	.239	.179	0.694	0.646	0.594	0.659	0.806	0.764	0.674	0.697
T = 50																	
ML	-.030	-.029	-.035	-.051	-.040	.036	.049	.136	.111	0.894	0.986	0.854	0.835	0.632	0.812	0.856	0.840
IntL	.000	.001	-.004	-.019	-.009	.022	.035	.131	.106	0.869	0.929	0.795	0.776	0.902	0.920	0.850	0.842
SPJ	.003	.005	.000	-.004	.006	.024	.039	.146	.123	0.808	0.837	0.691	0.657	0.886	0.912	0.798	0.764
SPJ_IntL	.001	.003	-.002	-.011	.001	.023	.039	.146	.122	0.811	0.841	0.700	0.663	0.894	0.914	0.790	0.770
T = 100																	
ML	-.016	-.015	-.016	-.025	-.020	.021	.032	.091	.074	0.883	0.889	0.912	0.920	0.766	0.864	0.894	0.910
IntL	.000	.000	-.001	-.010	-.005	.016	.028	.089	.072	0.863	0.858	0.877	0.884	0.894	0.886	0.882	0.902
SPJ	.000	.001	.001	-.001	.002	.016	.030	.095	.077	0.854	0.813	0.802	0.818	0.892	0.874	0.842	0.870
SPJ_IntL	.000	.001	.000	-.002	.000	.016	.030	.095	.076	0.852	0.814	0.807	0.820	0.888	0.872	0.842	0.876

Table 2

Simulation results for the Dynamic AR(1) panel model of Section 8.2. See Table 1 for more information on the dependence settings considered here. ML is the sample bias of the maximum likelihood estimator, B'_{NT} and C_{NT} are the population values of the bias terms as defined in Theorems 4.3 and 4.5. Finally, $ML - B'_{NT} - C_{NT}$ gives the difference between the sample and population biases of the maximum likelihood estimator. All results are based on 500 replications.

$y_{it} = \lambda_{i0} + \theta_0 y_{i,t-1} + \varepsilon_{it}; \theta_0 = 0.5, N = 100$					
	Ind	Wk	Cls	Str	Fact
T = 10					
ML	-.163	-.165	-.167	-.227	-.210
B'_{NT}	-.129	-.129	-.127	-.129	-.131
C_{NT}	-.001	-.002	-.007	-.086	-.061
$ML - B'_{NT} - C_{NT}$	-.033	-.033	-.034	-.013	-.019
T = 24					
ML	-.065	-.069	-.071	-.093	-.090
B'_{NT}	-.060	-.059	-.060	-.060	-.057
C_{NT}	.000	-.001	-.003	-.041	-.024
$ML - B'_{NT} - C_{NT}$	-.005	-.009	-.008	.008	-.009
T = 50					
ML	-.030	-.029	-.035	-.051	-.040
B'_{NT}	-.030	-.029	-.029	-.029	-.029
C_{NT}	.000	.000	-.002	-.019	-.015
$ML - B'_{NT} - C_{NT}$.000	.001	-.004	-.002	.003
T = 100					
ML	-.016	-.015	-.016	-.025	-.020
B'_{NT}	-.015	-.014	-.015	-.016	-.014
C_{NT}	.000	.000	-.001	-.010	-.008
$ML - B'_{NT} - C_{NT}$	-.001	.000	.000	.000	.002

of Table 1) we note that, at small T the coverage rates for ML improve dramatically as the magnitude of cross-section dependence increases. This is in line with our theoretical result that the bias of the asymptotic distribution disappears under convergence rates below \sqrt{NT} . In general, coverage rates improve with T for any dependence setting and estimation method.

The ratios of estimated standard errors to simulation standard deviation (SE/SD) are given in columns 11–14 of Table 1. These suggest that the variance estimator works reasonably well, and tends to improve with T . Also, variance is almost always underestimated. The case $T = 10$ is an exception, with some cases of significant overestimation. However, at such small sample sizes the variance estimator is subject to significant estimation error.

In this application, it is possible to obtain closed-form expressions for the bias terms B'_{NT} , C_{NT} and D_{NT} . Indeed, as derived in Section S3.1 in the Supplementary Appendix,

$$B'_{NT} = \frac{(1 - \theta_0^2)}{N} \sum_{i=1}^N \left(\frac{1}{T} \sum_{t=1}^T \varepsilon_{it} \frac{\lambda_{i0}}{1 - \theta_0} - \frac{1}{T^2} \sum_{t=1}^T \sum_{s=1}^T \varepsilon_{it} y_{i,s-1} \right), \quad (12)$$

$$C_{NT} = (1 - \theta_0^2)^2 \frac{1}{N} \sum_{i=1}^N \left[\frac{2\lambda_{i0}}{1 - \theta_0} \left(\frac{1}{T} \sum_{t=1}^T y_{i,t-1} \right) - \frac{1}{T} \sum_{t=1}^T y_{i,t-1}^2 - \frac{\lambda_{i0}^2}{(1 - \theta_0)^2} + \frac{1}{1 - \theta_0^2} \right] \\ \times \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \varepsilon_{it} \left(y_{i,t-1} - \frac{\lambda_{i0}}{1 - \theta_0} \right), \quad (13)$$

$$D_{NT} = 0. \quad (14)$$

We note that D_{NT} is equal to zero because it depends on the third-order partial derivatives of the log-likelihood, which are all equal to zero due to the linear structure of the model. As such, D_{NT} acts as a non-linearity bias which is absent here.

The closed-form expressions (12)–(14) allow us to directly calculate the population values of the bias terms. In Table 2 we reproduce the sample bias of ML, along with the actual values of B'_{NT} and C_{NT} . We also report the difference between the sample bias and the population bias $B'_{NT} + C_{NT}$. Not surprisingly, B'_{NT} generally remains fixed across dependence types, as it is generated by the error in estimating λ_{i0} , where cross-section dependence is of no consequence. As expected, C_{NT} is virtually equal to zero under weaker dependence types. Importantly, under the setting “strong”, the ratio of C_{NT} to B_{NT} jumps to about 2/3, whereas under the “strong-factor” setting this becomes 1/2. These ratios remain the same across different T . Hence, especially when $T = 10$, the extra bias term C_{NT} is of a significant magnitude relative to the incidental parameter bias. Finally, the difference between the sample and population biases is quite small, and disappears almost completely at $T = 24$, which confirms the validity of the theoretical expressions for B'_{NT} and C_{NT} .

8.3. Dynamic probit model

We next consider the dynamic probit model with an exogenous explanatory variable:

$$y_{it} = 1 \{ \lambda_{i0} + \theta_0 y_{i,t-1} + \delta_0 x_{it} \geq \varepsilon_{it} \}, \\ x_{it} = -\sqrt{2/3} \lambda_{i0} + 0.5 x_{i,t-1} + \epsilon_{it}.$$

Our setting mimics Section 3.3 of DJ15; in particular, ϵ_{it} and λ_i are i.i.d. draws from $\mathcal{N}(0, 1)$. Simulation results are based on 500 replications of the above model for $N = 500$, $T \in \{8, 16, 24, 50\}$, $\delta_0 = 0.5$ and $\theta_0 \in \{0.5, 1\}$. The results are presented in Tables 3–6; on each table, columns 2–6 report bias, 7–10 report RMSE, 11–14 present SE/SD, while the last four columns report the 95% confidence interval coverage rates.

We first consider estimation of δ_0 . Clearly, stronger types of dependence lead to a substantial change in the magnitude of bias for all estimators. All correction methods are successful in reducing a substantial portion of the bias, with SPJ and SPJ_IntL being the clear winners. Importantly, augmenting integrated likelihood with the SPJ method (SPJ_IntL) leads to significant reduction in bias across all dependence types when $T = 8$. We continue to observe the same under stronger types of dependence, even as T increases. As for RMSE, the results are clearly in favour of SPJ under weaker dependence types, whereas IntL delivers better results under stronger dependence types. Again, the likely cause for this is that the SPJ-based methods utilise $T/2$ time observations only. Accordingly, the effect of this diminishes as T increases. As for confidence intervals, SPJ and in some cases SPJ_IntL perform best under weaker types of dependence, whereas IntL is the clear choice under stronger dependence types, again for similar reasons. For a given estimation method and dependence type, the coverage rates generally tend to improve with T . In terms of variance estimation, we observe that IntL almost always delivers the best results. Overall, variance is almost always underestimated.

The results for θ_0 are similar. Stronger types of dependence cause a shift in the bias of $\hat{\theta}$, although interestingly this is not observed with the same intensity for ML. SPJ_IntL almost always leads to improvement over IntL across all T . Unlike for $\hat{\delta}$, this observation is valid even under weaker dependence types. As before, as far as bias is concerned, the best options seem to be SPJ and SPJ_IntL. In terms of RMSE, SPJ is to be preferred under weaker dependence types, while IntL is the clear choice under stronger forms of dependence. We see the same clear distinction between SPJ/SPJ_IntL and

Table 3

Simulation results for estimation of $\theta_0 = 0.5$ in the dynamic probit model of Section 8.3. See Table 1 for more information on the notation and the dependence settings considered here. All results are based on 500 replications.

$y_{it} = 1 \{ \lambda_i + \theta y_{i,t-1} + \delta x_{it} \geq \varepsilon_{it} \}; \quad \theta = 0.5, \delta = 0.5$																	
	Bias ($\hat{\theta}$)					RMSE ($\hat{\theta}$)				SE/SD ($\hat{\theta}$)				CI ($\hat{\theta}$)			
	Ind	Wk	Cls	Str	Fact	Wk	Cls	Str	Fact	Wk	Cls	Str	Fact	Wk	Cls	Str	Fact
T = 8																	
ML	-.386	-.379	-.387	-.366	-.389	.384	.413	.790	.643	1.341	0.819	0.593	0.621	0.002	0.150	0.688	0.666
IntL	-.249	-.244	-.247	-.089	-.156	.252	.274	.437	.374	2.038	1.278	1.046	1.036	0.534	0.662	0.928	0.942
SPJ	.118	.125	.118	.188	.186	.149	.209	.910	.671	2.289	1.158	0.455	0.454	0.988	0.922	0.678	0.684
SPJ_IntL	-.144	-.134	-.152	-.100	-.115	.210	.276	.725	.602	1.947	1.439	0.707	0.719	1.000	0.992	0.840	0.872
T = 16																	
ML	-.180	-.180	-.176	-.199	-.169	.185	.201	.500	.380	0.957	0.819	0.759	0.776	0.008	0.418	0.812	0.826
IntL	-.109	-.110	-.107	-.076	-.081	.117	.141	.368	.303	1.289	0.918	0.977	0.915	0.424	0.730	0.948	0.902
SPJ	.023	.024	.029	.022	.044	.052	.110	.522	.382	1.326	0.811	0.583	0.602	0.972	0.854	0.754	0.794
SPJ_IntL	.030	.029	.030	-.036	-.010	.063	.125	.501	.389	1.599	0.908	0.656	0.654	0.990	0.902	0.832	0.810
T = 24																	
ML	-.116	-.117	-.116	-.111	-.121	.121	.139	.379	.293	0.936	0.870	0.847	0.848	0.050	0.558	0.880	0.868
IntL	-.065	-.067	-.066	-.044	-.065	.074	.100	.324	.254	1.125	0.923	0.956	0.924	0.530	0.804	0.938	0.922
SPJ	.012	.009	.011	.015	.005	.035	.080	.375	.284	1.069	0.827	0.739	0.720	0.948	0.872	0.840	0.836
SPJ_IntL	.025	.023	.025	-.018	-.009	.043	.085	.404	.301	1.375	0.912	0.703	0.699	0.968	0.884	0.832	0.818
T = 50																	
ML	-.054	-.056	-.053	-.056	-.066	.060	.074	.244	.191	0.945	0.931	0.948	0.952	0.278	0.764	0.922	0.912
IntL	-.029	-.028	-.026	-.027	-.039	.039	.060	.232	.183	0.838	0.914	0.975	0.957	0.694	0.890	0.936	0.928
SPJ	.003	.001	.004	.005	-.007	.023	.054	.242	.190	0.910	0.868	0.885	0.854	0.906	0.896	0.902	0.908
SPJ_IntL	.005	.008	.010	-.004	-.007	.037	.062	.262	.201	0.684	0.789	0.820	0.809	0.794	0.866	0.872	0.882

Table 4

Simulation results for estimation of $\delta_0 = 0.5$ in the dynamic probit model of Section 8.3. See Table 1 for more information on the notation and the dependence settings considered here. All results are based on 500 replications.

$y_{it} = 1 \{ \lambda_i + \theta y_{i,t-1} + \delta x_{it} \geq \varepsilon_{it} \}; \quad \theta = 0.5, \delta = 0.5$																	
	Bias ($\hat{\delta}$)					RMSE ($\hat{\delta}$)				SE/SD ($\hat{\delta}$)				CI ($\hat{\delta}$)			
	Ind	Wk	Cls	Str	Fact	Wk	Cls	Str	Fact	Wk	Cls	Str	Fact	Wk	Cls	Str	Fact
T = 8																	
ML	.109	.107	.113	.203	.170	.111	.125	.363	.257	0.652	0.549	0.416	0.443	0.024	0.172	0.594	0.502
IntL	.076	.076	.084	.126	.116	.087	.103	.263	.209	0.671	0.596	0.563	0.532	0.260	0.382	0.670	0.596
SPJ	-.038	-.038	-.041	-.198	-.117	.055	.072	.456	.271	0.522	0.489	0.380	0.272	0.510	0.516	0.504	0.442
SPJ_IntL	-.025	-.024	-.017	-.046	-.020	.127	.124	.329	.244	0.420	0.450	0.522	0.421	0.546	0.630	0.642	0.538
T = 16																	
ML	.050	.052	.051	.078	.067	.055	.060	.174	.128	0.778	0.764	0.679	0.703	0.170	0.472	0.798	0.752
IntL	.026	.028	.027	.045	.039	.034	.041	.149	.110	0.947	0.847	0.743	0.744	0.638	0.768	0.828	0.804
SPJ	-.008	-.007	-.011	-.049	-.023	.022	.035	.179	.116	0.692	0.681	0.541	0.555	0.764	0.764	0.690	0.718
SPJ_IntL	-.026	-.023	-.027	-.034	-.030	.038	.051	.159	.114	0.670	0.622	0.629	0.612	0.652	0.678	0.748	0.744
T = 24																	
ML	.033	.032	.033	.046	.049	.035	.042	.124	.095	0.838	0.836	0.788	0.811	0.342	0.636	0.864	0.834
IntL	.014	.013	.015	.024	.030	.021	.029	.112	.083	0.892	0.869	0.821	0.839	0.780	0.856	0.880	0.894
SPJ	-.004	-.004	-.004	-.019	-.001	.016	.026	.112	.083	0.820	0.770	0.764	0.719	0.880	0.848	0.836	0.852
SPJ_IntL	-.014	-.014	-.014	-.019	-.006	.024	.031	.113	.082	0.780	0.761	0.759	0.736	0.774	0.800	0.834	0.850
T = 50																	
ML	.015	.015	.015	.021	.023	.018	.023	.078	.058	0.930	0.909	0.875	0.910	0.584	0.818	0.910	0.894
IntL	.007	.007	.007	.012	.015	.015	.020	.075	.055	0.723	0.831	0.881	0.904	0.766	0.862	0.912	0.912
SPJ	-.001	-.001	-.001	-.006	.002	.010	.017	.074	.054	0.915	0.869	0.860	0.861	0.928	0.896	0.910	0.896
SPJ_IntL	.001	.001	.001	-.002	.004	.020	.025	.077	.057	0.494	0.627	0.818	0.800	0.662	0.770	0.898	0.882

IntL for confidence interval estimation. It must be underlined that, as in the dynamic AR(1) application, the confidence interval coverage rates for ML improve dramatically (for both parameters) as the magnitude of dependence increases. This again confirms our theoretical findings. Finally, in terms of variance estimation IntL is mostly the better option, with the exception of some cases under weaker dependence types.

As noted previously, following DJ15 we estimate the variances for the SPJ and SPJ_IntL estimators by taking the average of the variance estimates for each half-panel. However, given that both SPJ-methods have the same asymptotic variance as the maximum likelihood estimator, it is also fine to estimate the variance directly, using the full sample. We note

Table 5

Simulation results for estimation of $\theta_0 = 1$ in the dynamic probit model of Section 8.3. See Table 1 for more information on the notation and the dependence settings considered here. All results are based on 500 replications.

$y_{it} = 1 \{ \lambda_i + \theta y_{i,t-1} + \delta x_{it} \geq \varepsilon_{it} \}; \quad \theta = 1, \delta = 0.5$																	
	Bias ($\hat{\theta}$)					RMSE ($\hat{\theta}$)				SE/SD ($\hat{\theta}$)				CI ($\hat{\theta}$)			
	Ind	Wk	Cls	Str	Fact	Wk	Cls	Str	Fact	Wk	Cls	Str	Fact	Wk	Cls	Str	Fact
T = 8																	
ML	-.441	-.436	-.446	-.436	-.413	.441	.473	.876	.683	1.622	0.878	0.567	0.635	0.006	0.200	0.664	0.678
IntL	-.353	-.347	-.354	-.295	-.290	.355	.388	.625	.533	1.727	1.006	0.808	0.807	0.202	0.414	0.824	0.782
SPJ	.076	.078	.066	-.167	-.003	.117	.200	.926	.669	2.360	1.150	0.635	0.500	0.998	0.952	0.660	0.646
SPJ_IntL	-.021	-.007	-.047	-.245	-.133	.158	.268	.865	.727	1.648	1.060	0.661	0.553	0.996	0.964	0.704	0.694
T = 16																	
ML	-.215	-.213	-.217	-.200	-.205	.217	.241	.534	.425	1.040	0.852	0.751	0.741	0.014	0.346	0.832	0.818
IntL	-.157	-.155	-.159	-.139	-.147	.161	.188	.464	.375	1.278	0.943	0.837	0.801	0.170	0.612	0.860	0.852
SPJ	.014	.017	.011	-.034	-.002	.053	.112	.555	.407	1.536	0.913	0.606	0.599	0.992	0.892	0.756	0.748
SPJ_IntL	.041	.047	.031	-.055	-.010	.076	.119	.582	.423	1.561	1.012	0.583	0.599	0.970	0.910	0.750	0.762
T = 24																	
ML	-.137	-.139	-.140	-.124	-.119	.144	.160	.391	.319	0.919	0.925	0.887	0.826	0.038	0.510	0.890	0.862
IntL	-.095	-.098	-.098	-.084	-.079	.104	.124	.362	.295	1.038	0.975	0.934	0.859	0.278	0.712	0.898	0.894
SPJ	.012	.010	.009	.012	.023	.041	.084	.419	.327	1.148	0.881	0.707	0.671	0.956	0.892	0.844	0.818
SPJ_IntL	.024	.021	.019	.013	.029	.047	.087	.422	.332	1.321	0.947	0.701	0.671	0.980	0.910	0.834	0.826
T = 50																	
ML	-.064	-.066	-.066	-.052	-.076	.070	.090	.265	.214	0.965	0.869	0.935	0.926	0.198	0.730	0.920	0.912
IntL	-.040	-.043	-.042	-.028	-.052	.051	.074	.257	.204	0.882	0.878	0.953	0.941	0.578	0.838	0.932	0.926
SPJ	.004	.002	.002	.006	-.014	.025	.063	.272	.209	0.963	0.809	0.842	0.833	0.928	0.878	0.892	0.888
SPJ_IntL	.013	.009	.010	.016	-.005	.038	.068	.276	.212	0.714	0.773	0.828	0.819	0.806	0.860	0.884	0.878

Table 6

Simulation results for estimation of $\delta_0 = 0.5$ in the dynamic probit model of Section 8.3. See Table 1 for more information on the notation and the dependence settings considered here. All results are based on 500 replications.

$y_{it} = 1 \{ \lambda_i + \theta y_{i,t-1} + \delta x_{it} \geq \varepsilon_{it} \}; \quad \theta = 1, \delta = 0.5$																	
	Bias ($\hat{\delta}$)					RMSE ($\hat{\delta}$)				SE/SD ($\hat{\delta}$)				CI ($\hat{\delta}$)			
	Ind	Wk	Cls	Str	Fact	Wk	Cls	Str	Fact	Wk	Cls	Str	Fact	Wk	Cls	Str	Fact
T = 8																	
ML	.135	.138	.135	.263	.207	.144	.148	.410	.300	0.576	0.537	0.371	0.356	0.014	0.148	0.514	0.424
IntL	.113	.119	.113	.186	.165	.128	.130	.305	.251	0.586	0.566	0.503	0.436	0.080	0.246	0.564	0.458
SPJ	-.023	-.022	-.031	-.156	-.123	.051	.076	.437	.324	0.547	0.458	0.448	0.247	0.638	0.572	0.492	0.366
SPJ_IntL	-.039	-.026	-.036	-.014	-.032	.110	.124	.328	.271	0.420	0.439	0.503	0.350	0.610	0.604	0.610	0.464
T = 16																	
ML	.062	.062	.061	.132	.093	.066	.071	.228	.149	0.783	0.750	0.561	0.626	0.086	0.396	0.670	0.656
IntL	.037	.038	.037	.095	.065	.043	.050	.191	.127	0.938	0.835	0.620	0.664	0.538	0.700	0.708	0.724
SPJ	-.008	-.008	-.010	-.040	-.026	.024	.037	.232	.133	0.722	0.671	0.431	0.449	0.806	0.778	0.630	0.616
SPJ_IntL	-.038	-.038	-.040	-.027	-.042	.050	.054	.209	.132	0.625	0.704	0.493	0.486	0.464	0.606	0.690	0.640
T = 24																	
ML	.041	.041	.041	.058	.056	.044	.049	.144	.109	0.805	0.841	0.687	0.701	0.206	0.586	0.806	0.752
IntL	.020	.020	.020	.034	.033	.027	.034	.128	.095	0.892	0.885	0.725	0.734	0.718	0.816	0.822	0.806
SPJ	-.005	-.005	-.006	-.039	-.017	.018	.028	.165	.101	0.789	0.781	0.491	0.557	0.846	0.852	0.724	0.716
SPJ_IntL	-.020	-.020	-.021	-.041	-.029	.028	.036	.155	.100	0.814	0.783	0.528	0.580	0.702	0.766	0.712	0.716
T = 50																	
ML	.019	.019	.020	.036	.024	.021	.026	.086	.061	0.977	0.927	0.879	0.898	0.516	0.780	0.878	0.876
IntL	.009	.009	.009	.024	.014	.016	.021	.080	.057	0.781	0.854	0.893	0.906	0.814	0.852	0.890	0.900
SPJ	-.001	-.001	-.001	-.004	-.004	.010	.018	.082	.058	0.967	0.898	0.790	0.806	0.928	0.906	0.878	0.868
SPJ_IntL	-.002	-.002	-.003	-.002	-.004	.020	.025	.083	.060	0.545	0.654	0.767	0.766	0.688	0.786	0.870	0.852

the following important points: First, in the absence of cross-section dependence, both $\sum_{i=1}^N \ell_{it}(\theta, \lambda_i)$ and $\sum_{i=1}^N \ell_{it}^I(\theta)$ are correctly-specified likelihood functions. Then, by the information equality, the asymptotic variance becomes equal to the inverse Hessian. This object is much easier to estimate than the sandwich-type asymptotic variance matrix under dependence. In addition, under independence the cross-section still carries significant information. Therefore, the inverse sample Hessian based on the half-panels will still be a reasonably accurate estimator. Under dependence, however, the situation is more difficult: one has to obtain both a sample Hessian and a HAC estimator. In addition, it is likely that the cross-section dimension delivers little information. As such, variance estimates based on half-panels can potentially be

Table 7
95% confidence interval coverage rates. Based on two different variance estimation methods. Results for SPJ and SPJ_IntL are based on the average of the variance estimates calculated for each half-panel (as in [Dhaene and Jochmans \(2015\)](#)) – these results have been replicated from [Tables 3–6](#). Results for SPJ* and SPJ_IntL* are based on the variance estimated using the full-sample. See [Table 1](#) for more information on the notation and the dependence settings considered here. All results are based on 500 replications.

	$y_{it} = 1 \{ \lambda_i + \theta y_{i,t-1} + \delta x_{it} \geq \varepsilon_{it} \}; \quad \theta = 0.5, \delta = 0.5$								$y_{it} = 1 \{ \lambda_i + \theta y_{i,t-1} + \delta x_{it} \geq \varepsilon_{it} \}; \quad \theta = 1, \delta = 0.5$							
	CI ($\hat{\delta}$)				CI ($\hat{\theta}$)				CI ($\hat{\delta}$)				CI ($\hat{\theta}$)			
	Wk	Cls	Str	Fact	Wk	Cls	Str	Fact	Wk	Cls	Str	Fact	Wk	Cls	Str	Fact
T = 8																
SPJ	0.510	0.516	0.504	0.442	0.988	0.922	0.678	0.684	0.638	0.572	0.492	0.366	0.998	0.952	0.660	0.646
SPJ*	0.994	0.966	0.670	0.728	1.000	0.996	0.886	0.922	0.986	0.958	0.596	0.662	1.000	1.000	0.810	0.876
SPJ_IntL	0.546	0.630	0.642	0.538	1.000	0.992	0.840	0.872	0.610	0.604	0.610	0.464	0.996	0.964	0.704	0.694
SPJ_IntL*	0.774	0.790	0.670	0.716	0.790	0.768	0.896	0.856	0.874	0.860	0.610	0.630	0.962	0.884	0.740	0.776
T = 16																
SPJ	0.764	0.764	0.690	0.718	0.972	0.854	0.754	0.794	0.806	0.778	0.630	0.616	0.992	0.892	0.756	0.748
SPJ*	0.980	0.938	0.782	0.830	1.000	0.972	0.848	0.866	0.976	0.950	0.686	0.788	1.000	0.980	0.852	0.864
SPJ_IntL	0.652	0.678	0.748	0.744	0.990	0.902	0.832	0.810	0.464	0.606	0.690	0.640	0.970	0.910	0.750	0.762
SPJ_IntL*	0.948	0.876	0.786	0.832	0.994	0.934	0.864	0.816	0.956	0.906	0.706	0.786	1.000	0.984	0.820	0.862
T = 24																
SPJ	0.880	0.848	0.836	0.852	0.948	0.872	0.840	0.836	0.846	0.852	0.724	0.716	0.956	0.892	0.844	0.818
SPJ*	0.958	0.914	0.866	0.898	0.992	0.940	0.878	0.876	0.962	0.950	0.774	0.802	0.996	0.956	0.898	0.888
SPJ_IntL	0.774	0.800	0.834	0.850	0.968	0.884	0.832	0.818	0.702	0.766	0.712	0.716	0.980	0.910	0.834	0.826
SPJ_IntL*	0.918	0.896	0.882	0.896	0.992	0.944	0.866	0.860	0.918	0.914	0.782	0.812	0.996	0.966	0.888	0.878
T = 50																
SPJ	0.928	0.896	0.910	0.896	0.906	0.896	0.902	0.908	0.928	0.906	0.878	0.868	0.928	0.878	0.892	0.888
SPJ*	0.950	0.924	0.914	0.904	0.968	0.942	0.918	0.920	0.956	0.922	0.904	0.902	0.968	0.902	0.910	0.918
SPJ_IntL	0.662	0.770	0.898	0.882	0.794	0.866	0.872	0.882	0.688	0.786	0.870	0.852	0.806	0.860	0.884	0.878
SPJ_IntL*	0.694	0.804	0.906	0.906	0.860	0.878	0.890	0.912	0.744	0.834	0.890	0.890	0.892	0.896	0.912	0.908

misleading. To investigate this, we also estimate the variances and confidence intervals for the SPJ-based methods using the full panel: in [Table 7](#) we present a comparison of the confidence interval coverage rates (the full set of results is in [Tables S.1 and S.2](#) in [Section S3.2](#) in the [Supplementary Appendix](#)). Clearly, using the full-sample variance estimator always leads to better coverage rates for $\hat{\delta}$, across all dependence types. As for $\hat{\theta}$, the coverage rates improve under stronger dependence types, while the situation is ambiguous for cluster dependence. This suggests that in empirical studies there may be a point in employing both approaches.¹¹

9. Empirical example

We consider the empirical application of female labour force participation decision, which has previously been studied in the bias-correction literature (e.g. [Carro \(2007\)](#), [Fernández-Val \(2009\)](#) and [DJ15](#)). We revisit this application using the new methods proposed in this paper, and compare these to previously obtained results in the literature. As discussed below, it is possible that this application is subject to cross-section dependence of some form. Indeed, estimation results fit a pattern associated with this scenario.

We consider the same setting as in [DJ15](#). In particular, the model of interest is

$$y_{it} = 1 \{ \lambda_{i0} + \theta_0 y_{i,t-1} + x'_{it} \delta_0 \geq \varepsilon_{it} \},$$

where ε_{it} are standard normal innovations, y_{it} is a binary variable which equals 1 if individual i is in the labour force at time t , λ_{i0} is an individual fixed effect, and x_{it} is a vector of covariates. The covariates are (i) the number of children between the ages of 0–2, (ii) the number of children between the ages of 3–5, (iii) the number of children between the ages of 6–17, (iv) log of husband income (in thousands of 1995 US Dollars), (v) age, and (vi) the square of age. As in the mentioned previous studies, we use waves 13–22 of Panel Study of Income Dynamics (PSID), which covers the period 1979–1988. The sample contains 1461 women, whose ages were between 18 and 60 in 1985, and who were married to a husband who was in the labour force for the entire sample period. We consider only those individuals who changed their participation status at least once within the sample period. This yields a final dataset of 664 individuals.

The typical assumption here is that ε_{it} are cross-sectionally independent. However, it is not difficult to think of factors which can cause arbitrary patterns of dependence between individuals. For instance, the same labour market shock can

¹¹ Confidence interval construction involves both bias-correction and variance estimation, which are affected by the numerical/computational properties of the underlying estimators. Our results suggest that other variance estimation approaches can potentially yield better coverage rates, especially for SPJ-based methods. Two prominent alternatives are the bootstrap approach as in [Gonçalves \(2011\)](#) and the t-statistic based approach of [Ibragimov and Müller \(2010\)](#). Investigation of these options is left to future work.

affect individuals with different levels or fields of education, or working in different industries, differently. Furthermore, individuals living in particular states or groups of states may exhibit common behaviour. If such factors are not already accounted for by the covariates, they can lead to cluster-type dependence. These factors will not necessarily be absorbed by additive individual fixed effects either, as individuals may move between states, education levels, industries etc. The same will also hold if, e.g., the effect of education depends on age, implying the presence of some interaction effect. It is worth noting that such scenarios can generate intricate dependence patterns where cluster membership varies from wave to wave. We also note that economy-wide macroeconomic shocks can act as common factors, causing (possibly strong) cross-section dependence.

The results of Section 6 suggest that under weaker types of dependence SPJ, IntL and SPJ_IntL should exhibit similar behaviour, as all these estimators are free of a $O_p(T^{-1})$ term. While under strong dependence there remains a $O_p(T^{-1})$ term, for SPJ and SPJ_IntL this term is zero-mean and so makes no $O(T^{-1})$ contribution to $\mathbb{E}[\hat{\theta}^j - \theta_0]$ and $\mathbb{E}[\hat{\theta}_{IL}^j - \theta_0]$. This is not the case for IntL. Therefore, employing either SPJ-based method seems to be the safe option, even when the econometrician cannot be sure of the existence or nature of dependence. All remaining methods (including IntL), on the other hand, are likely to be subject to a non-centred $O_p(T^{-1})$ term since these methods are not designed to deal with the extra bias terms. This also suggests that the presence of sufficiently strong cross-section dependence can manifest itself as systematically different results by SPJ-based vs non-SPJ-based estimators.

Estimation results are presented in Table 8: the top panel provides the results by ML, IntL, SPJ and SPJ_IntL. In the estimation of the SPJ-based methods, we follow DJ15 and consider the $T_1/T_2 = 4/5$ and $T_1/T_2 = 5/4$ partitions, and then take the average of the resulting SPJ estimates. Asymptotic variance can be estimated similarly, by taking the weighted average of the estimated variances from the four sub-samples. This is the case for SPJ and SPJ_IntL. However, following the discussion in Section 8.3 we also consider variance estimation based on the full sample. Indeed, SPJ* and SPJ_IntL* are the same as SPJ and SPJ_IntL, respectively, except that their variance estimates are based on this latter approach. For all the cases on the top panel, standard errors are calculated using $\hat{v}_{NT}^{-1} \hat{W}_{NT} \hat{v}_{NT}^{-1}$, where $\hat{v}_{NT} = d_{\theta\theta} \ell_{NT}(\hat{\theta}, \hat{\lambda}_1(\hat{\theta}), \dots, \hat{\lambda}_N(\hat{\theta}))$ and \hat{W}_{NT} is the HAC estimator of Theorem 7.2. The bottom panel reproduces the results presented in Table 11 of DJ15 for reference. HK is the correction method of HK11, FV is the method of Fernández-Val (2009), SPJ_mlike is the SPJ method of DJ15 applied to the likelihood function, AH is the method of Arellano and Hahn (2016), and C is the method of Carro (2007).¹²

SPJ and SPJ_IntL generally yield highly similar results. Moreover, especially for lagged participation and the first four covariates, there is a clear pattern of difference between SPJ/SPJ_IntL and the remaining correction methods. As noted before, systematically different results for SPJ-based estimators would certainly be the case under strong cross-section dependence. However, it would be rash to reach a conclusion at this stage. As mentioned in DJ15, the model at hand is potentially misspecified and subject to non-stationarity in covariates, which are other possible reasons for obtaining different results under different methods. While our results point to an interesting possibility, a full analysis is unfortunately beyond the scope of this paper. Interestingly, despite being an SPJ-based method, SPJ_mlike does not exhibit the same behaviour as other SPJ-based estimators (in fact, for # Children 0–2, # Children 6–17 and log husband income, SPJ_mlike and IntL yield almost identical results). However, we have not studied this estimator and so it is not clear whether it is also capable of removing the extra bias terms.

Regarding statistical significance, it is interesting that SPJ* and SPJ_IntL* yield identical results, while SPJ and SPJ_IntL largely disagree. This resonates with our discussion in Section 8.3, where it was observed that SPJ* and SPJ_IntL* are more reliable in inference. Looking across all methods, there is almost unanimous agreement that lagged participation, # Children 0–2, # Children 3–5, and log husband income are all relevant variables. Also, the significance results of IntL largely overlap with those of earlier correction methods. Whether age and its square are indeed irrelevant, as suggested by SPJ* and SPJ_IntL*, is difficult to say, especially as almost all the other methods suggest otherwise.

10. Conclusion

This paper studied the bias of the composite integrated likelihood and composite profile likelihood methods in nonlinear dynamic panel models, under dependence across both time and cross-section. Our results show that (i) the asymptotic distribution becomes free of the classical incidental parameter bias under dependence, unless dependence is sufficiently weak; and (ii) the $O_p(T^{-1})$ small-sample bias includes two new terms when dependence is sufficiently strong. In addition, we provide bias correction and inference methods which do not require the econometrician to have knowledge of the underlying dependence structure.

¹² The standard errors reproduced on this bottom panel are all based on the inverse sample Hessian, as a result of the cross-section independence assumption in DJ15. This is also why the ML and SPJ standard errors on the top panel (which are based on $\hat{v}_{NT}^{-1} \hat{W}_{NT} \hat{v}_{NT}^{-1}$) are different than those reported in DJ15.

Table 8

Modelling female labour force participation decision. Estimation results are based on the probit model with $y_{it} = 1 \{ \lambda_{i0} + \theta_0 y_{i,t-1} + x'_{it} \delta_0 \geq \varepsilon_{it} \}$, where ε_{it} are standard normal innovations, y_{it} is a binary variable which equals 1 if individual i is in the labour force at time t , λ_{i0} is an individual fixed effect, and x_{it} is vector of covariates, which consist of (i) the number of children between the ages of 0–2, (ii) the number of children between the ages of 3–5, (iii) the number of children between the ages of 6–17, (iv) log of husband income, (v) age, and (vi) the square of age. The considered methods are maximum likelihood (ML), integrated likelihood (IntL), SPJ applied to ML (SPJ), SPJ applied to IntL (SPJ_IntL), the SPJ method of [Dhaene and Jochmans \(2015\)](#) applied to the likelihood function (SPJ_mlike), as well as the correction methods of [Hahn and Kuersteiner \(2011\)](#) (HK), [Fernández-Val \(2009\)](#) (FV), [Arellano and Hahn \(2016\)](#) (AH), and [Carro \(2007\)](#) (C). The coefficient estimates of age and age squared (and their standard errors) are multiplied by 10 and 100, respectively. SPJ* and SPJ_IntL* are the same as SPJ and SPJ_IntL, respectively, except for the estimation of standard errors. The former use the full-sample whereas the latter take the weighted average of sub-sample variances. Data source: PSID waves 13–22, covering 1979–1988. The sample contains 664 individuals who have changed their participation status at least once.

	ML	IntL	SPJ	SPJ*	SPJ_IntL	SPJ_IntL*
Lagged participation	0.756*** (0.166)	0.863*** (0.190)	1.345*** (0.208)	1.345*** (0.454)	1.518*** (0.505)	1.518*** (0.555)
# Children 0–2	−0.554*** (0.056)	−0.534*** (0.057)	−0.634*** (0.055)	−0.634*** (0.100)	−0.592*** (0.200)	−0.592*** (0.083)
# Children 3–5	−0.279*** (0.023)	−0.269*** (0.026)	−0.338*** (0.043)	−0.338*** (0.102)	−0.327 (0.209)	−0.327*** (0.097)
# Children 6–17	−0.075* (0.041)	−0.068 (0.043)	−0.150*** (0.042)	−0.150** (0.059)	−0.130 (0.238)	−0.130* (0.059)
Log husband income	−0.246*** (0.052)	−0.251*** (0.053)	−0.308*** (0.036)	−0.308*** (0.078)	−0.333*** (0.094)	−0.333*** (0.096)
Age	2.050*** (0.347)	1.987*** (0.404)	1.794*** (0.389)	1.794 (1.953)	1.663 (10.814)	1.663 (1.645)
Age squared	−0.250*** (0.041)	−0.241*** (0.048)	−0.197*** (0.048)	−0.197 (0.161)	−0.183 (0.641)	−0.183 (0.123)
	HK		FV	SPJ_mlike	AH	C
Lagged participation	0.992*** (0.043)		1.031*** (0.043)	1.052*** (0.053)	0.978*** (0.043)	1.095*** (0.043)
# Children 0–2	−0.477*** (0.058)		−0.436*** (0.058)	−0.535*** (0.086)	−0.472*** (0.058)	−0.409*** (0.058)
# Children 3–5	−0.213*** (0.054)		−0.193*** (0.054)	−0.245*** (0.091)	−0.162*** (0.053)	−0.178*** (0.054)
# Children 6–17	−0.056 (0.043)		−0.050 (0.043)	−0.063 (0.078)	0.054 (0.043)	0.040 (0.043)
Log husband income	−0.232*** (0.055)		−0.209*** (0.055)	−0.253*** (0.074)	−0.038 (0.054)	−0.211*** (0.056)
Age	1.844*** (0.392)		1.616*** (0.392)	1.875** (0.874)	−0.173 (0.387)	1.615*** (0.394)
Age squared	−0.224*** (0.052)		−0.196*** (0.052)	−0.228* (0.117)	0.036 (0.052)	−0.194*** (0.053)

*Denote significance at 10%.

**Denote significance at 5%.

***Denote significance at 1%.

Acknowledgements

This paper has previously been circulated under the title “Bias Reduction under Dependence, in a Nonlinear and Dynamic Panel Setting: the Case of GARCH Panels” and is based on two chapters of my DPhil thesis at the University of Oxford. I thank Neil Shephard for his support and guidance. I would also like to express my gratitude to Manuel Arellano, Stéphane Bonhomme, Banu Demir, Shin Kanaya, Kasper Lund-Jensen, Ulrich Müller, Bent Nielsen, Whitney Newey, Andrew Patton, Anders Rahbek, Johannes Ruf, Enrique Sentana, Kevin Sheppard, Michael Streatfield, Martin Weidner, Kemal Yıldız, the Editor, the Associate Editor and three anonymous referees for constructive comments. The hedge fund data used in the Supplementary Appendix has been consolidated by Michael Streatfield and Sushant Vale. I gratefully acknowledge financial support from the European Commission, in the form of Marie Curie Actions Career Integration Grant [Project No 618562]. Part of this work has been undertaken during my visits to CEMFI and Princeton University Department of Economics, and the unbounded hospitality at both institutions is greatly acknowledged. All errors are mine.

Appendix A. Mathematical appendix

This part introduces the notation used in the derivations and contains the proofs of the results presented in the paper, with the exception of the proofs of [Theorems 4.1, 4.4](#) and [6.2](#) which are presented in the online Supplementary Appendix. The Supplementary Appendix also contains Lemmas S1.1–S1.23 and their proofs.

A.1. Index notation

Since θ is a $(P \times 1)$ vector-valued parameter, higher order likelihood expansions around θ get complicated very quickly. To alleviate this problem, we use the *index notation*, which makes it possible to work on multi-dimensional arrays in the same fashion as scalars (see [McCullagh \(1987\)](#)). To establish the main ideas, consider the following examples: let $v = (v_1, \dots, v_P)'$ be some P -dimensional vector. In index notation this vector can be written as $[v_p]$, $p = 1, \dots, P$. Similarly, for a $P \times Q$ matrix A , where the row i column j entry is denoted by A_{ij} , the index notation representation is

given by $[A_{i,j}]$. In the case at hand, $\theta = [\theta_p]$ where $p = 1, \dots, P$. In the following, we will drop the brackets whenever the meaning is clear from the context.

To make the likelihood notation less cumbersome, we will omit the arguments of functions whenever they are evaluated at $(\theta_0, \bar{\lambda}_{iT}(\theta_0)) = (\theta_0, \lambda_{i0})$. We also let

$$\begin{aligned} \ell_i &= \ell_{iT}(\theta_0), \quad \ell_{i,r} = \frac{d\ell_{iT}(\theta)}{d\theta_r} \Big|_{\theta=\theta_0}, \quad \ell_{i,r,s} = \frac{d^2\ell_{iT}(\theta)}{d\theta_r d\theta_s} \Big|_{\theta=\theta_0}, \\ \ell &= \frac{1}{N} \sum_{i=1}^N \ell_{iT}(\theta_0), \quad \ell_r = \frac{1}{N} \sum_{i=1}^N \frac{d\ell_{iT}(\theta)}{d\theta_r} \Big|_{\theta=\theta_0}, \quad \ell_{r,s} = \frac{1}{N} \sum_{i=1}^N \frac{d^2\ell_{iT}(\theta)}{d\theta_r d\theta_s} \Big|_{\theta=\theta_0}, \end{aligned}$$

where $(r, s) \in \{1, \dots, P\}^2$, and similarly for higher order derivatives and the integrated likelihood function $\ell_{iT}^l(\theta)$. Moreover, we use

$$\begin{aligned} \nu_{r,s} &= \mathbb{E}[\ell_{r,s}], \quad \nu_{r,s,t} = \mathbb{E}[\ell_{r,s,t}], \quad \mathcal{H}_{r,s} = \ell_{r,s} - \nu_{r,s}, \quad \mathcal{H}_{r,s,t} = \ell_{r,s,t} - \nu_{r,s,t}, \\ U_i(\theta) &= \ell_{iT}^{\lambda}(\theta), \quad V_i(\theta) = V_{iT}^{\lambda\lambda}(\theta), \quad E_i(\theta) = \mathbb{E}[\ell_{iT}^{\lambda\lambda}(\theta)], \quad \Pi_i(\theta) = \ln \pi_i(\bar{\lambda}_{iT}(\theta)|\theta). \end{aligned}$$

The derivatives of $U_i(\theta)$, $V_i(\theta)$, $E_i(\theta)$ and $\Pi_i(\theta)$ with respect to θ are defined in the same way as for the other likelihood terms. For example,

$$U_{i,r} = \frac{d\ell_{iT}^{\lambda}(\theta)}{d\theta_r} \Big|_{\theta=\theta_0} = \left\{ \frac{d}{d\theta_r} \left[\frac{\partial \ell_{iT}(\theta, \lambda)}{\partial \lambda} \Big|_{\lambda=\bar{\lambda}_{iT}(\theta)} \right] \right\} \Big|_{\theta=\theta_0}.$$

In what follows, when $[\nu_{r,s}]$ is concerned superscripts indicate the corresponding entry of the inverse of $[\nu_{r,s}]$. Then, the row \bar{r} and column \bar{s} entry of the inverse of $[\nu_{r,s}]$ is given by $\nu^{\bar{r},\bar{s}}$. Lastly, $\delta_r = (\hat{\theta}_{IL} - \theta_0)_r$, the r th entry of the $(P \times 1)$ vector $(\hat{\theta}_{IL} - \theta_0)$.

We also use the *Einstein summation convention*. To illustrate, let $p = 1, \dots, P$, $q = 1, \dots, Q$ and consider $\sum_{p=1}^P x_p y_{p,q}$ where $[x_p]$ is a $(P \times 1)$ vector while $[y_{p,q}]$ is a $(P \times Q)$ matrix. Using the Einstein notation, this summation would be written as $x_p y_{p,q}$, making the summation across p implicit. The main idea is that, whenever an index appears twice in a product of arrays, it is to be implicitly understood that the product is summed across this index. Indices that are not repeated within the same term are called *free indices*, and the number of these indices determines the dimension of the resulting array. Indices that are repeated, on the other hand, are called *dummy indices*. As such, $x_p y_{p,q}$ is a vector (one free index, q), while $x_{p,r,s,t} y_{p,q} z_{r,t} = \sum_{p=1}^P \sum_{r=1}^R \sum_{t=1}^T x_{p,r,s,t} y_{p,q} z_{r,t}$ is a two-dimensional matrix (two free indices, q and s). Note that the letters assigned to different indices can be changed freely as long as their relationship is left intact. For example, $x_{p,q} y_{q,r}$ is identical to $x_{q,p} y_{p,r}$, which is equal to xy in standard notation; however, $x_{p,q} y_{p,r}$ is equal to $x'y$.

A.2. Theorems 4.1–4.5

Proof of Theorem 4.1. See the Supplementary Appendix. ■

Proof of Theorem 4.2. Under the maintained assumptions we can invoke Lemmas S1.2, S1.18, S1.19 and S1.20 (notice from the first part of the Proof of Theorem 4.1 that Assumption 3.2(ii) is not needed for obtaining $\max_{1 \leq i \leq N} \sup_{\theta \in \Theta} |\hat{\lambda}_i(\theta) - \bar{\lambda}_{iT}(\theta)| \xrightarrow{P} 0$). By Lemmas S1.18 and S1.20 we obtain

$$\begin{aligned} \ell_{iT}^l(\theta) - \ell_{iT}(\theta) &= \frac{1}{2T} \ln \left(\frac{2\pi}{T} \right) - \frac{1}{2T} \ln [-\ell_{iT}^{\lambda\lambda}(\theta, \hat{\lambda}_i(\theta))] + \frac{1}{T} \ln \pi_i(\hat{\lambda}_i(\theta)|\theta) + \frac{1}{2} \frac{(\ell_{iT}^{\lambda}(\theta))^2}{\mathbb{E}[-\ell_{iT}^{\lambda\lambda}(\theta)]} \\ &\quad + o_p \left(\frac{1}{T^{6/5}} \right). \end{aligned}$$

Using Lemma S1.19, this yields

$$\begin{aligned} \ell_{iT}^l(\theta) - \ell_{iT}(\theta) &= \frac{1}{2T} \ln \left(\frac{2\pi}{T} \right) + \frac{1}{2} \frac{(\ell_{iT}^{\lambda}(\theta))^2}{\mathbb{E}[-\ell_{iT}^{\lambda\lambda}(\theta)]} - \frac{1}{2T} \ln \{-\mathbb{E}[\ell_{iT}^{\lambda\lambda}(\theta)]\} + \frac{1}{T} \ln \pi_i(\bar{\lambda}_{iT}(\theta)|\theta) \\ &\quad - \frac{1}{2T} \underbrace{\frac{V_{iT}^{\lambda\lambda}(\theta)}{\mathbb{E}[\ell_{iT}^{\lambda\lambda}(\theta)]}}_{o_p(T^{-2/5})} + \frac{1}{2T} \underbrace{\frac{\ell_{iT}^{\lambda}(\theta) \mathbb{E}[\ell_{iT}^{\lambda\lambda}(\theta)]}{\{\mathbb{E}[\ell_{iT}^{\lambda\lambda}(\theta)]\}^2}}_{o_p(T^{-2/5})} - \frac{1}{T} \underbrace{\frac{\partial \ln \pi_i(\bar{\lambda}_{iT}(\theta)|\theta)}{\partial \lambda_i} \frac{\ell_{iT}^{\lambda}(\theta)}{\mathbb{E}[\ell_{iT}^{\lambda\lambda}(\theta)]}}_{o_p(T^{-2/5})} \\ &\quad + o_p \left(\frac{1}{T^{6/5}} \right), \end{aligned}$$

where the $o_p(T^{-2/5})$ rates follow from Lemma S1.2 and Assumptions 3.1, 3.3 and 3.5. The stated result follows. ■

Remark A.1 (Remark on the Proof of Theorem 4.3). The standard approach for obtaining an asymptotic expansion for $(\hat{\theta}_{IL} - \theta_0)_r$ is to expand $N^{-1} \sum_{i=1}^N \ell'_{i,r}(\hat{\theta}_{IL})$ about $\hat{\theta}_{IL} = \theta_0$, and rearrange the resulting expansion to obtain an asymptotic representation for $(\hat{\theta}_{IL} - \theta_0)_r$. The difficulty with this approach is that, unlike the profile likelihood, the asymptotic behaviour of the derivatives of the integrated likelihood is not easily obtainable. However, one can approximate the integrated likelihood function by the target likelihood function, which can be dealt with more conveniently. This is done in Lemmas A.1 and A.2. The Proof of Theorem 4.3 combines these results to obtain the desired result.

Lemma A.1. Suppose $\max_{1 \leq i \leq N} \sup_{\theta} |\hat{\lambda}_i(\theta) - \bar{\lambda}_{IT}(\theta)| \xrightarrow{P} 0$ and that Assumptions 3.1 and 3.3–3.5 hold. Then, for any $(r_1, r_2, r_3) \in \{1, \dots, P\}^3$, as $N, T \rightarrow \infty$ we have

$$\begin{aligned} \ell'_{r_1}(\hat{\theta}_{IL}) &= \ell_{r_1} + \frac{1}{2N} \sum_{i=1}^N \left[\frac{U_i^2 E_{i,r_1}}{E_i^2} - 2 \frac{U_i U_{i,r_1}}{E_i} \right] + \frac{1}{NT} \sum_{i=1}^N \left[\frac{E_{i,r_1}}{2E_i} + \Pi_{i,r_1} \right] \\ &\quad + \delta_{r_2} \ell_{r_1,r_2} + \frac{1}{2} \delta_{r_2} \delta_{r_3} \ell_{r_1,r_2,r_3} + o_p \left(\frac{1}{T^{6/5}} \right). \end{aligned} \quad (15)$$

Proof of Lemma A.1. From Theorem 4.2 we have

$$\begin{aligned} \ell^I(\theta) &= \ell(\theta) + \frac{1}{2T} \ln \left(\frac{2\pi}{T} \right) + \frac{1}{2N} \sum_{i=1}^N \frac{(\ell'_{IT}(\theta))^2}{\mathbb{E}[-\ell''_{IT}(\theta)]} - \frac{1}{2NT} \sum_{i=1}^N \ln \{ \mathbb{E}[-\ell''_{IT}(\theta)] \} \\ &\quad + \frac{1}{NT} \sum_{i=1}^N \ln \pi_i(\bar{\lambda}_{IT}(\theta) | \theta) + o_p \left(\frac{1}{T^{6/5}} \right). \end{aligned}$$

Switching to index notation and using Lemmas S1.21 and S1.22, we get

$$\begin{aligned} \ell'_{r_1} &= \ell_{r_1} + \frac{1}{2N} \sum_{i=1}^N \left[\frac{U_i^2 E_{i,r_1}}{E_i^2} - 2 \frac{U_i U_{i,r_1}}{E_i} \right] + \frac{1}{NT} \sum_{i=1}^N \left[\frac{E_{i,r_1}}{2E_i} + \Pi_{i,r_1} \right] + o_p \left(\frac{1}{T^{6/5}} \right), \\ \ell'_{r_1,r_2} &= \ell_{r_1,r_2} + o_p \left(\frac{1}{T^{4/5}} \right) \quad \text{and} \quad \ell'_{r_1,r_2,r_3} = \ell_{r_1,r_2,r_3} + o_p \left(\frac{1}{T^{2/5}} \right), \end{aligned}$$

where $(r_1, r_2, r_3) \in \{1, \dots, P\}^3$. By Theorem 4.1, $\delta_r = o_p(1)$ for all $r \in \{1, \dots, P\}$. Also, by Assumption 3.8, for any $\tilde{\theta}$ such that $\tilde{\theta} \xrightarrow{P} \theta_0$, $\ell'_{r_1,r_2,r_3,r_4}(\tilde{\theta})$ is asymptotically bounded for any $(r_1, r_2, r_3, r_4) \in \{1, \dots, P\}^4$. Then, expanding $N^{-1} \sum_{i=1}^N \ell'_{i,r_1}(\hat{\theta}_{IL})$ about θ_0 we obtain,

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N \ell'_{i,r_1}(\hat{\theta}_{IL}) &= \frac{1}{N} \sum_{i=1}^N \ell'_{i,r_1} + \delta_{r_2} \frac{1}{N} \sum_{i=1}^N \ell'_{i,r_1,r_2} + \delta_{r_2} \delta_{r_3} \frac{1}{2N} \sum_{i=1}^N \ell'_{i,r_1,r_2,r_3} \\ &\quad + \delta_{r_2} \delta_{r_3} \delta_{r_4} \frac{1}{6N} \sum_{i=1}^N \ell'_{i,r_1,r_2,r_3,r_4}(\tilde{\theta}), \end{aligned}$$

where $\tilde{\theta}$ is some mean value between $\hat{\theta}_{IL}$ and θ_0 . Substituting the asymptotic approximations for ℓ'_{r_1} , ℓ'_{r_1,r_2} and ℓ'_{r_1,r_2,r_3} yields the stated result. ■

Lemma A.2. Let Assumptions 3.1–3.8 hold. Then, for any $(a, b, c, d, e, m) \in \{1, \dots, P\}^6$,

$$\delta_m = -\ell_a v^{a,m} - \mathcal{D}_a v^{a,m} + \ell_a v^{a,b} \mathcal{H}_{c,b} v^{c,m} - \frac{1}{2} \ell_a v^{a,b} \ell_c v^{c,d} v_{e,b,d} v^{e,m} + o_p \left(\frac{1}{T^{6/5}} \right)$$

as $N, T \rightarrow \infty$, where

$$\mathcal{D}_a = \frac{1}{N} \sum_{i=1}^N \frac{U_i^2 E_{i,a}}{2E_i^2} - \frac{1}{N} \sum_{i=1}^N \frac{U_i U_{i,a}}{E_i} + \frac{1}{TN} \sum_{i=1}^N \frac{E_{i,a}}{2E_i} + \frac{1}{TN} \sum_{i=1}^N \Pi_{i,a}.$$

Proof of Lemma A.2. Notice that $\ell_{r_1}^l(\hat{\theta}_{ll}) = 0$ by definition for all $r_1 \in \{1, \dots, P\}$. Using $\ell_{r_1, r_2} = v_{r_1, r_2} + \mathcal{H}_{r_1, r_2}$ and $\ell_{r_1, r_2, r_3} = v_{r_1, r_2, r_3} + \mathcal{H}_{r_1, r_2, r_3}$, and rearranging (15), we obtain

$$\begin{aligned} -\delta_{r_2} v_{r_1, r_2} &= \underbrace{\ell_{r_1}}_{O_p(1/\sqrt{N^{\rho_1} T})} + \underbrace{\delta_{r_2} \mathcal{H}_{r_1, r_2}}_{O_p(1/\sqrt{N^{\rho_1 + \rho_2} T^2})} + \underbrace{\frac{1}{NT} \sum_{i=1}^N \left(\frac{E_{i, r_1}}{2E_i} + \Pi_{i, r_1} \right)}_{O_p(1/T)} \\ &\quad + \underbrace{\frac{1}{N} \sum_{i=1}^N \left(\frac{U_i^2 E_{i, r_1}}{2E_i^2} - \frac{U_i U_{i, r_1}}{E_i} \right)}_{O_p(1/T)} + \underbrace{\frac{1}{2} \delta_{r_2} \delta_{r_3} v_{r_1, r_2, r_3}}_{O_p(1/\sqrt{N^2 \rho_1 T^2})} + o_p\left(\frac{1}{T^{6/5}}\right), \end{aligned} \quad (16)$$

as $N, T \rightarrow \infty$. The orders of individual terms follow from Assumptions 3.3, 3.5–3.7. The next step is to invert (16), in order to obtain an expansion for δ_m which is not a function of itself. To do this, first δ_m has to be isolated on the left-hand side. This cannot be done simply by replacing r_2 by m as δ_m appears on both sides of (16). However, notice that if $x^{-1} = [x^{rs}]$ is the inverse of $x = [x_{rs}]$, then $x^{rs} x_{st} = \kappa_t^r$, where $\kappa_t^r = 1$ if $r = t$, and $\kappa_t^r = 0$ otherwise (this is also known as Kronecker delta). Hence,

$$\delta_{r_2} v_{r_1, r_2} v^{r_1, m} = \delta_{r_2} \kappa_{r_2}^m = \begin{cases} \delta_m & \text{if } r_2 = m \\ 0 & \text{if } r_2 \neq m \end{cases}.$$

Then, multiplying both sides of (16) by $v^{r_1, m}$,

$$\delta_m = -\ell_{r_1} v^{r_1, m} - \mathcal{D}_{r_1} v^{r_1, m} - \delta_{r_2} \mathcal{H}_{r_1, r_2} v^{r_1, m} - \frac{1}{2} \delta_{r_2} \delta_{r_3} v_{r_1, r_2, r_3} v^{r_1, m} - o_p\left(\frac{1}{T^{6/5}}\right). \quad (17)$$

Define the following copies of (17): $\delta_{r_2} = -\ell_a v^{a, r_2} + o_p(T^{-4/5})$ and $\delta_{r_3} = -\ell_b v^{b, r_3} + o_p(T^{-4/5})$. Substituting δ_{r_2} and δ_{r_3} in the right-hand side of (17) and renaming the dummy indices in order to simplify the final expression yields the stated result. ■

Proof of Theorem 4.3. From Lemma A.2, we have that

$$\delta_m = -\underbrace{\ell_a v^{a, m}}_{O_p(1/\sqrt{N^{\rho_1} T})} - \underbrace{\mathcal{D}_a v^{a, m}}_{O_p(1/T)} + \underbrace{\ell_a v^{a, b} \mathcal{H}_{c, b} v^{c, m}}_{O_p(1/\sqrt{N^{\rho_1 + \rho_2} T^2})} - \underbrace{\frac{1}{2} \ell_a v^{a, b} \ell_c v^{c, d} v_{e, b, d} v^{e, m}}_{O_p(1/\sqrt{N^2 \rho_1 T^2})} + o_p\left(\frac{1}{T^{6/5}}\right),$$

independent of the particular values of ρ_1 , ρ_2 and ρ_3 . Remember that since we are using the Einstein summation notation, whenever an index appears twice in an expression, it is to be understood that the expression is summed across that index. Now, ℓ_a is row a of the $(P \times 1)$ vector S while $v^{a, m}$ is the row a column m entry of v^{-1} . Therefore, $\ell_a v^{a, m} = \sum_{a=1}^P \ell_a v^{a, m} = (v^{-1})' S = v^{-1} S$, since v^{-1} is symmetric, by definition. Consider now $\ell_a v^{a, b} \ell_c v^{c, d} v_{e, b, d} v^{e, m}$. Notice that $v_{e, b, d}$ is a three dimensional array of the expectations of third-order derivatives of ℓ_{NT} . Now, let P_b and P'_d be the row b entry of $v^{-1} S$ and column d entry of $S' v^{-1}$, respectively. Also, fix e to \bar{e} . Then, $v_{\bar{e}, b, d}$ becomes a $(P \times P)$ matrix and $\ell_a v^{a, b} \ell_c v^{c, d} v_{\bar{e}, b, d}$ is equal to $\sum_{b=1}^P \sum_{c=1}^P \sum_{d=1}^P P_b v_{\bar{e}, b, d} P'_d = M_{\bar{e}}$, the row \bar{e} entry of M using previously defined notation. Hence, $\ell_a v^{a, b} \ell_c v^{c, d} v_{e, b, d} v^{e, m} = \sum_{e=1}^P M_e v^{e, m} = v^{-1} M$. The remaining two terms can be analysed in the same way. The orders of the individual terms have already been determined in the Proof of Lemma A.2. Hence, (8) follows. ■

Proof of Theorem 4.4. See the Supplementary Appendix. ■

Proof of Theorem 4.5. The proof strategy is similar to that of the Proof of Lemma A.2. Let $\bar{\delta}_r$ be the r th entry of $(\hat{\theta} - \theta_0)$. By expanding $\ell_{i; r_1}(\hat{\theta}, \hat{\lambda}_i(\hat{\theta}))$ around $\hat{\theta} = \theta_0$ and averaging both sides of the expansion over i , we obtain

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N \ell_{i; r_1}(\hat{\theta}, \hat{\lambda}_i(\hat{\theta})) &= \frac{1}{N} \sum_{i=1}^N \ell_{i; r_1}(\theta_0, \hat{\lambda}_i(\theta_0)) + \bar{\delta}_{r_2} \frac{1}{N} \sum_{i=1}^N \ell_{i; r_1, r_2}(\theta_0, \hat{\lambda}_i(\theta_0)) \\ &\quad + \bar{\delta}_{r_2} \bar{\delta}_{r_3} \frac{1}{2N} \sum_{i=1}^N \ell_{i; r_1, r_2, r_3}(\theta_0, \hat{\lambda}_i(\theta_0)) \\ &\quad + \bar{\delta}_{r_2} \bar{\delta}_{r_3} \bar{\delta}_{r_4} \frac{1}{6N} \sum_{i=1}^N \ell_{i; r_1, r_2, r_3, r_4}(\tilde{\theta}, \hat{\lambda}_i(\tilde{\theta})), \end{aligned} \quad (18)$$

where $\tilde{\theta}$ is some mean value.¹³ Remember that, by definition, $N^{-1} \sum_{i=1}^N \ell_{i;r_1}(\hat{\theta}, \hat{\lambda}_i(\hat{\theta})) = 0$. Then, using Lemma S1.23 in (18), and rearranging the resulting expression we obtain

$$\begin{aligned} -\bar{\delta}_{r_2} v_{r_1, r_2} &= \ell_{r_1} + \frac{1}{N} \sum_{i=1}^N \left(\frac{U_i^2 E_{i; r_1}}{2E_i^2} - \frac{U_i U_{i; r_1}}{E_i} \right) \\ &\quad + \bar{\delta}_{r_2} \mathcal{H}_{r_1, r_2} + \frac{1}{2} \bar{\delta}_{r_2} \bar{\delta}_{r_3} v_{r_1, r_2, r_3} + \frac{1}{2} \bar{\delta}_{r_2} \bar{\delta}_{r_3} \mathcal{H}_{r_1, r_2, r_3} \\ &\quad + \bar{\delta}_{r_2} o_p \left(\frac{1}{T^{4/5}} \right) + \bar{\delta}_{r_2} \bar{\delta}_{r_3} o_p \left(\frac{1}{T^{2/5}} \right) + o_p \left(\frac{1}{T^{6/5}} \right) \\ &= \underbrace{\ell_{r_1}}_{O_p(1/\sqrt{N^{\rho_1} T})} + \underbrace{\frac{1}{N} \sum_{i=1}^N \left(\frac{U_i^2 E_{i; r_1}}{2E_i^2} - \frac{U_i U_{i; r_1}}{E_i} \right)}_{O_p(1/T)} \\ &\quad + \underbrace{\bar{\delta}_{r_2} \mathcal{H}_{r_1, r_2}}_{O_p(1/\sqrt{N^{\rho_1 + \rho_2} T^2})} + \underbrace{\frac{1}{2} \bar{\delta}_{r_2} \bar{\delta}_{r_3} v_{r_1, r_2, r_3}}_{O_p(1/\sqrt{N^{2\rho_1} T^2})} + o_p \left(\frac{1}{T^{6/5}} \right), \end{aligned} \quad (19)$$

where the orders of individual terms follow from Lemma S1.2 and Assumptions 4.3, 3.6 and 3.7, as in the Proof of Lemma A.2. Using the same methods as for the integrated likelihood estimator, we next invert this result to obtain an expansion for $\bar{\delta}_q$. As explained before, we have $v_{r_1, r_2} v_{r_1, r_2}^{r_1, q} \bar{\delta}_{r_2} = \bar{\delta}_q$, by definition of Kronecker's Delta. Then, multiplying both sides of (19) by $v^{r_1, q}$, using the same inversion idea as in the Proof of Lemma A.2, and finally redefining the indices we obtain

$$\begin{aligned} \bar{\delta}_m &= -\ell_a v^{a, m} + v^{a, m} \frac{1}{N} \sum_{i=1}^N \left(\frac{U_i U_{i; a}}{E_i} - \frac{U_i^2 E_{i; a}}{2E_i^2} \right) \\ &\quad + \ell_d v^{d, b} \mathcal{H}_{a, b} v^{a, m} - \frac{1}{2} \ell_d v^{d, b} \ell_e v^{e, c} v^{a, m} v_{a, b, c} + o_p \left(\frac{1}{T^{6/5}} \right), \end{aligned}$$

which yields the stated result (see the Proof of Theorem 4.3 for conversion of the terms above from index notation to matrix notation). ■

A.3. Theorems 5.1–5.6

Proof of Theorems 5.1, 5.3 and 5.5. The proofs of the three theorems are given together as they follow exactly the same steps.

We start with proving the conditions of Assumption 3.7, under the setting of Theorem 5.1. Consider the following: let $X_{it} = f(x_{it}, \phi_i)$ be some measurable zero-mean function of the random variable x_{it} , where x_{it} is subject to the mixing assumptions of the first part of Theorem 5.1. Then we can invoke Lemma S1.4 to obtain that for any $\varepsilon > 0$,

$$P \left(\left| \frac{\sqrt{NT}}{NT} \sum_{i=1}^N \sum_{t=1}^T X_{it} \right| > \varepsilon \right) \leq \frac{1}{\varepsilon^2} \frac{NT}{N^2 T^2} \sum_{i=1}^N \sum_{j=1}^N \sum_{s=1}^T \sum_{t=1}^T |\mathbb{E}[X_{it} X_{js}]| = \frac{O(1)}{\varepsilon^2},$$

which implies that $(NT)^{-1/2} \sum_{i=1}^N \sum_{t=1}^T X_{it} = O_p(1)$. This result is sufficient to prove that the conditions of Assumption 3.7 hold. This is because, the three terms under consideration in Assumption 3.7 are all zero-mean and inherit the mixing properties of the underlying data. Hence, replacing $(NT)^{-1/2} \sum_{i=1}^N \sum_{t=1}^T X_{it}$ by $\sqrt{NT} d_\theta \ell_{NT}(\theta_0)$, $\sqrt{NT} \{d_{\theta\theta} \ell_{NT}(\theta_0) - \mathbb{E}[d_{\theta\theta} \ell_{NT}(\theta_0)]\}$ and $\sqrt{NT} \{d_{\theta\theta\theta} \ell_{NT}(\theta_0) - \mathbb{E}[d_{\theta\theta\theta} \ell_{NT}(\theta_0)]\}$, one obtains the rate $O_p(1)$ for each of these terms. This implies that under the conditions of the first part of Theorem 5.1, Assumption 3.7 holds with $\rho_1 = \rho_2 = \rho_3 = 1$. The same approach can be used to prove the corresponding results in Theorems 5.3 and 5.5. In particular, under the conditions of Theorem 5.3, one can invoke Lemma S1.5 and prove that Assumption 3.7 holds with $\rho_1 = \rho_2 = \rho_3 = 0$. Finally, under the conditions of Theorem 5.5, Lemma S1.6 can be used to prove that Assumption 3.7 holds with $\rho_1 = \rho_2 = \rho_3 = \alpha$. Observe next that both $\{\mathbb{E}[d_{\theta\theta} \ell_{NT}(\theta_0)]\}^{-1}$ and $\mathbb{E} \left\{ \left[d_{\theta\theta} \frac{d\ell_{NT}(\theta)}{d\theta} \right] \Big|_{\theta=\theta_0} \right\}$ are $O(1)$ as $N, T \rightarrow \infty$ under the maintained assumptions.

¹³ Note that the derivative $\ell_{i; r_1, r_2, r_3, r_4}(\tilde{\theta}, \hat{\lambda}_i(\tilde{\theta}))$ requires the existence of fifth-order partial derivatives of the likelihood. To see this, first remember that, by definition, for any θ we have $\ell_{i; r_1}^{\lambda}(\theta, \hat{\lambda}_i(\theta)) = 0$. Then, $d_\theta \ell_{i; r_1}^{\lambda}(\theta, \hat{\lambda}_i(\theta)) = 0$. At the same time, $d_\theta \ell_{i; r_1}^{\lambda}(\theta, \hat{\lambda}_i(\theta)) = \ell_{i; r_1}^{\lambda\theta}(\theta, \hat{\lambda}_i(\theta)) + \ell_{i; r_1}^{\lambda\lambda}(\theta, \hat{\lambda}_i(\theta)) d_\theta \hat{\lambda}_i(\theta)$. Hence, $d_\theta \hat{\lambda}_i(\theta) = -\ell_{i; r_1}^{\lambda\theta}(\theta, \hat{\lambda}_i(\theta)) / \ell_{i; r_1}^{\lambda\lambda}(\theta, \hat{\lambda}_i(\theta))$. Now, $d_\theta \ell_{i; r_1}(\theta, \hat{\lambda}_i(\theta)) = \ell_{i; r_1}^{\theta}(\theta, \hat{\lambda}_i(\theta)) + \ell_{i; r_1}^{\lambda}(\theta, \hat{\lambda}_i(\theta)) d_\theta \hat{\lambda}_i(\theta)$ and so, due to the term $d_\theta \hat{\lambda}_i(\theta)$, $d_\theta \ell_{i; r_1}(\theta, \hat{\lambda}_i(\theta))$ requires the existence of second-order partial derivatives. By using the same steps, one can show that the same pattern will hold for higher-order derivatives.

Then, given that [Assumption 3.7](#) holds, it directly follows that C_{NT} and D_{NT} are both $O_p(N^{-\rho}T^{-1})$, where $\rho = 1$, $\rho = 0$ and $\rho = \alpha$ under the conditions of [Theorems 5.1, 5.3](#) and [5.5](#), respectively.

Next, we deal with $\mathbb{E}[B_{NT}]$, $\mathbb{E}[B'_{NT}]$, $\mathbb{E}[C_{NT}]$ and $\mathbb{E}[D_{NT}]$. Firstly, by [Assumptions 4.3](#) and [3.5](#), E_i , $E_{i;a}$ and $\Pi_{i;a}$ are all uniformly bounded. Moreover, by Lemma S1.3, under any of the three dependence settings both $\mathbb{E}[U_i^2]$ and $\mathbb{E}[U_i U_{i;a}]$ are $O(T^{-1})$, uniformly over i . Hence, $\mathbb{E}[B'_{NT}] = O(T^{-1})$ and $\mathbb{E}[B_{NT}] = O(T^{-1})$. Finally, we have $\mathbb{E}[C_{NT}] = \mathbb{E}[\ell_a \mathcal{H}_{c,b}] \nu^{a,b} \nu^{c,m}$ and $\mathbb{E}[D_{NT}] = \mathbb{E}[\ell_a \ell_c] \nu^{a,b} \nu^{c,d} \nu_{e,b,d} \nu^{e,m}$, in index notation. The orders of these terms are again determined by Lemmas S1.4–S1.6. In particular, (i) under the assumptions of [Theorem 5.1](#), $\mathbb{E}[\ell_a \mathcal{H}_{c,b}]$ and $\mathbb{E}[\ell_a \ell_c]$ are both $O(N^{-1}T^{-1})$ by Lemma S1.4; (ii) under the assumptions of [Theorem 5.3](#), $\mathbb{E}[\ell_a \mathcal{H}_{c,b}]$ and $\mathbb{E}[\ell_a \ell_c]$ are both $O(T^{-1})$ by Lemma S1.5; and (iii) under the assumptions of [Theorem 5.5](#), $\mathbb{E}[\ell_a \mathcal{H}_{c,b}]$ and $\mathbb{E}[\ell_a \ell_c]$ are both $O(N^{-\alpha}T^{-1})$ by Lemma S1.6.

Finally, we consider the analysis of B'_{NT} and B_{NT} . In index notation, these are given by

$$B'_{NT} = -\nu^{a,m} \left(\frac{1}{N} \sum_{i=1}^N \frac{U_i^2 E_{i;a}}{2E_i^2} - \frac{1}{N} \sum_{i=1}^N \frac{U_i U_{i;a}}{E_i} \right),$$

$$B_{NT} = -\nu^{a,m} \left(\frac{1}{N} \sum_{i=1}^N \frac{U_i^2 E_{i;a}}{2E_i^2} - \frac{1}{N} \sum_{i=1}^N \frac{U_i U_{i;a}}{E_i} + \frac{1}{TN} \sum_{i=1}^N \frac{E_{i;a}}{2E_i} + \frac{1}{TN} \sum_{i=1}^N \Pi_{i;a} \right).$$

Under the assumptions of [Theorems 5.1, 5.3](#) or [5.5](#), Lemma S1.14 can be invoked to prove that the conditions of [Assumption 3.6](#) hold, and consequently that $\frac{1}{N} \sum_{i=1}^N \frac{U_i^2 E_{i;a}}{2E_i^2}$ and $\frac{1}{N} \sum_{i=1}^N \frac{U_i U_{i;a}}{E_i}$ are both uniformly $O_p(T^{-1})$ as $N, T \rightarrow \infty$.

Also, $\nu^{a,m}$ exists by [Assumption 3.1\(iv\)](#). This proves that $B'_{NT} = O_p(T^{-1})$ under each of the three dependence settings. Moreover, as mentioned before, $E_{i;a}$, E_i and $\Pi_{i;a}$ are all uniformly bounded. Hence, we also obtain that $B_{NT} = O_p(T^{-1})$ under each of the three dependence settings. ■

Proof of Theorems 5.2, 5.4 and 5.6. The proofs of these three theorems use similar notation. In particular, throughout λ is some $(k \times 1)$ real vector such that $\lambda' \lambda = 1$ – this must not be confused with the incidental parameter. Also, K is some generic positive finite constant, which can be scalar or matrix-valued, depending on the context.

Part I – Proof of Theorem 5.2: First, notice that by Assumptions (ii) and (iii) we can invoke Lemmas S1.4 and S1.8 to obtain

$$\frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \mathbb{E}[S_{it,T} S'_{jt,T}] = O\left(\frac{1}{N}\right), \quad (20)$$

$$\frac{1}{NT} \sum_{i=1}^N \sum_{j=1}^N \sum_{s=1}^T \sum_{t=1}^T \mathbb{E}[S_{is,T} S'_{jt,T}] = O(1), \quad (21)$$

$$\frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^N \sum_{l=1}^N \mathbb{E}[S'_{it,T} S_{jt,T} S'_{kt,T} S_{lt,T}] = O(1), \quad (22)$$

as $N, T \rightarrow \infty$, where the rates in (20) and (22) hold uniformly over t . Now, let $Z_{t,NT} = N^{-1/2} \sum_{i=1}^N S_{it,T}$ and $v_{t,NT} = \lambda' W_{NT}^{-1/2} Z_{t,NT}$. Then, $\lambda' W_{NT}^{-1/2} Z_{t,NT}$, $t = 1, \dots, T$, is a zero-mean α -mixing sequence for any N and T , with the same mixing coefficients (and size) as $S_{it,T}$. Since $\min_{\kappa, N, T} e_{\kappa, N, T} > 0$ by assumption,

$$\lambda' (W_{NT})^{-1} \lambda \leq \max_{\kappa, N, T} \frac{1}{e_{\kappa, N, T}} = \frac{1}{\min_{\kappa, N, T} e_{\kappa, N, T}} \leq K < \infty,$$

and therefore,

$$\left\| \lambda' W_{NT}^{-1/2} \right\| = \sqrt{\text{tr} \left(\lambda' W_{NT}^{-1/2} W_{NT}^{-1/2} \lambda \right)} = \sqrt{\lambda' W_{NT}^{-1} \lambda} \leq K.$$

Next,

$$\begin{aligned} \mathbb{E} \left[|v_{t,NT}|^4 \right] &= \mathbb{E} \left[\left| \lambda' W_{NT}^{-1/2} Z_{t,NT} \right|^4 \right] \\ &= \mathbb{E} \left[\left(\text{tr} \left(\lambda' W_{NT}^{-1/2} Z_{t,NT} \right) \right)^4 \right] \\ &\leq \mathbb{E} \left[\left\| \lambda' W_{NT}^{-1/2} \right\|^4 \times \|Z_{t,NT}\|^4 \right] \\ &= \left\| \lambda' W_{NT}^{-1/2} \right\|^4 \mathbb{E} \left[\left(\|Z_{t,NT}\|^2 \right)^2 \right] \end{aligned}$$

$$\begin{aligned}
&\leq K \mathbb{E} \left[|Z'_{t,NT} Z_{t,NT}|^2 \right] \\
&= K \mathbb{E} \left[\left| \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N S'_{it,T} S_{jt,T} \right|^2 \right] \\
&\leq K \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^N \sum_{l=1}^N \mathbb{E} [S'_{it,T} S_{jt,T} S'_{kt,T} S_{lt,T}], \tag{23}
\end{aligned}$$

where the second inequality follows from $\|\lambda' W_{NT}^{-1/2}\| \leq K$. By result 4.1.2(2) of [Lütkepohl \(1996\)](#), if A and B are both $(m \times n)$, then $|tr(A'B)| \leq \|A\| \|B\|$. Then, using this result and Hölder's inequality, one can show that (23) is bounded from above by

$$K \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^N \sum_{l=1}^N \left\{ \mathbb{E} [|S_{it,T}|^4] \mathbb{E} [|S_{jt,T}|^4] \mathbb{E} [|S_{kt,T}|^4] \mathbb{E} [|S_{lt,T}|^4] \right\}^{1/4},$$

which is finite for all t, T and N , by the assumption that $\max_{i,t,T} \mathbb{E} [|S_{it,T}|^{4+\varepsilon}] < \infty$ for some $\varepsilon > 0$. Moreover by (22), (23) is $O(1)$ uniformly over t as $N, T \rightarrow \infty$. Hence, $\mathbb{E} [|v_{t,NT}|^4] < \infty$ for all t, T and N (note that by exactly the same arguments and using (20), one can prove the same for $Var(v_{t,NT})$). Finally,

$$\begin{aligned}
Var \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T v_{t,NT} \right) &= Var \left(\lambda' W_{NT}^{-1/2} \frac{1}{\sqrt{T}} \sum_{t=1}^T Z_{t,NT} \right) \\
&= Var \left(\lambda' W_{NT}^{-1/2} \frac{1}{\sqrt{NT}} \sum_{t=1}^T \sum_{i=1}^N S_{it,T} \right) \\
&= \lambda' W_{NT}^{-1/2} Var \left(\frac{1}{\sqrt{NT}} \sum_{t=1}^T \sum_{i=1}^N S_{it,T} \right) W_{NT}^{-1/2} \lambda \\
&= \lambda' W_{NT}^{-1/2} W_{NT} W_{NT}^{-1/2} \lambda \\
&= 1.
\end{aligned}$$

By the results obtained so far and by the assumed mixing size of $-(2 + \delta)/\delta$, we can apply Theorem 5.20 of [White \(2001\)](#) on $\lambda' W_{NT}^{-1/2} Z_{t,NT}$ to obtain $T^{-1/2} \sum_{t=1}^T \lambda' W_{NT}^{-1/2} Z_{t,NT} \xrightarrow{d} \mathcal{N}(0, 1)$. Since this holds for any λ defined as above, by the Cramér-Wold device we have

$$W_{NT}^{-1/2} \frac{1}{\sqrt{NT}} \sum_{t=1}^T \sum_{i=1}^N S_{it,T} \xrightarrow{d} \mathcal{N}(0, I).$$

Part II – Proof of Theorem 5.4: First, notice that by Assumptions (ii) and (iii), we can use Lemma S1.5 to obtain

$$\frac{1}{N^2 T} \sum_{i=1}^N \sum_{j=1}^N \sum_{s=1}^T \sum_{t=1}^T \mathbb{E} [S_{is,T} S'_{jt,T}] = O(1), \tag{24}$$

as $N, T \rightarrow \infty$. Let $Z_{t,NT} = \frac{1}{N} \sum_{i=1}^N S_{it,T}$ and $v_{t,NT} = \lambda' W_{NT}^{-1/2} Z_{t,NT}$. Then, $v_{t,NT}$, $t = 1, \dots, T$, is a zero-mean and α -mixing sequence for any N and T . As in the Proof of [Theorem 5.2](#), since $\min_{\kappa, N, T} e_{\kappa, N, T} > 0$ by assumption, $\lambda' (W_{NT})^{-1} \lambda \leq \frac{1}{\min_{\kappa, N, T} e_{\kappa, N, T}} < \infty$, and so $\|\lambda' W_{NT}^{-1/2}\| \leq K$. Then, for any $\delta > 0$,

$$\begin{aligned}
\mathbb{E} [|v_{t,NT}|^{2+\delta}] &= \mathbb{E} \left[|\lambda' W_{NT}^{-1/2} Z_{t,NT}|^{2+\delta} \right] \\
&= \mathbb{E} \left[\left(\left| tr \left(\lambda' W_{NT}^{-1/2} Z_{t,NT} \right) \right| \right)^{2+\delta} \right] \\
&\leq \mathbb{E} \left[\left\| \lambda' W_{NT}^{-1/2} \right\|^{2+\delta} \times \|Z_{t,NT}\|^{2+\delta} \right] \\
&= \left\| \lambda' W_{NT}^{-1/2} \right\|^{2+\delta} \mathbb{E} \left[\left(\|Z_{t,NT}\|^2 \right)^{1+\delta/2} \right]
\end{aligned}$$

$$\begin{aligned}
&\leq K \mathbb{E} \left[\left| \text{tr} \left(Z'_{t,NT} Z_{t,NT} \right) \right|^{1+\delta/2} \right] \\
&= K \mathbb{E} \left[\left| Z'_{t,NT} Z_{t,NT} \right|^{1+\delta/2} \right] \\
&= K \mathbb{E} \left[\left| \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N S'_{it,T} S_{jt,T} \right|^{1+\delta/2} \right] \\
&\leq K \left(\frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \left\{ \mathbb{E} \left[\left| S'_{it,T} S_{jt,T} \right|^{1+\delta/2} \right] \right\}^{1/(1+\delta/2)} \right)^{1+\delta/2} \\
&\leq K \left(\frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \left\{ \mathbb{E} \left[\left| S_{it,T} \right|^{1+\delta/2} \left| S_{jt,T} \right|^{1+\delta/2} \right] \right\}^{1/(1+\delta/2)} \right)^{1+\delta/2} \\
&\leq K \left(\frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \left\{ \mathbb{E} \left[\left| S_{it,T} \right|^{2+\delta} \right] \mathbb{E} \left[\left| S_{jt,T} \right|^{2+\delta} \right] \right\}^{1/(2+\delta)} \right)^{1+\delta/2},
\end{aligned}$$

where the second, third and fifth inequalities follow from $\left\| \lambda' W_{NT}^{-1/2} \right\| \leq K$, Minkowski's inequality and Hölder's inequality, respectively. As mentioned in the Proof of Theorem 5.2, if A and B are both $(m \times n)$, then $\left| \text{tr}(A'B) \right| \leq \|A\| \|B\|$, which leads to the fourth inequality. Then, by the maintained assumption that $\max_{i,t,T} E \left[\left| S_{it,T} \right|^{2+\delta} \right] < \infty$, we have that $\mathbb{E} \left[\left| v_{t,NT} \right|^{2+\delta} \right] < \infty$ uniformly over t, T and N as $N, T \rightarrow \infty$. By the same arguments it can also be shown that $\text{Var}(v_{t,NT}) < \infty$ uniformly over t, T and N as $N, T \rightarrow \infty$. Next,

$$\begin{aligned}
\text{Var} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T v_{t,NT} \right) &= \text{Var} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T \lambda' W_{NT}^{-1/2} Z_{t,NT} \right) \\
&= \lambda' W_{NT}^{-1/2} \text{Var} \left(\frac{1}{N\sqrt{T}} \sum_{t=1}^T \sum_{i=1}^N S_{it,T} \right) W_{NT}^{-1/2} \lambda \\
&= \lambda' W_{NT}^{-1/2} W_{NT} W_{NT}^{-1/2} \lambda \\
&= 1.
\end{aligned}$$

Hence, under the results obtained so far and by the assumed α -mixing uniform coefficient rate of $-(2+\delta)/\delta$ for some $\delta > 0$, we can invoke Theorem 5.20 of White (2001) to obtain $T^{-1/2} \sum_{t=1}^T \lambda' W_{NT}^{-1/2} Z_{t,NT} \xrightarrow{d} N(0, 1)$, and since this holds for any λ as defined above, by the Cramér-Wold device we have

$$W_{NT}^{-1/2} \frac{1}{N\sqrt{T}} \sum_{t=1}^T \sum_{i=1}^N S_{it,T} \xrightarrow{d} N(0, I), \quad \text{as } N, T \rightarrow \infty.$$

Part III – Proof of Theorem 5.6: We drop the subscript N from L_N and G_N , to keep the notation concise. Using Assumptions (ii) and (iii), we invoke Lemmas S1.5 and S1.9 to obtain

$$\frac{1}{T} \sum_{s=1}^T \sum_{t=1}^T \mathbb{E} [S_{is,T} S'_{jt,T}] = O(1), \quad (25)$$

$$\frac{1}{T^2} \sum_{s=1}^T \sum_{t=1}^T \sum_{q=1}^T \sum_{r=1}^T \mathbb{E} [S'_{is,T} S_{jt,T} S'_{kq,T} S_{lr,T}] = O(1), \quad (26)$$

respectively, which both hold uniformly over i, j, k and l as $N, T \rightarrow \infty$. Let $Z_{g,LT} = L^{-1} T^{-1/2} \sum_{i \in \mathcal{G}_g} \sum_{t=1}^T S_{it,T}$, $\Sigma_{GLT} = G^{-1} \sum_{g=1}^G \text{Var} (Z_{g,LT})$ and $v_{g,GLT} = \lambda' \Sigma_{GLT}^{-1/2} Z_{g,LT}$. Then $v_{g,GLT}$ ($g = 1, \dots, G$) is a sequence of independently distributed and zero-mean variables, for any G, L and T . Since $\min_{k,N,T} e_{k,N,T} > 0$ by assumption, we have $\lambda' \Sigma_{GLT}^{-1} \lambda \leq \frac{1}{\min_{k,N,T} e_{k,N,T}} \leq K$, and so $\left\| \lambda' \Sigma_{GLT}^{-1} \lambda \right\| \leq K$. Then,

$$\begin{aligned}
v_{g,GLT}^2 &= \lambda' \Sigma_{GLT}^{-1/2} Z_{g,LT} Z'_{g,LT} \Sigma_{GLT}^{-1/2} \lambda \\
&= \text{tr} \left(\lambda' \Sigma_{GLT}^{-1/2} Z_{g,LT} Z'_{g,LT} \Sigma_{GLT}^{-1/2} \lambda \right)
\end{aligned}$$

$$\begin{aligned}
&= \text{tr} \left(Z_{g,LT}' Z_{g,LT}' \Sigma_{GLT}^{-1/2} \lambda \lambda' \Sigma_{GLT}^{-1/2} \right) \\
&\leq \|Z_{g,LT}' Z_{g,LT}'\| \times \left\| \Sigma_{GLT}^{-1/2} \lambda \lambda' \Sigma_{GLT}^{-1/2} \right\| \\
&= \sqrt{\text{tr} \left(Z_{g,LT}' Z_{g,LT}' Z_{g,LT}' Z_{g,LT}' \right)} \sqrt{\text{tr} \left(\Sigma_{GLT}^{-1/2} \lambda \lambda' \Sigma_{GLT}^{-1/2} \Sigma_{GLT}^{-1/2} \lambda \lambda' \Sigma_{GLT}^{-1/2} \right)} \\
&= \sqrt{\text{tr} \left(Z_{g,LT}' Z_{g,LT}' Z_{g,LT}' Z_{g,LT}' \right)} \sqrt{\text{tr} \left(\lambda' \Sigma_{GLT}^{-1/2} \Sigma_{GLT}^{-1/2} \lambda \lambda' \Sigma_{GLT}^{-1/2} \Sigma_{GLT}^{-1/2} \lambda \right)} \\
&= \sqrt{\text{tr} \left(Z_{g,LT}' Z_{g,LT}' Z_{g,LT}' Z_{g,LT}' \right)} \sqrt{\text{tr} \left(\lambda' \Sigma_{GLT}^{-1} \lambda \lambda' \Sigma_{GLT}^{-1} \lambda \right)} \\
&= \|Z_{g,LT}'\|^2 \|\lambda' \Sigma_{GLT}^{-1} \lambda\| \\
&\leq K \|Z_{g,LT}'\|^2.
\end{aligned}$$

Now, for any $\varepsilon > 0$,

$$\begin{aligned}
\mathbb{E} \left[G^{-1} v_{g,GLT}^2 \mathbf{1} \{ G^{-1} v_{g,GLT}^2 > \varepsilon \} \right] &= G^{-1} \mathbb{E} \left[v_{g,GLT}^2 \mathbf{1} \{ v_{g,GLT}^2 > \varepsilon G \} \right] \\
&\leq KG^{-1} \mathbb{E} \left[\|Z_{g,LT}'\|^2 \mathbf{1} \left\{ \|Z_{g,LT}'\|^2 > \frac{\varepsilon G}{K} \right\} \right] \\
&\leq KG^{-1} \mathbb{E} \left[\|Z_{g,LT}'\|^2 \mathbf{1} \left\{ \|Z_{g,LT}'\|^4 > \frac{\varepsilon^2 G^2}{K^2} \right\} \right] \\
&\leq KG^{-1} \sqrt{\mathbb{E} \left[\|Z_{g,LT}'\|^4 \right]} \sqrt{\mathbb{E} \left[\mathbf{1} \left\{ \|Z_{g,LT}'\|^4 > \frac{\varepsilon^2 G^2}{K^2} \right\}^2 \right]} \\
&\leq KG^{-1} \sqrt{\mathbb{E} \left[\|Z_{g,LT}'\|^4 \right]} \sqrt{\mathbb{E} \left[\mathbf{1} \left\{ \|Z_{g,LT}'\|^4 > \frac{\varepsilon^2 G^2}{K^2} \right\} \right]} \\
&= KG^{-1} \sqrt{\mathbb{E} \left[\|Z_{g,LT}'\|^4 \right]} \sqrt{P \left(\|Z_{g,LT}'\|^4 > \frac{\varepsilon^2 G^2}{K^2} \right)} \\
&\leq KG^{-1} \sqrt{\mathbb{E} \left[\|Z_{g,LT}'\|^4 \right]} \sqrt{\frac{K^2}{\varepsilon^2 G^2} \mathbb{E} \left[\|Z_{g,LT}'\|^4 \right]} \\
&= \frac{K}{\varepsilon G^2} \mathbb{E} \left[\|Z_{g,LT}'\|^4 \right].
\end{aligned} \tag{27}$$

Notice that

$$\begin{aligned}
\mathbb{E} \left[\|Z_{g,LT}'\|^4 \right] &= \mathbb{E} \left[\left(\sqrt{\text{tr} \left(Z_{g,LT}' Z_{g,LT}' \right)} \right)^4 \right] \\
&= \mathbb{E} \left[\text{tr} \left(Z_{g,LT}' Z_{g,LT}' \right)^2 \right] \\
&= \mathbb{E} \left[Z_{g,LT}' Z_{g,LT}' Z_{g,LT}' Z_{g,LT}' \right] \\
&= \frac{1}{L^4 T^2} \sum_{i \in \mathcal{G}_g} \sum_{j \in \mathcal{G}_g} \sum_{k \in \mathcal{G}_g} \sum_{l \in \mathcal{G}_g} \sum_{q=1}^L \sum_{r=1}^T \sum_{s=1}^T \sum_{t=1}^T \mathbb{E} \left[S'_{iq,T} S_{jr,T} S'_{ks,T} S_{lt,T} \right] \\
&= O(1),
\end{aligned} \tag{28}$$

uniformly over g as $N, T \rightarrow \infty$, by (26). Then, by (27) and (28), for any $\varepsilon > 0$ we have $\mathbb{E} \left[G^{-1} v_{g,GLT}^2 \mathbf{1} \{ G^{-1} v_{g,GLT}^2 > \varepsilon \} \right] = O(G^{-2})$, and so,

$$\lim_{G \rightarrow \infty} \sum_{g=1}^G \mathbb{E} \left[G^{-1} v_{g,GLT}^2 \mathbf{1} \{ G^{-1} v_{g,GLT}^2 > \varepsilon \} \right] = 0,$$

which confirms that the Lindeberg condition holds for $v_{g,GLT}$ (notice that $G^{-1} \sum_{g=1}^G \text{Var}(v_{g,GLT}) = 1$). We can therefore apply Theorem 5.6 of White (2001) $v_{g,GLT}$ to obtain $G^{-1/2} \sum_{g=1}^G v_{g,GLT} \xrightarrow{d} N(0, 1)$. Since this holds for any λ defined as

above, by the Cramér-Wold device we obtain

$$\Sigma_{GLT}^{-1/2} \frac{1}{L\sqrt{GT}} \sum_{g=1}^G \sum_{i \in \mathcal{G}_g}^L \sum_{t=1}^T S_{it,T} \xrightarrow{d} N(0, I),$$

or equivalently,

$$\Sigma_{GLT}^{-1/2} \frac{\sqrt{GT}}{NT} \sum_{i=1}^N \sum_{t=1}^T S_{it,T} \xrightarrow{d} N(0, I),$$

as desired. ■

A.4. Theorems 6.1 and 6.2

Proof of Theorem 6.1. Let $\bar{B}'_{NT} = B'_{NT} - \mathbb{E}[B'_{NT}]$, $\bar{C}_{NT} = C_{NT} - \mathbb{E}[C_{NT}]$ and $\bar{D}_{NT} = D_{NT} - \mathbb{E}[D_{NT}]$. Under any of the three cases (i)–(iii), the conditions of Assumptions 3.6 and 3.7 are obtained. Therefore, all assumptions of Theorem 4.5 hold, which we invoke to obtain the expansion (9). Then,

$$\begin{aligned} \hat{\theta} - \theta_0 &= A_{NT} + B'_{NT} + C_{NT} + D_{NT} + o_p\left(\frac{1}{T^{6/5}}\right) \\ &= A_{NT} + \frac{\bar{B}_{\infty}}{T} + \underbrace{\bar{B}'_{NT}}_{o_p(N^{-\rho/2}T^{-1})} + \underbrace{\mathbb{E}[C_{NT}] + \mathbb{E}[D_{NT}]}_{O(N^{-\rho}T^{-1})} + \underbrace{\bar{C}_{NT} + \bar{D}_{NT}}_{o_p(N^{-\rho}T^{-1})} \\ &\quad + o_p\left(\frac{1}{T}\right), \end{aligned} \quad (29)$$

where $\rho = 1$ under weak dependence, $\rho = 0$ under strong dependence, and $\rho = \alpha$ under cluster dependence. Here, the rates of $\mathbb{E}[C_{NT}]$ and $\mathbb{E}[D_{NT}]$ follow from Theorems 5.1, 5.3 and 5.5 under weak, strong and cluster dependence settings, respectively. Moreover, the rates of \bar{B}'_{NT} , \bar{C}_{NT} and \bar{D}_{NT} follow from Lemmas S1.14 and S1.15, which can be invoked under the assumptions of any of the three cases. Before moving on to the analysis of $\hat{\theta}^j$, we note that under the stationarity assumption $\nu = \mathbb{E}\left[N^{-1} \sum_{i=1}^N d_{\theta\theta} \ell_{it}(\theta_0, \bar{\lambda}_i(\theta_0))\right]$, where $\bar{\lambda}_i(\theta) = \arg \max_{\lambda} \mathbb{E}[\ell_{it}(\theta, \lambda)]$. So, ν does not depend on T . Similarly, $S_{it,T}$ becomes

$$S_{it} = \left\{ \ell_{it}^{\theta}(\theta_0, \bar{\lambda}_i(\theta_0)) - \frac{\mathbb{E}[\ell_{it}^{\lambda\theta}(\theta_0, \bar{\lambda}_i(\theta_0))]}{\mathbb{E}[\ell_{it}^{\lambda\lambda}(\theta_0, \bar{\lambda}_i(\theta_0))]} \ell_{it}^{\lambda}(\theta_0, \bar{\lambda}_i(\theta_0)) \right\}. \quad (30)$$

Hence, under stationarity

$$\begin{aligned} A_{NT} &= -\nu^{-1} \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T S_{it}, \\ A_{NT_1} &= -\nu^{-1} \frac{2}{NT} \sum_{i=1}^N \sum_{t=1}^{T/2} S_{it}, \quad A_{NT_2} = -\nu^{-1} \frac{2}{NT} \sum_{i=1}^N \sum_{t=T/2+1}^T S_{it}, \end{aligned}$$

where we have used $T_1 = T_2 = T/2$. Then, noting that $A_{NT_1} + A_{NT_2} = 2A_{NT}$, we obtain

$$2A_{NT} - \frac{1}{2}A_{NT_1} - \frac{1}{2}A_{NT_2} = A_{NT}. \quad (31)$$

This result is important since it implies that the SPJ estimator will still be subject to the same CLT as the uncorrected estimator.

We now consider the behaviour of $\hat{\theta}^j$ under each dependence scenario. Under (i) we have weak dependence, where $\rho = 1$. Then, (29) becomes

$$\hat{\theta} = \theta_0 + A_{NT} + \frac{\bar{B}_{\infty}}{T} + o_p\left(\frac{1}{T}\right).$$

The same expansion is also valid for $\hat{\theta}_{T_1}$ and $\hat{\theta}_{T_2}$, after adjusting for the size of the time dimension:

$$\begin{aligned} \hat{\theta}_{T_1} &= \theta_0 + A_{NT_1} + \frac{\bar{B}_{\infty}}{T_1} + o_p\left(\frac{1}{T}\right), \\ \hat{\theta}_{T_2} &= \theta_0 + A_{NT_2} + \frac{\bar{B}_{\infty}}{T_2} + o_p\left(\frac{1}{T}\right). \end{aligned}$$

Then,

$$\begin{aligned}\hat{\theta}^j - \theta_0 &= 2A_{NT} - \frac{1}{2}A_{NT_1} - \frac{1}{2}A_{NT_2} + \frac{2\bar{B}_\infty}{T} - \frac{1}{2}\frac{\bar{B}_\infty}{T_1} - \frac{1}{2}\frac{\bar{B}_\infty}{T_2} + o_p\left(\frac{1}{T}\right) \\ &= A_{NT} + \frac{2\bar{B}_\infty}{T} - \frac{\bar{B}_\infty}{T} - \frac{\bar{B}_\infty}{T} + o_p\left(\frac{1}{T}\right) \\ &= A_{NT} + o_p\left(\frac{1}{T}\right),\end{aligned}$$

where we have used (31). This confirms that β_{NT} is identically equal to zero in this case. Moreover, by Theorem 5.2

$$\sqrt{NT}(\hat{\theta}^j - \theta_0) = \sqrt{NT}A_{NT} + o_p(1) \xrightarrow{d} \mathcal{N}(0, \bar{v}_\infty^{-1}\bar{W}_\infty\bar{v}_\infty^{-1}),$$

where \bar{W}_∞ is as defined in the conditions of case (i).

Next, we consider strong dependence, which corresponds to case (ii). Since $\rho = 0$, (29) becomes

$$\begin{aligned}\hat{\theta} &= \theta_0 + A_{NT} + \frac{\bar{B}_\infty}{T} + \underbrace{\mathbb{E}[C_{NT}] + \mathbb{E}[D_{NT}]}_{o(T^{-1})} + \underbrace{\bar{B}'_{NT} + \bar{C}_{NT} + \bar{D}_{NT}}_{o_p(T^{-1})} + o_p\left(\frac{1}{T}\right) \\ &= \theta_0 + A_{NT} + \frac{\bar{B}_\infty + \bar{C}_\infty + \bar{D}_\infty}{T} + \bar{B}'_{NT} + \bar{C}_{NT} + \bar{D}_{NT} + o_p\left(\frac{1}{T}\right),\end{aligned}$$

and similarly,

$$\begin{aligned}\hat{\theta}_{T_1} &= \theta_0 + A_{NT_1} + \frac{\bar{B}_\infty + \bar{C}_\infty + \bar{D}_\infty}{T_1} + \bar{B}'_{NT_1} + \bar{C}_{NT_1} + \bar{D}_{NT_1} + o_p\left(\frac{1}{T}\right), \\ \hat{\theta}_{T_2} &= \theta_0 + A_{NT_2} + \frac{\bar{B}_\infty + \bar{C}_\infty + \bar{D}_\infty}{T_2} + \bar{B}'_{NT_2} + \bar{C}_{NT_2} + \bar{D}_{NT_2} + o_p\left(\frac{1}{T}\right).\end{aligned}$$

Then, using (31) as before, and remembering that $T_1 = T_2 = T/2$, one obtains

$$\hat{\theta}^j - \theta_0 = A_{NT} + \tilde{B}_{NT} + \tilde{C}_{NT} + \tilde{D}_{NT} + o_p\left(\frac{1}{T}\right),$$

where $\tilde{B}_{NT} = 2\bar{B}'_{NT} - (\bar{B}'_{NT_1} + \bar{B}'_{NT_2})/2$, and \tilde{C}_{NT} and \tilde{D}_{NT} are defined similarly. Let $\beta_{NT} = \tilde{B}_{NT} + \tilde{C}_{NT} + \tilde{D}_{NT}$. Notice that \tilde{B}_{NT} , \tilde{C}_{NT} and \tilde{D}_{NT} are all $O_p(T^{-1})$ and, by definition, zero-mean; hence

$$\hat{\theta}^j - \theta_0 = A_{NT} + \beta_{NT} + o_p\left(\frac{1}{T}\right),$$

where $\beta_{NT} = O_p(T^{-1})$ and $\mathbb{E}[\beta_{NT}] = 0$, as stated. Moreover, using Theorem 5.4

$$\sqrt{T}(\hat{\theta}^j - \theta_0) = \sqrt{T}A_{NT} + O_p\left(\frac{1}{\sqrt{T}}\right) \xrightarrow{d} \mathcal{N}(0, \bar{v}_\infty^{-1}\bar{W}_\infty\bar{v}_\infty^{-1}),$$

where \bar{W}_∞ is as defined in the conditions of case (ii).

Finally, we consider case (iii) which is cluster dependence. We have $\rho = \alpha$ with $0 < \alpha < 1$, so (29) yields

$$\begin{aligned}\hat{\theta} &= \theta_0 + A_{NT} + \frac{\bar{B}_\infty}{T} + o_p(T^{-1}), \\ \hat{\theta}_{T_1} &= \theta_0 + A_{NT_1} + \frac{\bar{B}_\infty}{T_1} + o_p(T^{-1}), \\ \hat{\theta}_{T_2} &= \theta_0 + A_{NT_2} + \frac{\bar{B}_\infty}{T_2} + o_p(T^{-1}).\end{aligned}$$

Therefore,

$$\hat{\theta}^j - \theta_0 = A_{NT} + o_p(T^{-1}),$$

which confirms that β_{NT} is identically equal to zero in this case. Finally, by Theorem 5.6

$$\sqrt{N^\alpha T}(\hat{\theta}^j - \theta_0) = \sqrt{N^\alpha T}A_{NT} + o_p(1) \xrightarrow{d} \mathcal{N}(0, \bar{v}_\infty^{-1}\bar{W}_\infty\bar{v}_\infty^{-1}),$$

where \bar{W}_∞ is as defined in case (iii). ■

Proof of Theorem 6.2. See the Supplementary Appendix. ■

A.5. Theorems 7.1 and 7.2

Proof of Theorem 7.1. Under the assumptions of Theorems 5.2, 5.4 and 5.6 we have

$$-\nu^{-1} \frac{\sqrt{N^\rho T}}{NT} \sum_{i=1}^N \sum_{t=1}^T S_{it,T} \xrightarrow{d} \mathcal{N}(0, \bar{\nu}_\infty^{-1} \bar{W}_\infty \bar{\nu}_\infty^{-1}),$$

where $\bar{W}_\infty = \lim_{N,T \rightarrow \infty} \text{Var} \left(\sqrt{N^\rho T} (NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T S_{it,T} \right)$. Then, it follows that

$$N^\rho T (R\hat{\theta}_* - r)' (R\bar{\nu}_\infty^{-1} \bar{W}_\infty \bar{\nu}_\infty^{-1} R')^{-1} (R\hat{\theta}_* - r) \xrightarrow{d} \chi_Q^2. \quad (32)$$

Define $W = \text{Var} \left(\sqrt{T} (NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T S_{it,T} \right)$ and notice that $N^\rho W - \bar{W}_\infty = o_p(1)$ as $N, T \rightarrow \infty$. Also, $\hat{\nu}_{NT} - \nu = o_p(1)$ and $\hat{W}_{NT} - W = o_p(1)$ by assumption. Moreover, $\nu - \bar{\nu}_\infty = o(1)$. Therefore,

$$\begin{aligned} & T(R\hat{\theta}_* - r)' \left(R\hat{\nu}_{NT}^{-1} \hat{W}_{NT} \hat{\nu}_{NT}^{-1} R' \right)^{-1} (R\hat{\theta}_* - r) \\ &= T(R\hat{\theta}_* - r)' (R\bar{\nu}_\infty^{-1} W \bar{\nu}_\infty^{-1} R')^{-1} (R\hat{\theta}_* - r) + o_p(1) \\ &= N^\rho T (R\hat{\theta}_* - r)' (R\bar{\nu}_\infty^{-1} N^\rho W \bar{\nu}_\infty^{-1} R')^{-1} (R\hat{\theta}_* - r) + o_p(1) \\ &= N^\rho T (R\hat{\theta}_* - r)' (R\bar{\nu}_\infty^{-1} \bar{W}_\infty \bar{\nu}_\infty^{-1} R')^{-1} (R\hat{\theta}_* - r) + o_p(1) \xrightarrow{d} \chi_Q^2, \end{aligned}$$

as stated, where convergence in distribution to χ_Q^2 follows from (32). ■

Proof of Theorem 7.2. The proof is based on the same steps as the Proof of Theorem 2 of Newey and West (1987) – referred to as NW henceforth. As in their approach, we focus on the scalar case here. Let

$$\bar{S}_{NT} = \frac{1}{T} \sum_{t=1}^T \hat{h}_{t,NT}^2 + \frac{2}{T} \sum_{j=1}^m w(j, m) \sum_{t=j+1}^T \hat{h}_{t,NT} \hat{h}_{t-j,NT},$$

and

$$S_{NT} = \frac{1}{T} \sum_{s=1}^T \sum_{t=1}^T \mathbb{E} [h_{t,NT} h_{s,NT}],$$

where

$$\begin{aligned} h_{t,NT}(\theta, \lambda_1, \dots, \lambda_N) &= \frac{1}{N} \sum_{i=1}^N S_{it,T}(\theta, \lambda_i), \\ \hat{h}_{t,NT} &= h_{t,NT}(\hat{\theta}_*, \hat{\lambda}_1(\hat{\theta}_*), \dots, \hat{\lambda}_N(\hat{\theta}_*)) \quad \text{and} \quad h_{t,NT} = h_{t,NT}(\theta_0, \bar{\lambda}_{1T}(\theta_0), \dots, \bar{\lambda}_{NT}(\theta_0)). \end{aligned}$$

Notice that \bar{S}_{NT} , S_{NT} , $\hat{h}_{t,NT}$ and $h_{t,NT}$ are akin to \bar{S}_T , S_T , \hat{h}_t and h_t of NW, respectively. Let K be some generic non-negative and finite constant, the exact value of which may change from usage to usage. Now, by Assumption 3.3 we have $\sup_{\theta, \lambda_i} |\ell_{it}^\theta(\theta, \lambda_i)| \leq d(x_{it})$ and $\sup_{\theta, \lambda_i} |\ell_{it}^\lambda(\theta, \lambda_i)| \leq d(x_{it})$. In addition, let $C = \inf_{i,T} \inf_{\theta, \lambda_i} \mathbb{E} [|\ell_{it}^{\lambda\lambda}(\theta, \lambda_i)|]$. From Assumption 3.1(iv) we have $C > 0$. Next,

$$\sup_{\theta, \lambda_i} \mathbb{E} [|\ell_{it}^{\lambda\theta}(\theta, \lambda_i)|] \leq \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\sup_{\theta, \lambda_i} |\ell_{it}^{\lambda\theta}(\theta, \lambda_i)| \right] \leq \frac{1}{T} \sum_{t=1}^T \mathbb{E} [|d(x_{it})|] \leq K < \infty,$$

since $\sup_{i,t} \mathbb{E} [|d(x_{it})|] \leq K < \infty$, by the last part of Assumption 3.3. Hence,

$$\begin{aligned} \sup_{\theta, \lambda_i} |S_{it,T}(\theta, \lambda_i)| &= \sup_{\theta, \lambda_i} \left| \ell_{it}^\theta(\theta, \lambda_i) - \frac{\mathbb{E} [\ell_{it}^{\lambda\theta}(\theta, \lambda_i)]}{\mathbb{E} [\ell_{it}^{\lambda\lambda}(\theta, \lambda_i)]} \ell_{it}^\lambda(\theta, \lambda_i) \right| \\ &\leq \sup_{\theta, \lambda_i} |\ell_{it}^\theta(\theta, \lambda_i)| + \sup_{\theta, \lambda_i} \left| \frac{\mathbb{E} [\ell_{it}^{\lambda\theta}(\theta, \lambda_i)]}{\mathbb{E} [\ell_{it}^{\lambda\lambda}(\theta, \lambda_i)]} \right| \sup_{\theta, \lambda_i} |\ell_{it}^\lambda(\theta, \lambda_i)| \\ &\leq d(x_{it}) + \frac{\sup_{\theta, \lambda_i} \mathbb{E} [|\ell_{it}^{\lambda\theta}(\theta, \lambda_i)|]}{\inf_{\theta, \lambda_i} \mathbb{E} [|\ell_{it}^{\lambda\lambda}(\theta, \lambda_i)|]} d(x_{it}) \end{aligned}$$

$$\begin{aligned}
&\leq d(x_{it}) + \frac{K}{C} d(x_{it}) \\
&\leq Kd(x_{it}).
\end{aligned} \tag{33}$$

Next, using

$$\begin{aligned}
d_{\theta} S_{it,T}(\theta, \bar{\lambda}_{iT}(\theta)) &= \ell_{it}^{\theta\theta}(\theta, \bar{\lambda}_{iT}(\theta)) + \ell_{it}^{\theta\lambda}(\theta, \bar{\lambda}_{iT}(\theta)) \bar{\lambda}'_{iT}(\theta) \\
&\quad + \bar{\lambda}_{iT}''(\theta) \ell_{it}^{\lambda\lambda}(\theta, \bar{\lambda}_{iT}(\theta)) + \bar{\lambda}'_{iT}(\theta) \ell_{it}^{\theta\lambda}(\theta, \bar{\lambda}_{iT}(\theta)) + [\bar{\lambda}'_{iT}(\theta)]^2 \ell_{it}^{\lambda\lambda}(\theta, \bar{\lambda}_{iT}(\theta)), \\
d_{\lambda} S_{it,T}(\theta, \bar{\lambda}_{iT}(\theta)) &= \ell_{it}^{\theta\lambda}(\theta, \bar{\lambda}_{iT}(\theta)) + \bar{\lambda}'_{iT}(\theta) \ell_{it}^{\lambda\lambda}(\theta, \bar{\lambda}_{iT}(\theta)),
\end{aligned}$$

where

$$\begin{aligned}
\bar{\lambda}'_{iT}(\theta) &= \frac{\mathbb{E}[\ell_{it}^{\lambda\theta}(\theta, \bar{\lambda}_{iT}(\theta))]}{\mathbb{E}[-\ell_{it}^{\lambda\lambda}(\theta, \bar{\lambda}_{iT}(\theta))]}, \\
\bar{\lambda}''_{iT}(\theta) &= \frac{\mathbb{E}[\ell_{it}^{\lambda\theta\theta}(\theta, \bar{\lambda}_{iT}(\theta))] + \{\mathbb{E}[\ell_{it}^{\lambda\theta\lambda}(\theta, \bar{\lambda}_{iT}(\theta))] + \mathbb{E}[\ell_{it}^{\lambda\lambda\lambda}(\theta, \bar{\lambda}_{iT}(\theta))]\} \bar{\lambda}'_{iT}(\theta)}{\mathbb{E}[-\ell_{it}^{\lambda\lambda}(\theta, \bar{\lambda}_{iT}(\theta))]},
\end{aligned}$$

and utilising the same arguments as for $S_{it,T}(\theta, \lambda_i)$ one can also show that

$$\sup_{\theta, \lambda_i} |d_{\lambda} S_{it,T}(\theta, \lambda_i)| \leq Kd(x_{it}), \quad \text{and} \quad \sup_{\theta, \lambda_i} |d_{\theta} S_{it,T}(\theta, \lambda_i)| \leq Kd(x_{it}). \tag{34}$$

As a direct result of (33) and (34), we have

$$\sup_{\theta, \lambda_i} \frac{1}{N} \sum_{i=1}^N |S_{it,T}(\theta, \lambda_i)| \leq Kd_{tN}, \tag{35}$$

$$\sup_{\theta, \lambda_i} \frac{1}{N} \sum_{i=1}^N |d_{\lambda} S_{it,T}(\theta, \lambda_i)| \leq Kd_{tN}, \quad \sup_{\theta, \lambda_i} \frac{1}{N} \sum_{i=1}^N |d_{\theta} S_{it,T}(\theta, \lambda_i)| \leq Kd_{tN}, \tag{36}$$

where $d_{tN} = N^{-1} \sum_{i=1}^N d(x_{it})$. Notice that by [Assumption 3.3](#) $\sup_{i,t} \mathbb{E}[(d(x_{it}))^2] < \infty$ and so, using Hölder's Inequality one obtains $\sup_{t,N} \mathbb{E}[K^2 d_{tN}^2] < \infty$ as $N, T \rightarrow \infty$. Finally, again by [Assumption 3.3](#), it can be shown that for some $\delta > 0$ we have $\mathbb{E}[|h_{t,NT}|^{4+\delta}] < \infty$. These last two results, along with (35) and (36) are counterparts of the conditions of assumption (ii) of NW. Notice also that, since $S_{it,T}(\theta, \lambda_i)$ is a measurable function of x_{it} , it is also α -mixing with the same size properties. Then, since the mixing coefficient size is uniform over i , $N^{-1} \sum_{i=1}^N S_{it,T}(\theta, \lambda_i)$, $t = 1, \dots, T$, is a sequence of α -mixing random variables for any $(\theta, \lambda_1, \dots, \lambda_N)$ with coefficients of size $(4 + 2r)/r$, for all N and T .

NW's proof is based on decomposing $|\bar{S}_{NT} - S_{NT}|$ as a sum of four terms, and proving that all these terms converge to zero. This decomposition is given in their equation (9). Since this is an algebraic result, the same decomposition also holds for $|\bar{S}_{NT} - S_{NT}|$, with appropriately modified notation. In NW's proof, the last three terms of this decomposition are shown to converge to zero by using results for mixing processes. Since $h_{t,NT}$ satisfies the same properties as NW's h_t , we can obtain convergence to zero for these three terms by the same approach. In particular, the second term converges to zero by Lemma 6.19 of [White \(2001\)](#), whereas the third term can be dealt with by using an appropriate covariance inequality (e.g. Corollary 6.17 of [White \(2001\)](#)) and the same reasoning as in NW. Finally, the fourth term can be shown to converge to zero by using the same technique as in the last part of the Proof of Lemma 6.6 of [Gallant and White \(1988\)](#). This leaves the first term, which is given by

$$\tau_1 = \left| \bar{S}_{NT} - \left[\frac{1}{T} \sum_{t=1}^T h_{t,NT}^2 + \frac{2}{T} \sum_{j=1}^m w(j, m) \sum_{t=j+1}^T h_{t,NT} h_{t-j,NT} \right] \right|. \tag{37}$$

As in NW, we will deal with this term by using asymptotic expansions. However, unlike in NW, $h_{t,NT}(\theta, \lambda_1, \dots, \lambda_N)$ is a function of both common and individual-specific parameters. As such, NW's expansions are not valid here. This is the main difference between the two proofs. Now,

$$\begin{aligned}
\frac{1}{N} \sum_{i=1}^N S_{it,T}(\hat{\theta}_*, \hat{\lambda}_i(\hat{\theta}_*)) &= \frac{1}{N} \sum_{i=1}^N S_{it,T}(\theta_0, \hat{\lambda}_i(\theta_0)) + \frac{1}{N} \sum_{i=1}^N d_{\theta} S_{it,T}(\tilde{\theta}, \hat{\lambda}_i(\tilde{\theta}))(\hat{\theta}_* - \theta_0) \\
&= \frac{1}{N} \sum_{i=1}^N S_{it,T}(\theta_0, \bar{\lambda}_{iT}(\theta_0)) + \mathcal{P}_{t,NT},
\end{aligned} \tag{38}$$

where

$$\mathcal{P}_{t,NT} = \frac{1}{N} \sum_{i=1}^N d_{\lambda} S_{it,T}(\theta_0, \bar{\lambda}_{iT}(\theta_0))[\hat{\lambda}_i(\theta_0) - \bar{\lambda}_{iT}(\theta_0)] + \frac{1}{N} \sum_{i=1}^N d_{\theta} S_{it,T}(\tilde{\theta}, \hat{\lambda}_i(\tilde{\theta}))(\hat{\theta}_* - \theta_0).$$

The first equality in (38) follows from expanding $S_{it,T}(\hat{\theta}_*, \hat{\lambda}_i(\hat{\theta}_*))$ about θ_0 , whereas the second line follows from expanding $S_{it,T}(\theta_0, \hat{\lambda}_i(\theta_0))$ about $\hat{\lambda}_{iT}(\theta_0)$. Here $\hat{\theta}$ and $\hat{\lambda}_{iT}(\cdot)$ are mean value terms. Next, substituting (38) into \bar{S}_{NT} , and then using the resulting expression in (37) we obtain $\mathcal{T}_1 = |\mathcal{P}'_{NT}|$ where

$$\begin{aligned} \mathcal{P}'_{NT} = & \frac{2}{T} \sum_{t=1}^T h_{t,NT} \mathcal{P}_{t,NT} + \frac{1}{T} \sum_{t=1}^T \mathcal{P}_{t,NT}^2 + \frac{2}{T} \sum_{j=1}^m w(j, m) \sum_{t=j+1}^T h_{t,NT} \mathcal{P}_{t-j,NT} \\ & + \frac{2}{T} \sum_{j=1}^m w(j, m) \sum_{t=j+1}^T \mathcal{P}_{t,NT} h_{t-j,NT} + \frac{2}{T} \sum_{j=1}^m w(j, m) \sum_{t=j+1}^T \mathcal{P}_{t,NT} \mathcal{P}_{t-j,NT}. \end{aligned}$$

Now, by (36), and also since $\hat{\theta}_* - \theta = O_p(T^{-1/2})$ and $\max_{1 \leq i \leq N} |\hat{\lambda}_i(\theta_0) - \bar{\lambda}_{iT}(\theta_0)| = o_p(T^{-2/5})$ by assumption, we have

$$|\mathcal{P}_{t,NT}| = \left(o_p\left(\frac{1}{T^{2/5}}\right) + O_p\left(\frac{1}{T^{1/2}}\right) \right) K d_{tN}.$$

Moreover, by (35), $|h_{t,NT}| \leq K d_{tN}$. Using these results and also the triangle inequality,

$$\begin{aligned} |\mathcal{P}'_{NT}| \leq & \left(o_p\left(\frac{1}{T^{2/5}}\right) + O_p\left(\frac{1}{T^{1/2}}\right) \right) \frac{K}{T} \sum_{t=1}^T d_{tN}^2 \\ & + \left(o_p\left(\frac{1}{T^{2/5}}\right) + O_p\left(\frac{1}{T^{1/2}}\right) \right) \frac{K}{T} \sum_{j=1}^m w(j, m) \sum_{t=j+1}^T 2 d_{tN} d_{t-j,N}, \end{aligned} \quad (39)$$

where the exact value of K varies from line to line. We note the following: Using the generalised Chebyshev's Inequality and the earlier result that $\sup_{t,N} \mathbb{E}[d_{tN}^2] < \infty$, we have $T^{-1} \sum_{t=1}^T d_{tN}^2 = O_p(1)$ as $N, T \rightarrow \infty$. Also, observe that $\sum_{t=j+1}^T 2 d_{tN} d_{t-j,N} \leq \sum_{t=j+1}^T (d_{tN}^2 + d_{t-j,N}^2)$ and $\sum_{j=1}^m w(j, m) \leq m$. Then, (39) becomes

$$\begin{aligned} |\mathcal{P}'_{NT}| \leq & \left(o_p\left(\frac{1}{T^{2/5}}\right) + O_p\left(\frac{1}{T^{1/2}}\right) \right) \sum_{j=1}^m w(j, m) \underbrace{\frac{K}{T} \sum_{t=j+1}^T (d_{tN}^2 + d_{t-j,N}^2)}_{KT^{-1} \sum_{t=1}^T 2 d_{tN}^2} + o_p(1) \\ \leq & o_p(1) + \left(o_p\left(\frac{1}{T^{2/5}}\right) + O_p\left(\frac{1}{T^{1/2}}\right) \right) m, \end{aligned}$$

which is $o_p(1)$ as $N, T \rightarrow \infty$, since $m = o(T^{1/4})$ by assumption. Hence, finally, $|\mathcal{T}_1| = o_p(1)$ as $N, T \rightarrow \infty$, and consequently $|\bar{S}_{NT} - S_{NT}| = o_p(1)$ as $N, T \rightarrow \infty$, as desired. ■

Appendix B. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.jeconom.2019.05.020>.

References

- Arellano, M., Bonhomme, S., 2009. Robust priors in nonlinear panel data models. *Econometrica* 77, 489–536.
- Arellano, M., Hahn, J., 2007. Understanding bias in nonlinear panel models: some recent developments. In: Blundell, R., Newey, W., Persson, T. (Eds.), *Advances in Economics and Econometrics: Theory and Applications*, Ninth World Congress - Vol. III. Cambridge University Press, pp. 381–409.
- Arellano, M., Hahn, J., 2016. A likelihood-based approximate solution to the incidental parameter problem in dynamic nonlinear models with multiple effects. *Glob. Econ. Rev.* 45, 251–274.
- Bai, J., 2009. Panel data models with interactive fixed effects. *Econometrica* 77, 1229–1279.
- Bailey, N., Kapetanios, G., Pesaran, M.H., 2016. Exponent of cross-sectional dependence: estimation and inference. *J. Appl. Econometrics* 31, 929–960.
- Barndorff-Nielsen, O.E., 1983. On a formula for the distribution of the maximum likelihood estimator. *Biometrika* 70, 343–365.
- Bertrand, M., Duflo, E., Mullainathan, S., 2004. How much should we trust differences-in-differences estimates. *Q. J. Econ.* 119, 249–275.
- Bester, C.A., Conley, T.G., Hansen, C.B., 2011. Inference with dependent data using cluster covariance estimators. *J. Econometrics* 165, 137–151.
- Bester, C.A., Hansen, C.B., 2009. A penalty function approach to bias reduction in nonlinear panel models with fixed effects. *J. Bus. Econom. Statist.* 27, 131–148.
- Cameron, A.C., Miller, D.L., 2011. Robust inference with clustered data. In: Ullah, A., Giles, D.E. (Eds.), *Handbook of Empirical Economics and Finance*. CRC Press, pp. 1–28.
- Carro, J.M., 2007. Estimating dynamic panel data discrete choice models with fixed effects. *J. Econometrics* 140, 503–528.
- Chandrasekhar, A., Lewis, R., 2016. Econometrics of sampled networks. Working paper.
- Conley, T.G., 1999. GMM estimation with cross sectional dependence. *J. Econometrics* 92, 1–45.
- Cox, D.R., Reid, N., 1987. Parameter orthogonality and approximate conditional inference (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* 49, 1–39.
- Cox, D.R., Reid, N., 2004. A note on pseudolikelihood constructed from marginal densities. *Biometrika* 91, 729–737.
- Dhaene, G., Jochmans, K., 2015. Split-panel jackknife estimation of fixed-effects models. *Rev. Econom. Stud.* 82, 991–1030.

- Driscoll, J.C., Kraay, A.C., 1998. Consistent covariance matrix estimation with spatially dependent panel data. *Rev. Econ. Stat.* 80, 549–560.
- Fernández-Val, I., 2009. Fixed effects estimation of structural parameters and marginal effects in panel probit models. *J. Econometrics* 150, 71–85.
- Fernández-Val, I., Weidner, M., 2016. Individual and time effects in nonlinear panel data models with large n . *J. Econometrics* 192, 291–312.
- Gallant, A.R., White, H., 1988. *A Unified Theory of Estimation and Inference for nonlinear Dynamic Models*. Basil Blackwell, Oxford.
- Goldsmith-Pinkham, P., Imbens, G.W., 2013. Social networks and the identification of peer effects. *J. Bus. Econom. Statist.* 31, 253–264.
- Gonçalves, S., 2011. The moving blocks bootstrap for panel linear regression models with individual fixed effects. *Econometric Theory* 27, 1048–1082.
- Hahn, J., Kuersteiner, G., 2002. Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and t are large. *Econometrica* 70, 1639–1657.
- Hahn, J., Kuersteiner, G., 2011. Bias reduction for dynamic nonlinear panel models with fixed effects. *Econometric Theory* 27, 1152–1191.
- Hahn, J., Newey, W.K., 2004. Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica* 72 (4), 1295–1319.
- Hansen, C.B., 2007. Asymptotic properties of a robust variance matrix estimator for panel data when t is large. *J. Econometrics* 141, 597–620.
- Ibragimov, R., Müller, U.K., 2010. T-statistic based correlation and heterogeneity robust inference. *J. Bus. Econom. Statist.* 28, 453–468.
- Jenish, N., Prucha, I.R., 2009. Central limit theorems and uniform laws of large numbers for arrays of random fields. *J. Econometrics* 150, 86–98.
- Kass, R.E., Tierney, L., Kadane, J.B., 1990. The validity of posterior expansions based on Laplace's method. In: Geisser, S., Hodges, J., Press, S., Zellner, A. (Eds.), *Essays in Honor of George Bernard*. Amsterdam: North Holland, pp. 473–488.
- Kelejian, H.H., Prucha, I.R., 2007. HAC estimation in a spatial framework. *J. Econometrics* 140, 131–154.
- Lancaster, T., 2000. The incidental parameter problem since 1948. *J. Econometrics* 95, 391–413.
- Lindsay, B.G., 1988. Composite likelihood methods. In: Prabhu, N.U. (Ed.), *Statistical Inference from Stochastic Processes*. American Mathematical Society, Providence, RI, pp. 221–239.
- Lütkepohl, H., 1996. *Handbook of Matrices*. Wiley, Chichester.
- McCullagh, P., 1987. *Tensor Methods in Statistics*. Chapman & Hall, London.
- Newey, W.K., West, K.D., 1987. A simple positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55, 703–708.
- Neyman, J., Scott, E.L., 1948. Consistent estimates based on partially consistent observations. *Econometrica* 16, 1–16.
- Pace, L., Salvan, A., 2006. Adjustments of the profile likelihood from a new perspective. *J. Statist. Plann. Inference* 136, 3554–3564.
- Phillips, P.C.B., Sul, D., 2007. Bias in dynamic panel estimation with fixed effects, incidental trends and cross section dependence. *J. Econometrics* 137, 162–188.
- Rilstone, P., Srivastava, V.K., Ullah, A., 1996. The second-order bias and mean squared error of nonlinear estimators. *J. Econometrics* 75, 369–395.
- Sartori, N., 2003. Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika* 90, 533–549.
- Severini, T.A., 1999. On the relationship between Bayesian and non-Bayesian elimination of nuisance parameters. *Statist. Sinica* 9, 713–724.
- Severini, T.A., 2000. *Likelihood Methods in Statistics*. Oxford University Press, New York.
- Severini, T.A., 2007. Integrated likelihood functions for non-Bayesian inference. *Biometrika* 94, 529–542.
- Severini, T.A., Wong, W.H., 1992. Profile likelihood and conditionally parametric models. *Ann. Statist.* 20, 1768–1802.
- White, H., 2001. *Asymptotic Theory for Econometricians*, second ed. Academic Press, Orlando.
- Wooldridge, J.M., 2003. Cluster-sample methods in applied econometrics. *Amer. Econ. Rev.* 93, 133–188.
- Woutersen, T., 2002. Robustness against incidental parameters. Working paper.