SPLIT-PANEL JACKKNIFE ESTIMATION OF FIXED-EFFECT MODELS

SUPPLEMENTARY APPENDIX

Geert Dhaene

University of Leuven, Department of Economics, Naamsestraat 69, B-3000 Leuven, Belgium
geert.dhaene@kuleuven.be

Koen Jochmans

Sciences Po, Department of Economics, 28 rue des Saints-Pères, 75007 Paris, France
koen.jochmans@sciencespo.fr

1. PROOFS OF THEOREMS IN THE MAIN TEXT

Proof. (Proof of Theorem 2.1.) Note that

$$\widetilde{\theta} = \frac{1}{m}\sum_{j=1}^{m} \widetilde{\theta}_{\mathcal{S}_j}, \qquad \widetilde{\theta}_{\mathcal{S}_j} \equiv \frac{g}{g-1}\widehat{\theta} - \frac{1}{g-1}\overline{\theta}_{\mathcal{S}_j},$$

where the $\overline{\theta}_{\mathcal{S}_j}$ are as in (2.2). Clearly, averaging over the equivalence class of $\mathcal{S}$ does not affect the asymptotic properties of $\widetilde{\theta}$. Thus, it suffices to consider the asymptotic behavior of $\widetilde{\theta}_{\mathcal{S}}$. First note that, for any two distinct $S, S' \in \mathcal{S}$, because S and S' are disjoint, $\sqrt{NT}(\widehat{\theta}_S - \theta_{|S|})$ and $\sqrt{NT}(\widehat{\theta}_{S'} - \theta_{|S'|})$ are jointly asymptotically normal as $N, T \to \infty$, with large N, T covariance equal to zero by Assumptions 2.1 and 2.2. Then, given that $\mathcal{S}$ is a partition of $\{1, 2, \dots, T\}$ and $\min_{S\in\mathcal{S}} |S|/T$ is bounded away from zero, it follows that, as $N, T \to \infty$,

$$\sqrt{NT}\begin{pmatrix} \widehat{\theta} - \theta_T \\ \overline{\theta}_{\mathcal{S}} - \text{plim}_{N\to\infty}\overline{\theta}_{\mathcal{S}} \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma^{-1} & \Sigma^{-1} \\ \Sigma^{-1} & \Sigma^{-1} \end{pmatrix}\right)$$

and, in turn, $\sqrt{NT}(\widetilde{\theta}_{\mathcal{S}} - \text{plim}_{N\to\infty}\widehat{\theta}_{\mathcal{S}}) \xrightarrow{d} \mathcal{N}(0, \Sigma^{-1})$. By the construction of $\widetilde{\theta}_{\mathcal{S}}$ and by Assumption 2.3,

$$\sqrt{NT}(\text{plim}_{N\to\infty}\widetilde{\theta}_{\mathcal{S}} - \theta_0) = \sqrt{NT}o(T^{-1}) \to 0$$

provided $N, T \to \infty$ with $N/T \to \rho$. Therefore, the bias is asymptotically negligible. This completes the proof. □

Proof. (Proof of Theorem 2.2.) Again, the averaging over the equivalence class of $\mathcal{S}$ can be ignored, so it suffices to show that the maximizer of

$$\dot{l}_{\mathcal{S}}(\theta) \equiv \frac{g}{g-1}\widehat{l}(\theta) - \frac{1}{g-1}\overline{l}_{\mathcal{S}}(\theta)$$

satisfies the conclusions of the theorem. Let $\overline{s}_{\mathcal{S}}(\theta) \equiv \nabla_\theta \overline{l}_{\mathcal{S}}(\theta)$ and $\dot{s}_{\mathcal{S}}(\theta) \equiv \nabla_\theta \dot{l}_{\mathcal{S}}(\theta)$. As in the proof of Theorem 2.1, Assumptions 2.1, 2.4, and 2.5 imply that, for all $\theta \in \mathcal{N}_0$, $\sqrt{NT}(\dot{s}_{\mathcal{S}}(\theta) - s_0(\theta)) \xrightarrow{d} \mathcal{N}(0, \Delta(\theta))$ as $N, T \to \infty$ with $N/T \to \rho$. Let $\dot{\theta}$ be the maximizer of $\dot{l}_{\mathcal{S}}(\theta)$ on Θ. Because $\dot{s}_{\mathcal{S}}(\dot{\theta}) = 0$ with probability approaching one, a Taylor expansion around θ_0 yields

$$\dot{\theta} - \theta_0 = -\dot{H}_{\mathcal{S}}(\theta_0)^{-1}\dot{s}_{\mathcal{S}}(\theta_0) + o_p(1/\sqrt{NT}),$$

where $\dot{H}_{\mathcal{S}}(\theta) \equiv \nabla_{\theta\theta'}\dot{l}_{\mathcal{S}}(\theta)$. As $\dot{H}_{\mathcal{S}}(\theta_0) \xrightarrow{p} -\Sigma$, it then follows that $\sqrt{NT}(\dot{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma^{-1})$ as $N, T \to \infty$ with $N/T \to \rho$. This completes the proof. □

Proof. (Proof of Theorem 4.1.) First consider the infeasible estimator

$$\mu^\dagger \equiv \widehat{\mu}(\theta_0) = \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \mu_{it}(\widehat{\alpha}_i(\theta_0),\theta_0).$$

A third-order expansion around $\alpha_i(\theta_0)$ gives

$$\mu^\dagger - \mu_* = \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \nabla_\alpha \mu_{it}(\alpha_i(\theta_0),\theta_0)\,(\widehat{\alpha}_i(\theta_0)-\alpha_i(\theta_0)) + \frac{\nabla_{\alpha\alpha}\mu_{it}(\alpha_i(\theta_0),\theta_0)}{2}\,(\widehat{\alpha}_i(\theta_0)-\alpha_i(\theta_0))^2 + O_p\left(\frac{1}{\sqrt{NT}}\right).$$

Using the influence-function representation in Assumption (4.1),

$$\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \nabla_\alpha \mu_{it}(\alpha_i(\theta_0),\theta_0)\,(\widehat{\alpha}_i(\theta_0)-\alpha_i(\theta_0)) = \frac{\frac{1}{N}\sum_{i=1}^N \mathbb{E}[\nabla_\alpha \mu_{it}(\alpha_i(\theta_0),\theta_0)]\,\beta_i}{T} + \frac{\frac{1}{N}\sum_{i=1}^N \sum_{j=-\infty}^{+\infty}\mathbb{E}[\nabla_\alpha \mu_{it}(\alpha_i(\theta_0),\theta_0)\,\psi_{it-j}]}{T} + O_p\left(\frac{1}{\sqrt{NT}}\right),$$

where the convergence can be deduced from Assumption (4.2). Also,

$$\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \frac{\nabla_{\alpha\alpha}\mu_{it}(\alpha_i(\theta_0),\theta_0)}{2}\,(\widehat{\alpha}_i(\theta_0)-\alpha_i(\theta_0))^2 = \frac{\frac{1}{N}\sum_{i=1}^N \mathbb{E}[\nabla_{\alpha\alpha}\mu_{it}(\alpha_i(\theta_0),\theta_0)]\sigma_i^2/2}{T} + O_p\left(\frac{1}{\sqrt{NT}}\right)$$

follows in a similar manner; the higher-order bias can be ignored because $T^{-2} = o_p(1/\sqrt{NT})$ under rectangular-array asymptotics. Hence,

$$\mu^\dagger - \mu_* = \frac{D}{T} + o\left(\frac{1}{T}\right) + O_p\left(\frac{1}{\sqrt{NT}}\right).$$

As $\sqrt{N}(\mu_* - \mu_0) \xrightarrow{d} \mathcal{N}(0, \mathrm{var}(\mathbb{E}[\mu_{it}(\theta_0,\alpha_{i0})]))$, we equally obtain $\sqrt{N}(\mu^\dagger - \mu_0) \xrightarrow{d} \mathcal{N}(0, \mathrm{var}(\mathbb{E}[\mu_{it}(\theta_0,\alpha_{i0})]))$ as $N,T\to\infty$ and $N/T\to\rho$. To obtain the same result for the feasible estimator $\widehat{\mu}$, note that another expansion around θ_0 further yields $\widehat{\mu} - \mu^\dagger = \overline{\mathbb{E}}[\nabla_\theta \mu(\alpha_i(\theta_0),\theta_0)]\,(\widehat{\theta}-\theta_0) + o_p(1)$, so that

$$\widehat{\mu} = \mu_* + \frac{D+E}{T} + o\left(\frac{1}{T}\right) + O_p\left(\frac{1}{\sqrt{NT}}\right),$$

from which $\sqrt{N}(\widehat{\mu}-\mu_*) = o_p(1)$ readily follows. The corresponding results for the jackknife estimator follow in analogous fashion as before. □

2. ASYMPTOTIC CHARACTERIZATION OF BIAS CORRECTION VIA THE JACKKNIFE

2.1. *First-order bias correction*

Because

$$\widetilde{\theta} = \frac{1}{m}\sum_{j=1}^m \widetilde{\theta}_{\mathcal{S}_j}, \qquad \widetilde{\theta}_{\mathcal{S}_j} \equiv \frac{g}{g-1}\widehat{\theta} - \frac{1}{g-1}\overline{\theta}_{\mathcal{S}_j},$$

and the set of cardinalities is the same for all members of the equivalence class of $\mathcal{S}$, the large N, fixed T bias of $\widetilde{\theta}$ is the same as that of $\widetilde{\theta}_{\mathcal{S}}$. Hence, it suffices to consider $\widetilde{\theta}_{\mathcal{S}}$. Recall that $\mathcal{S} = \{S_1, S_2, \ldots, S_g\}$ is a collection of $g \geq 2$ subpanels partitioning $\{1,2,\ldots,T\}$ in such a way that the sequence $\min_{S\in\mathcal{S}} |S|/T$ is bounded away from zero as T grows.

Theorem S.2.1. *Let Assumption 2.1 hold and assume that (2.5) is satisfied for some $k \geq 2$. Then*

$$\mathrm{plim}_{N\to\infty}\widetilde{\theta}_{\mathcal{S}} = \theta_0 + \frac{B_2'}{T^2} + \frac{B_3'}{T^3} + \ldots + \frac{B_k'}{T^k} + o\left(T^{-k}\right)$$

where

$$B'_j \equiv \frac{g - T^{j-1}\sum_{S\in\mathcal{S}}|S|^{1-j}}{g-1}B_j = O(1),$$

$\operatorname{sign}(B'_j) = -\operatorname{sign}(B_j)$, *and* $|B'_j| \geq |B_j|\sum_{m=1}^{j-1} g^m$.

PROOF. Since $|S|/T$ is bounded away from zero for all $S \in \mathcal{S}$,

$$\operatorname{plim}_{N\to\infty}\overline{\theta}_{\mathcal{S}} = \theta_0 + \sum_{j=1}^{k}\sum_{S\in\mathcal{S}}\frac{|S|^{1-j}}{T}B_j + o(T^{-k}) = \theta_0 + \frac{g}{T}B_1 + \sum_{j=2}^{k}\sum_{S\in\mathcal{S}}\frac{|S|^{1-j}}{T}B_j + o(T^{-k})$$

where, by convention, $\sum_{j=2}^{k}(\cdot) = 0$ if $k = 1$. The result regarding $\operatorname{plim}_{N\to\infty}\widehat{\theta}_{\mathcal{S}}$ follows easily, since $B'_j = O(1)$ for $j = 2,\ldots,k$ because $T/|S| = O(1)$ for all $S \in \mathcal{S}$. Since $T/|S| > 1$, we have $T^{j-1}\sum_{S\in\mathcal{S}}|S|^{1-j} > g$ for all $j \geq 2$, so $\operatorname{sign}(B'_j) = -\operatorname{sign}(B_j)$. To prove that $|B'_j| \geq |B_j|\sum_{m=1}^{j-1} g^m$, it suffices to show that, for $j \geq 2$, we have

$$T^{j-1}\sum_{S\in\mathcal{S}}|S|^{1-j} - g \geq (g-1)\sum_{m=1}^{j-1} g^m. \tag{S.2.1}$$

By a property of the harmonic mean, for $j \geq 2$,

$$T^{j-1}\sum_{S\in\mathcal{S}}|S|^{1-j} \geq T^{j-1}\sum_{S\in\mathcal{S}}\left(\frac{T}{g}\right)^{1-j} = g^j,$$

from which (S.2.1) follows. □

2.2. Higher-order bias correction

Let

$$b_j(G) \equiv (-1)^h g_1\ldots g_h \sum_{\substack{k_1,\ldots,k_h\geq 0\\ k_1+\ldots+k_h\leq j-h-1}} g_1^{k_1}\ldots g_h^{k_h}, \qquad j = 1,2,\ldots, \tag{S.2.2}$$

with the standard convention that empty sums and products are 0 and 1, respectively, so that $b_j(G) = 0$ for $j \leq h = |G|$, and $b_j(\varnothing) = 1$ for all $j \geq 1$.

THEOREM S.2.2. *Let Assumption 2.1 hold and assume that (2.5) is satisfied for some $k \geq h$. If $k = h$, then $\operatorname{plim}_{N\to\infty}\widehat{\theta}_{1/G} = \theta_0 + o\left(T^{-h}\right)$. If $k > h$, then*

$$\operatorname{plim}_{N\to\infty}\widehat{\theta}_{1/G} = \theta_0 + \frac{B'_{h+1}(G)}{T^{h+1}} + \ldots + \frac{B'_k(G)}{T^k} + o\left(T^{-k}\right)$$

where $B'_j(G) = b_j(G)B_j + O(T^{-1})$.

PROOF. For all $g \in G$,

$$\operatorname{plim}_{N\to\infty}\overline{\theta}_{1/g} = \theta_0 + \sum_{j=1}^{k}\sum_{S\in\mathcal{S}_g}\frac{|S|^{1-j}}{T}B_j + o(T^{-k}).$$

Hence

$$\begin{aligned}\operatorname{plim}_{N\to\infty}\widehat{\theta}_{1/G} &= \theta_0 + \sum_{j=1}^{k}\left(\left(1 + \sum_{g\in G}a_{1/g}\right)\frac{1}{T^j} - \sum_{g\in G}a_{1/g}\sum_{S\in\mathcal{S}_g}\frac{|S|^{1-j}}{T}\right)B_j + o(T^{-k})\\ &= \theta_0 + \sum_{j=1}^{k}\frac{c_j(G)B_j}{T^j} + o(T^{-k}),\end{aligned}$$

where

$$\begin{aligned} c_j(G) &\equiv 1 + \sum_{g\in G} a_{1/g} \left(1 - \sum_{S\in\mathcal{S}_g} T^{j-1}|S|^{1-j}\right) \\ &= (1-\iota' A^{-1}\iota)^{-1} - \sum_{g\in G} a_{1/g} \sum_{S\in\mathcal{S}_g} T^{j-1}|S|^{1-j} \\ &= (1-\iota' A^{-1}\iota)^{-1} - \sum_{r=1}^{l} a_{1/g_r} \sum_{S\in\mathcal{S}_{g_r}} T^{j-1}|S|^{1-j} \\ &= (1-\iota' A^{-1}\iota)^{-1} \left(1 - \sum_{r=1}^{l}\left(\sum_{s=1}^{l} A^{rs}\right) \sum_{S\in\mathcal{S}_{g_r}} T^{j-1}|S|^{1-j}\right), \end{aligned} \tag{S.2.3}$$

and A^{rs} is the (r,s)th element of A^{-1}. For $j\leq l$,

$$\begin{aligned} c_j(G) &= (1-\iota' A^{-1}\iota)^{-1} \left(1 - \sum_{r=1}^{l}\left(\sum_{s=1}^{l} A^{rs}\right) A_{jr}\right) \\ &= (1-\iota' A^{-1}\iota)^{-1} \left(1 - \sum_{s=1}^{l}\sum_{r=1}^{l} A_{jr}A^{rs}\right) = 0. \end{aligned}$$

This proves that $\text{plim}_{N\to\infty}\widehat{\theta}_{1/G} = \theta_0 + o\left(T^{-l}\right)$ if $k=l$. Now consider the case $k>l$. We need to show that $c_j(G) = b_j(G) + O(T^{-1})$ for $l<j\leq k$. For all $g\in G$ and all $S\in\mathcal{S}_g$, $T|S|^{-1} = g + O(T^{-1})$, and, for $r=1,\ldots,k$, $\sum_{S\in\mathcal{S}_g} T^{r-1}|S|^{1-r} = g^r + O(T^{-1})$. Hence $A = \mathbf{A} + O(T^{-1})$, where $\mathbf{A}$ is the $l\times l$ matrix with elements $\mathbf{A}_{rs} = g_s^r$. Let $\pi_j \equiv (g_1^j,\ldots,g_l^j)'$. From (S.2.3), for $l<j\leq k$,

$$\begin{aligned} c_j(G) &= (1-\iota'\mathbf{A}^{-1}\iota)^{-1}\left(1 - \sum_{r=1}^{l}\left(\sum_{s=1}^{l}\mathbf{A}^{rs}\right) g_r^j\right) + O(T^{-1}) \\ &= (1-\iota'\mathbf{A}^{-1}\iota)^{-1}\left(1 - \pi_j'\mathbf{A}^{-1}\iota\right) + O(T^{-1}) \\ &= \frac{|\mathbf{A}|^{-1}\begin{vmatrix} \mathbf{A} & \iota \\ \pi_j' & 1 \end{vmatrix}}{|\mathbf{A}|^{-1}\begin{vmatrix} \mathbf{A} & \iota \\ \iota' & 1 \end{vmatrix}} + O(T^{-1}) = (-1)^l\frac{|V_j|}{|V|} + O(T^{-1}), \end{aligned}$$

where ι is an $l\times 1$ vector of ones and

$$|V| = \begin{vmatrix} 1 & \iota' \\ \iota & \mathbf{A}' \end{vmatrix}, \qquad V_j = \begin{vmatrix} \iota' & 1 \\ \mathbf{A}' & \pi_j \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 \\ 1 & \iota' & 1 \\ \iota & \mathbf{A}' & \pi_j \end{vmatrix}.$$

$|V|$ is a Vandermonde determinant given by

$$|V| = \prod_{0\leq p<q\leq l}(g_q - g_p), \qquad g_0\equiv 1.$$

Noting that the first row of V_{l+1} is $(0^0, 0^1, \ldots, 0^{l+1})$, $|V_{l+1}|$ is also a Vandermonde determinant, given by

$$|V_{l+1}| = \prod_{-1\leq p<q\leq l}(g_q - g_p) = |V|\prod_{1\leq q\leq l} g_q, \qquad g_{-1}\equiv 0.$$

For $j>l+1$, by the Jacobi-Trudi identity (see, e.g., Littlewood 1958, pp. 88), $|V_j|$ can be written as the product of $|V_{l+1}|$ and a homogeneous product sum of $g_{-1}, g_0, \ldots, g_l$,

$$|V_j| = |V_{l+1}| \sum_{\substack{k_{-1},k_0,\ldots,k_l\geq 0 \\ k_{-1}+k_0+\ldots+k_l = j-l-1}} g_{-1}^{k_{-1}} g_0^{k_0}\ldots g_l^{k_l} = |V_{l+1}| \sum_{\substack{k_1,\ldots,k_l\geq 0 \\ k_1+\ldots+k_l\leq j-l-1}} g_1^{k_1}\ldots g_l^{k_l},$$

which also holds for $j = l + 1$. On collecting results, $c_j(G) = b_j(G) + O(T^{-1})$ for $l < j \leq k$. □

Table S.1 gives the first few $b_j(G)$ for selected G. Together with Theorem S.2.2, the $b_j(G)$ indicate how the jackknife transforms the higher-order bias terms.

Table S.1. Higher-order bias inflation factors

G	$b_1(G)$	$b_2(G)$	$b_3(G)$	$b_4(G)$	$b_5(G)$
$\varnothing$	1	1	1	1	1
$\{2\}$	0	−2	−6	−14	−20
$\{2,3\}$	0	0	6	36	150
$\{2,3,4\}$	0	0	0	−24	−240

3. BIAS CORRECTION WITH OVERLAPPING SUBPANELS

We next consider h-order bias correction ($h \geq 1$), with $o \geq 0$ collections of two overlapping subpanels and $h - o \geq 0$ collections of non-overlapping subpanels. Let $G \equiv \{g_1, \ldots, g_h\}$, where $1 < g_1 < \ldots < g_o < 2 \leq g_{o+1} < \ldots < g_h$ and $g_{o+1}, \ldots, g_h$ are integers. We need T large enough so that $T \geq g_h T_{\min}$ and $\lceil T/g \rceil \neq \lceil T/g' \rceil$ for all distinct $g, g' \in G$. For each $g \in G$, let $\mathcal{S}_g$ be a collection of subpanels covering $\{1, \ldots, T\}$ such that (i) if $g < 2$, then $\mathcal{S}_g$ consists of two subpanels, each with $\lceil T/g \rceil$ elements; (ii) if $g \geq 2$, then $\mathcal{S}_g$ is a collection of g non-overlapping subpanels forming an almost equal partition of $\{1, \ldots, T\}$. With $\{\mathcal{S}_{gj}; j = 1, 2, \ldots, m_g\}$ the equivalence class of $\mathcal{S}_g$, define the split-panel jackknife estimator

$$\widetilde{\theta}_{1/G} \equiv \left(1 + \sum_{g \in G} a_{1/g}\right) \widehat{\theta} - \sum_{g \in G} a_{1/g} \overline{\theta}_{1/g}, \qquad \overline{\theta}_{1/g} \equiv \frac{1}{m_g} \sum_{j=1}^{m_g} \overline{\theta}_{\mathcal{S}_{gj}}, \qquad \overline{\theta}_{\mathcal{S}_{gj}} \equiv \sum_{S \in \mathcal{S}_g} \frac{|S|}{\sum_{S \in \mathcal{S}_g} |S|} \widehat{\theta}_S, \tag{S.3.1}$$

where a_{1/g_r} is the rth element of $(1 - \iota' A^{-1} \iota)^{-1} A^{-1} \iota$ and A is the $h \times h$ matrix with elements

$$[A]_{r,s} \equiv \frac{\sum_{S \in \mathcal{S}_{g_s}} (T/|S|)^{r-1}}{\sum_{S \in \mathcal{S}_{g_s}} |S|/T}, \qquad r, s = 1, \ldots, h. \tag{S.3.2}$$

Note that, for $o = 0$, $\widehat{\theta}_{1/G}$ reduces to the estimator given in (5.3). The coefficients $a_{1/g}$, again, solve a linear equation system so that the bias of $\widehat{\theta}_{1/G}$ is $o(T^{-h})$.

Let $b_j(G)$ be as in (S.2.2). Let

$$d_T(G) \equiv 1 + (1 - \iota' A^{-1} \iota)^{-2} \iota' A'^{-1} \Gamma A^{-1} \iota,$$

where Γ is the symmetric $h \times h$ matrix whose (r, s)th element, for $r \leq s$, is

$$\Gamma_{rs} \equiv \begin{cases} \frac{1}{2} (A_{1r} - 1)(2 - A_{1s}) & \text{if } s \leq o, \\ 0 & \text{otherwise.} \end{cases}$$

The following theorem shows that $d_T(G)$ is the variance inflation factor due to the subpanel overlap.

Theorem S.3.1. *Let Assumptions 2.1, 2.2, and 2.3 hold. Assume that (2.5) holds for some $k \geq 2$. If $k = h$, then* $\text{plim}_{N \to \infty} \widehat{\theta}_{1/G} = \theta_0 + o\left(T^{-h}\right)$. *If $k > h$, then*

$$\text{plim}_{N \to \infty} \widehat{\theta}_{1/G} = \theta_0 + \frac{B'_{h+1}(G)}{T^{h+1}} + \ldots + \frac{B'_k(G)}{T^k} + o\left(T^{-k}\right)$$

where $B_j'(G) = b_j(G)B_j + O(T^{-1})$. *Further,*

$$\sqrt{\frac{NT}{d_T(G)}}(\widehat{\theta}_{1/G} - \theta_0) \overset{d}{\to} \mathcal{N}(0, \Sigma^{-1}) \qquad \text{as } N, T \to \infty \text{ and } N/T \to \rho,$$

and $d(G) \equiv \lim_{T\to\infty} d_T(G) \geq 1$, *with equality if and only if* $o = 0$.

Proof. The first part is proved along the same lines as in Theorem S.2.2. We have

$$\text{plim}_{N\to\infty}\widehat{\theta}_{1/G} = \theta_0 + \sum_{j=1}^{k} \frac{c_j(G)B_j}{T^j} + o(T^{-k}),$$

where now

$$\begin{aligned} c_j(G) &\equiv 1 + \sum_{g\in G} a_{1/g}\left(1 - \sum_{S\in\mathcal{S}_g} \frac{T^j|S|^{1-j}}{\sum_{S\in\mathcal{S}_g}|S|}\right) \\ &= (1 - \iota' A^{-1}\iota)^{-1}\left(1 - \sum_{r=1}^{l}\left(\sum_{s=1}^{l} A^{rs}\right)\sum_{S\in\mathcal{S}_{g_r}} \frac{T^j|S|^{1-j}}{\sum_{S\in\mathcal{S}_{g_r}}|S|}\right). \end{aligned}$$

For $j \leq l$, $c_j(G) = 0$. Consider the case $k > l$. For all $g \in G$ and $r = 1, \ldots, k$,

$$\begin{aligned} \sum_{S\in\mathcal{S}_g} \frac{T^r|S|^{1-r}}{\sum_{S\in\mathcal{S}_g}|S|} &= \frac{T}{\sum_{S\in\mathcal{S}_g}|S|}\sum_{S\in\mathcal{S}_g} T^{r-1}|S|^{1-r} = \frac{g}{\sum_{S\in\mathcal{S}_g} 1} g^{r-1}\sum_{S\in\mathcal{S}_g} 1 + O(T^{-1}) \\ &= g^r + O(T^{-1}). \end{aligned}$$

Hence, $A = \mathbf{A} + O(T^{-1})$ and, for $l < j \leq k$,

$$c_j(G) = (1 - \iota'\mathbf{A}^{-1}\iota)^{-1}\left(1 - \pi_j'\mathbf{A}^{-1}\iota\right) + O(T^{-1}),$$

where $\pi_j \equiv (g_1^j, \ldots, g_l^j)'$. By the proof of Theorem 2, $c_j(G) = b_j(G) + O(T^{-1})$ for $l < j \leq k$, thus completing the proof of the first part. We now derive the asymptotic distribution of $\widehat{\theta}_{1/G}$. For any pair of subpanels S and S' such that, as $T \to \infty$, $T^{-1}|S| \to s > 0$, $T^{-1}|S'| \to s' > 0$, and $T^{-1}|S \cap S'| \to s_\cap \geq 0$, we have

$$\text{Avar}\begin{pmatrix} \widehat{\theta}_S \\ \widehat{\theta}_{S'} \end{pmatrix} = \begin{pmatrix} 1/s & s_\cap/(ss') \\ s_\cap/(ss') & 1/s' \end{pmatrix} \otimes \Sigma^{-1}, \tag{S.3.3}$$

where $\text{Avar}(\cdot)$ denotes the large N, T variance. Now consider $\overline{\theta}_{1/g} = \frac{1}{2}(\widehat{\theta}_{S_1} + \widehat{\theta}_{S_2})$ and $\overline{\theta}_{1/g'} = \frac{1}{2}(\widehat{\theta}_{S_1'} + \widehat{\theta}_{S_2'})$, where $1 < g < g' < 2$ and $1 \in S_1 \cap S_1'$. Then $T^{-1}|S_1| = T^{-1}|S_2| \to 1/g$, $T^{-1}|S_1'| = T^{-1}|S_2'| \to 1/g'$, $T^{-1}|S_1 \cap S_2| \to (2-g)/g$, $T^{-1}|S_1' \cap S_2'| \to (2-g')/g'$, $T^{-1}|S_1 \cap S_1'| = T^{-1}|S_2 \cap S_2'| \to 1/g'$, and $T^{-1}|S_1 \cap S_2'| = T^{-1}|S_2 \cap S_1'| \to (g + g' - gg')/(gg')$. Application of (S.3.3) gives

$$\text{Avar}\begin{pmatrix} \widehat{\theta}_{S_1} \\ \widehat{\theta}_{S_2} \\ \widehat{\theta}_{S_1'} \\ \widehat{\theta}_{S_2'} \end{pmatrix} = \begin{pmatrix} g & g(2-g) & g & g+g'-gg' \\ g(2-g) & g & g+g'-gg' & g \\ g & g+g'-gg' & g' & g'(2-g') \\ g+g'-gg' & g & g'(2-g') & g' \end{pmatrix} \otimes \Sigma^{-1},$$

and so

$$\text{Avar}\begin{pmatrix} \overline{\theta}_{1/g} \\ \overline{\theta}_{1/g'} \end{pmatrix} = \frac{1}{2}\begin{pmatrix} g(3-g) & 2g+g'-gg' \\ 2g+g'-gg' & g'(3-g') \end{pmatrix} \otimes \Sigma^{-1}.$$

Let $\overline{\theta}_{1/G} \equiv (\overline{\theta}_{1/g_1}, \ldots, \overline{\theta}_{1/g_l})$. Then $\text{Avar}(\text{vec}\,\overline{\theta}_{1/G}) = V \otimes \Sigma^{-1}$, where $\text{vec}(\cdot)$ is the stack operator and V is the symmetric $l \times l$ matrix whose $(r, s)^{\text{th}}$ element, for $r \leq s$, is

$$V_{rs} \equiv \begin{cases} g_r + \frac{1}{2}(g_s - g_r g_s) & \text{if } s \leq o, \\ 1 & \text{otherwise.} \end{cases}$$

Therefore, $\widehat{\theta}_{1/G} = (1 - \iota' A^{-1}\iota)^{-1}(\widehat{\theta} - \overline{\theta}_{1/G} A^{-1}\iota)$ is asymptotically normally distributed, centered at θ_0, and has large N, T variance

$$\begin{aligned}\text{Avar}(\widehat{\theta}_{1/G}) &= (1 - \iota'\mathbf{A}^{-1}\iota)^{-2}(1 - 2\iota'\mathbf{A}^{-1}\iota + \iota'\mathbf{A}'^{-1}V\mathbf{A}^{-1}\iota)\Sigma^{-1} \\ &= \left(1 + \frac{\iota'\mathbf{A}'^{-1}\left(V - \iota\iota'\right)\mathbf{A}^{-1}\iota}{(1 - \iota'\mathbf{A}^{-1}\iota)^2}\right)\Sigma^{-1} = d(G)\Sigma^{-1},\end{aligned}$$

since $V - \iota\iota' = \Gamma$. The proof is completed by showing that, if $o \geq 1$, the leading $o \times o$ submatrix of Γ is positive definite. Let L_o be 2 times this submatrix, so that

$$L_o = \begin{pmatrix} L_{o-1} & \lambda_{o-1} \\ \lambda'_{o-1} & \lambda_{oo} \end{pmatrix},$$

where

$$\lambda_{o-1} \equiv \begin{pmatrix} g_1 - 1 \\ \vdots \\ g_{o-1} - 1 \end{pmatrix}(2 - g_o), \qquad \lambda_{oo} \equiv (g_o - 1)(2 - g_o).$$

The (r, s)-th element of L_{o-1}^{-1}, for $r \leq s$, is

$$L_{o-1}^{rs} = \begin{cases} \frac{g_{r+1} - g_{r-1}}{(g_r - g_{r-1})(g_{r+1} - g_r)} & \text{if } r = s < o - 1, \\ \frac{2 - g_{o-2}}{(g_{o-1} - g_{o-2})(2 - g_{o-1})} & \text{if } r = s = o - 1, \\ -\frac{1}{g_{r+1} - g_r} & \text{if } r = s - 1, \\ 0 & \text{otherwise,} \end{cases}$$

where $g_0 \equiv 1$. Hence

$$\lambda'_{o-1} L_{o-1}^{-1} \lambda_{o-1} = (2 - g_o)^2 \left(\sum_{r=1}^{o-2} h_r + \frac{(g_{o-1} - 1)^2 (2 - g_{o-2})}{(g_{o-1} - g_{o-2})(2 - g_{o-1})} \right),$$

where

$$\begin{aligned} h_r &= \frac{(g_r - 1)^2 (g_{r+1} - g_{r-1})}{(g_r - g_{r-1})(g_{r+1} - g_r)} - 2\frac{(g_r - 1)(g_{r+1} - 1)}{g_{r+1} - g_r} \\ &= (g_r - 1)\left(\frac{g_{r-1} - 1}{g_r - g_{r-1}} - \frac{g_{r+1} - 1}{g_{r+1} - g_r} \right). \end{aligned}$$

After some algebra, $\sum_{r=1}^{o-2} h_r = -\frac{(g_{o-1} - 1)(g_{o-2} - 1)}{g_{o-1} - g_{o-2}}$, and so

$$\lambda_{oo} - \lambda'_{o-1} L_{o-1}^{-1} \lambda_{o-1} = \frac{(g_o - g_{o-1})(2 - g_o)}{2 - g_{o-1}}.$$

The determinant of L_o is

$$|L_o| = |L_{o-1}| \left(\lambda_{oo} - \lambda'_{o-1} L_{o-1}^{-1} \lambda_{o-1} \right) = (2 - g_o) \prod_{r=1}^{o} (g_r - g_{r-1}),$$

by induction. Clearly, $0 < |L_o| < |L_{o-1}| < \ldots < |L_1| < 1$. All leading submatrices of L_o have a positive determinant, so L_o is positive definite. □

Overlapping subpanels allow $|b_j(G)|$ to be much smaller than is possible with collections of non-overlapping subpanels because $|b_j(G)|$ increases rapidly in all $g \in G$. For the same reason, the optimal choice of $g_{o+1}, \ldots, g_h$, from the perspective of minimizing the higher-order bias terms, is $2, \ldots, h - o + 1$. With overlapping subpanels, however, the large N, T variance inflation factor, $d_T(G)$, increases rapidly with both the number of collections of overlapping subpanels, o, and the number of collections of non-overlapping subpanels, $h-o$. So, in practice, one would hardly ever consider using more than one collection of overlapping subpanels in combination with collections of non-overlapping subpanels. More generally, we would only recommend the split-panel jackknife estimator with overlapping subpanels in applications where N is very large and there is a great need for bias reduction, for example, when T is very small.

Intuitively, subpanel overlap causes large N,T variance inflation because the time periods, t, receive unequal weights in those $\overline{\theta}_{1/g}$ where $1<g<2$. In principle, it is possible to prevent variance inflation by adding to $\overline{\theta}_{1/g}$ a term that has zero probability limit and equalizes those weights. As an example, take $g=3/2$ and suppose T is a multiple of 3 and $T\geq 3T_{\min}$. Then

$$\overline{\theta}_{2/3}=\frac{1}{2}(\widehat{\theta}_{1:2}+\widehat{\theta}_{2:3}),$$

where $\widehat{\theta}_{1:2}$ and $\widehat{\theta}_{2:3}$ use the first two-thirds and the last two-thirds of the time periods, respectively. Now consider

$$\widetilde{\theta}_{2/3}\equiv\frac{1}{2}(\widehat{\theta}_{1:2}+\widehat{\theta}_{2:3})+\frac{1}{12}(\widehat{\theta}_{1:1}-2\widehat{\theta}_{2:2}+\widehat{\theta}_{3:3}),$$

where each t receives a weight $1/T$ and $\text{plim}_{N\to\infty}\widetilde{\theta}_{2/3}=\text{plim}_{N\to\infty}\overline{\theta}_{2/3}$ because the second term of $\widetilde{\theta}_{2/3}$ has zero probability limit. Hence, replacing $\overline{\theta}_{2/3}$ with $\widetilde{\theta}_{2/3}$ in $\widehat{\theta}_{1/G}$, with unchanged weights $a_{1/g}$, $g\in G$, will leave the asymptotic bias unaffected but will reduce the large N,T variance. It is possible, for any $T\geq 2T_{\min}$ and any g between 1 and 2 that divides T, to find $\widetilde{\theta}_{1/g}$, similar to $\widetilde{\theta}_{2/3}$, such that each t receives a weight $1/T$ and $\text{plim}_{N\to\infty}\widetilde{\theta}_{1/g}=\text{plim}_{N\to\infty}\overline{\theta}_{1/g}$. A drawback, however, is that the weights associated with certain subpanel estimators in the zero plim term may become large, especially when g is close to 1, similar to the weights of the delete-one estimates in the ordinary jackknife. In exploratory simulations with small T, we found that this may substantially increase the variance, so we have not pursued this idea further.

4. DETAILS FOR THE GAUSSIAN AUTOREGRESSION

The model is

$$y_{it}=\alpha_{i0}+\gamma_0 y_{it-1}+\varepsilon_{it},\qquad \varepsilon_{it}\sim\mathcal{N}(0,\sigma^2),\qquad |\gamma_0|<1,$$

and the initial observations are y_{i0}. The maximum-likelihood estimator of γ_0, conditional on the y_{i0}, is the within-group estimator. The large N, fixed-T inconsistency of this estimator has been derived by Nickell (1981) under stationarity and by Bun and Carree (2005) and Dhaene and Jochmans (2013) for arbitrary initial observations.

First assume that the data are stationary. Re-arranging Equation (18) in Nickell (1981) allows writing the inconsistency as

$$\gamma_T-\gamma_0=-(1+\gamma_0)\frac{A}{1-B},\qquad A\equiv\frac{1}{T-1}\left(1-\frac{1-\gamma_0^T}{T\,(1-\gamma_0)}\right),\qquad B\equiv 2rA,\qquad r\equiv\frac{\gamma_0}{1-\gamma_0}.$$

Let k be any positive integer. Since $\gamma_0^T=o(T^{-k})$ as $T\to\infty$, we have

$$A=\frac{1}{T}-r\left(\frac{1}{T^2}+\frac{1}{T^3}+\dots\right)+o(T^{-k})$$

for any k. Now it readily follows that the expansion $\gamma_T-\gamma_0=\sum_{j=1}^k B_j/T^j+O(T^{-k-1})$ holds for any k, with the first few terms given by

$$\gamma_T-\gamma_0=-\frac{1+\gamma_0}{T}-\frac{r\,(1+\gamma_0)}{T^2}+\frac{r\,(1+\gamma_0)}{T^3}+\frac{\left(r+4r^2+2r^3\right)(1+\gamma_0)}{T^4}+O(T^{-5}).\tag{S.4.1}$$

Although for every fixed $T \geq 2$, $\sum_{j=1}^{\infty} B_j/T^j$ is a convergent series, $\gamma_T - \gamma_0 \neq \sum_{j=1}^{\infty} B_j/T^j$ due to the term γ_0^T, which affects A and, therefore, γ_T but not B_j. Note, also, that the expansion extends to the case $\gamma_0 = 1$, because

$$\lim_{\gamma_0 \uparrow 1} (\gamma_T - \gamma_0) = -\frac{3}{T+1} = -\frac{3}{T} + \frac{3}{T^2} - \frac{3}{T^3} + \dots,$$

which justifies the jackknife as a bias-reduction method also in the case of a unit root. As a numerical illustration of how the jackknife affects the non-eliminated higher-order bias terms, consider the case $\gamma_0 = .5$. From (S.4.1) and Table S.1, we obtain the expansions

$$\begin{aligned}
\text{plim}_{N\to\infty} \widehat{\gamma} &= .5 - \frac{1.5}{T} - \frac{1.5}{T^2} + \frac{1.5}{T^3} + \frac{10.5}{T^4} + O(T^{-5}), \\
\text{plim}_{N\to\infty} \widehat{\gamma}_{1/2} &= .5 + \frac{3}{T^2} - \frac{9}{T^3} - \frac{147}{T^4} + O(T^{-5}), \\
\text{plim}_{N\to\infty} \widehat{\gamma}_{1/\{2,3\}} &= .5 + \frac{9}{T^3} + \frac{378}{T^4} + O(T^{-5}), \\
\text{plim}_{N\to\infty} \widehat{\gamma}_{1/\{2,3,4\}} &= .5 - \frac{252}{T^4} + O(T^{-5}),
\end{aligned}$$

assuming that T increases in multiples of 12. The expansions show that higher-order versions of the jackknife require a larger T before the leading non-eliminated bias term becomes dominant.

Now consider arbitrary initial observations, with (α_{i0}, y_{i0}) drawn from a distribution $\mathcal{G}$. It is useful to introduce the quantity

$$\psi^2 \equiv \overline{\mathbb{E}}\left[\frac{(y_{i0} - \mu_i)^2}{\varsigma^2}\right], \qquad \mu_i \equiv \frac{\alpha_{i0}}{1-\gamma_0}, \qquad \varsigma^2 \equiv \frac{\sigma_0^2}{1-\gamma_0^2},$$

which is a measure of outlyingness of the initial observations with respect to their respective stationary distributions. When the data are stationary, we have $\psi^2 = 1$. Using results of Dhaene and Jochmans (2013) we can write the inconsistency of $\widehat{\gamma}$ as

$$\gamma_T - \gamma_0 = -(1+\gamma_0)\frac{A}{1-D}, \qquad D \equiv B + (1-\psi^2)C, \qquad C \equiv \frac{1}{T-1}\left(\frac{1-\gamma_0^{2T}}{1-\gamma_0^2} - \frac{1}{T}\left(\frac{1-\gamma_0^T}{1-\gamma_0}\right)^2\right),$$

with A and B as above. As $T \to \infty$,

$$C = \frac{A}{1-\gamma_0} - \frac{r}{1+\gamma_0}\left(\frac{1}{T} + \frac{1}{T^2} + \dots\right) + o(T^{-k})$$

for any k. Therefore, $\gamma_T - \gamma_0$ can again be expanded in powers of $1/T$ to any order. The first two terms are given by

$$\gamma_T - \gamma_0 = -\frac{1+\gamma_0}{T} - \frac{r(1+\gamma_0)}{T^2} + \frac{(\psi^2-1)/(1-\gamma_0)}{T^2} + O(T^{-3}).$$

In the stationary case, the expansion reduces to (S.4.1). The expression further shows that (i) the first-order bias is independent of the distribution of the initial observations (see also Hahn and Kuersteiner 2002); and (ii) deviations from stationarity do show up in the second-order bias.

We now calculate the bias of the profile log-likelihood for arbitrary initial observations. The joint log-likelihood for all parameters is

$$l(\gamma, \sigma^2, \alpha_1, \dots, \alpha_N) = -\frac{1}{2}\log\sigma^2 - \frac{1}{2NT\sigma^2}\sum_{i=1}^{N}(y_i - \gamma y_{i-} - \alpha_i\iota)'(y_i - \gamma y_{i-} - \alpha_i\iota),$$

where $y_i \equiv (y_{i1}, \dots, y_{iT})'$, $y_{i-} \equiv (y_{i0}, \dots, y_{iT-1})'$, ι is a $T \times 1$ vector of ones, and additive constants are omitted throughout. Profiling out $\alpha_1, \ldots, \alpha_N$ and σ^2 gives $\widehat{\alpha}_i(\gamma) = (y_i - \gamma y_{i-})'\iota/T$ and

$$\widehat{l}\,(\gamma) = -\frac{1}{2}\log\left(\frac{1}{NT}\sum_{i=1}^{N}(y_i - \gamma y_{i-})' M (y_i - \gamma y_{i-})\right),$$

where $M = I_T - \frac{1}{T}\iota\iota'$. Using results in Dhaene and Jochmans (2013), as $N \to \infty$, $\widehat{l}\,(\gamma)$ converges in probability to

$$\begin{aligned} l_T(\gamma) &= -\frac{1}{2}\log\left(\frac{1}{T}\mathbb{E}(y_i - \gamma y_{i-})' M (y_i - \gamma y_{i-})\right) \\ &= -\frac{1}{2}\log\left(1 + \frac{(\gamma-\gamma_0)^2}{1-\gamma_0^2} + R_T\right), \qquad R_T \equiv (\gamma-\gamma_0)\frac{2A}{1-\gamma_0} + (\gamma-\gamma_0)^2\frac{-D}{1-\gamma_0^2}. \end{aligned}$$

If instead of $\widehat{\alpha}_i(\gamma)$ we use

$$\alpha_i(\gamma) = \text{plim}_{T\to\infty}\,(y_i - \gamma y_{i-})'\,\iota/T = (1-\gamma)\frac{\alpha_{i0}}{1-\gamma_0},$$

we obtain an infeasible log-likelihood whose probability limit, as $N \to \infty$, is found as

$$\begin{aligned} l_0(\gamma) &= -\frac{1}{2}\log\left(\frac{1}{T}\mathbb{E}(y_i - \gamma y_{i-} - \alpha_i(\gamma)\iota)'(y_i - \gamma y_{i-} - \alpha_i(\gamma))\right) \\ &= -\frac{1}{2}\log\left(1 + \frac{(\gamma-\gamma_0)^2}{1-\gamma_0^2} + R_0\right), \qquad R_0 \equiv \frac{(\gamma-\gamma_0)^2}{T(1-\gamma_0^2)}\left(\frac{1-\gamma_0^{2T}}{1-\gamma_0^2}\right)(\psi^2-1), \end{aligned}$$

after a tedious but straightforward calculation. Note that γ_0 maximizes $l_0(\gamma)$ for every T and ψ^2. The asymptotic bias of $\widehat{l}\,(\gamma)$ is

$$l_T(\gamma) - l_0(\gamma) = -\frac{1}{2}\log(1+R) = -\frac{1}{2}\left(R - \frac{1}{2}R^2\right) + O\left(T^{-3}\right)$$

where

$$R = \frac{R_T - R_0}{1 + \frac{(\gamma-\gamma_0)^2}{1-\gamma_0^2} + R_0} = O\left(T^{-1}\right),$$

given that R_T and R_0 are $O(T^{-1})$. Using $D = 2rA + (1-\psi^2)C$, we can write

$$R_T - R_0 = (\gamma-\gamma_0)\frac{2A}{1-\gamma_0} + (\gamma-\gamma_0)^2\frac{-2rA}{1-\gamma_0^2} + F$$

where

$$\begin{aligned} F &\equiv \frac{(\gamma-\gamma_0)^2}{1-\gamma_0^2}\left(C - \frac{1-\gamma_0^{2T}}{T(1-\gamma_0^2)}\right)(\psi^2-1) \\ &= \frac{(\gamma-\gamma_0)^2}{1-\gamma_0^2}\left(\frac{1}{T-1}\left(\frac{1}{1-\gamma_0^2} - \frac{1}{T}\left(\frac{1}{1-\gamma_0}\right)^2\right) - \frac{1}{T(1-\gamma_0^2)}\right)(\psi^2-1) + o(T^{-k}) \end{aligned}$$

for any k. Clearly, $F = O(T^{-2})$. It follows that the $O(T^{-1})$ term of $l_T(\gamma) - l_0(\gamma)$ is free of ψ^2. The effect of non-stationary initial observations again shows up in the second-order bias term of the profile log-likelihood.

Table S.2 presents simulation results for the Gaussian autoregression with non-stationary initial observations, where the jackknife is bias-reducing. We generated $y_{i0} \sim \mathcal{N}(\alpha_{i0}/(1-\gamma_0), \psi^2\sigma_0^2/(1-\gamma_0^2))$ with ψ set to 0 and 2. These values correspond, respectively, to inlying and outlying initial observations relative to the steady-state distributions. The results show that the bias-corrected estimators continue to remove most of the small-sample bias from $\widehat{\gamma}$. The jackknife estimator $\widetilde{\gamma}_{1/2}$ generally performs better than the plug-in estimator $\widetilde{\gamma}_{\text{HK}} = \widehat{\gamma} + (1+\widehat{\gamma})/T$. When $\gamma_0 = .5$, the 5%-level validity tests both overreject the null when T is small, but the overrejection rates decrease as T increases, as predicted by the theory. This is because in the early periods the time-series are moving toward their steady state. This move becomes larger as $|\psi|$ moves farther away from 1. The impact of ψ vanishes as $\gamma_0 \to 1$ (Dhaene and Jochmans 2013), which explains the much improved acceptance rates for very small T when γ_0 is increased to .95.

Table S.2. Small-sample performance in a non-stationary Gaussian autoregression

			bias				confidence				validity	
T	γ_0	ψ	$\widehat{\gamma}$	$\widetilde{\gamma}_{\text{HK}}$	$\widetilde{\gamma}_{1/2}$	$\dot{\gamma}_{1/2}$	$\widehat{\gamma}$	$\widetilde{\gamma}_{\text{HK}}$	$\widetilde{\gamma}_{1/2}$	$\dot{\gamma}_{1/2}$	$\tilde{t}_{1/2}$	$\dot{t}_{1/2}$
4	.5	0	−.537	−.296	−.191	−.239	.000	.022	.304	.090	.601	.458
6	.5	0	−.340	−.147	−.054	−.121	.000	.233	.726	.373	.639	.688
8	.5	0	−.243	−.086	−.012	−.070	.000	.507	.834	.626	.737	.818
12	.5	0	−.151	−.039	.007	−.031	.000	.778	.866	.832	.855	.904
4	.5	2	−.244	.070	.084	−.099	.001	.747	.681	.687	.376	.480
6	.5	2	−.178	.043	.064	−.059	.001	.798	.662	.769	.304	.593
8	.5	2	−.142	.028	.044	−.039	.002	.854	.711	.836	.373	.691
12	.5	2	−.102	.014	.023	−.020	.013	.907	.808	.895	.585	.809
4	.95	0	−.609	−.274	−.220	−.405	.000	.023	.220	.000	.950	.746
6	.95	0	−.441	−.189	−.128	−.290	.000	.016	.332	.000	.945	.870
8	.95	0	−.346	−.146	−.088	−.225	.000	.014	.419	.000	.934	.915
12	.95	0	−.243	−.101	−.051	−.154	.000	.014	.520	.000	.922	.940
4	.95	2	−.511	−.152	−.111	−.330	.000	.370	.620	.001	.947	.729
6	.95	2	−.347	−.079	−.025	−.219	.000	.513	.824	.001	.928	.865
8	.95	2	−.257	−.046	.008	−.159	.000	.660	.850	.002	.909	.896
12	.95	2	−.166	−.017	.028	−.098	.000	.809	.717	.008	.875	.927

Model: $y_{it} = \alpha_{i0} + \gamma_0 y_{it-1} + \varepsilon_{it}$, $\varepsilon_{it} \sim \mathcal{N}(0, \sigma_0^2)$. Data generated with $N = 100$, $\sigma_0^2 = 1$, $\alpha_{i0} \sim \mathcal{N}(0,1)$, $y_{i0} \sim \mathcal{N}(\alpha_{i0}/(1-\gamma_0), \psi^2\sigma_0^2/(1-\gamma_0^2))$. 10,000 Monte Carlo replications.

5. DETAILS FOR THE AVERAGE DERIVATIVE OF THE SURVIVAL FUNCTION

With $\phi(\cdot)$ the standard normal density, $\widehat{\gamma}$ the within-group estimate, $\widehat{\sigma}^2$ the mean squared residual, $\widehat{\alpha}_i(\gamma,\sigma) = \frac{1}{T}\sum_{t=1}^T (y_{it} - \gamma y_{it-1})$, and $\widetilde{y}_{it} \equiv y_{it} - \frac{1}{T}\sum_{t=1}^T y_{it}$, the relevant expressions to form an estimate of σ_{c}^2 are

$$\mu_{it} = \mu_{it}(\gamma,\sigma,\alpha_i) = \gamma\phi\left(q_{it}\right)/\sigma, \qquad q_{it} \equiv (\alpha_i + \gamma y_{it-1})/\sigma,$$
$$\nabla_{\alpha_i}\mu_{it} = -q_{it}\mu_{it}/\sigma, \qquad \nabla_{\gamma}\partial\mu_{it} = \mu_{it}\left(1 - \gamma q_{it} y_{it-1}/\sigma\right)/\gamma,$$
$$\nabla_{\sigma}\mu_{it} = \mu_{it}\left(q_{it}^2 - 1\right)/\sigma, \qquad \psi_{it} = y_{it} - \alpha_i - \gamma y_{it-1},$$
$$\nabla_{\gamma}\widehat{\alpha}_i(\gamma,\sigma) = -\frac{1}{T}\sum_{t=1}^{T} y_{it-1}, \qquad \nabla_{\sigma}\widehat{\alpha}_i(\gamma,\sigma) = 0,$$

and

$$\widehat{\Sigma} = \frac{T-1}{NT^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{\widehat{\sigma}^2}\begin{pmatrix} \widetilde{y}_{it-1}^2 & 0 \\ 0 & 3r_{it}^2 - 1 \end{pmatrix}, \qquad r_{it} \equiv \left(\widetilde{y}_{it} - \widehat{\gamma}\widetilde{y}_{it-1}\right)/\widehat{\sigma},$$

together with those given in the main text.

6. DETAILS FOR THE HIGHER-ORDER EXPANSIONS IN THE LOGIT MODEL

Suppose that the binary variable y_{it} is generated by

$$y_{it} = 1\{\alpha_i + \theta y_{it-1} + \varepsilon_{it} \geq 0\},$$

where ε_{it} is i.i.d. with cdf $\Lambda(u) = e^u/(1+e^u)$. We observe y_{it} for $t = 0, \ldots, T$. We work with the likehood function that conditions on the y_{i0}. The true values α_{i0} are assumed to be i.i.d. draws from some (unknown) distribution $\mathcal{G}$.

Suppose, initially, that $\mathcal{G}$ is degenerate. We omit the index i. Let $y \equiv (y_0, \ldots, y_T)$ be a binary sequence of length $T+1$. For every possible y, define the vector of relative transition frequencies $w \equiv w(y) \equiv$

$(w_{00}, w_{01}, w_{10}, w_{11})' \in \Delta^3$, where

$$w_{00} \equiv \tfrac{1}{T}\textstyle\sum_{t=1}^{T} 1(y_{t-1}=0, y_t=0), \qquad w_{01} \equiv \tfrac{1}{T}\textstyle\sum_{t=1}^{T} 1(y_{t-1}=0, y_t=1),$$
$$w_{10} \equiv \tfrac{1}{T}\textstyle\sum_{t=1}^{T} 1(y_{t-1}=1, y_t=0), \qquad w_{11} \equiv \tfrac{1}{T}\textstyle\sum_{t=1}^{T} 1(y_{t-1}=1, y_t=1),$$

and Δ^3 is the 3-dimensional unit simplex. Also, define the vector of model transition probabilities $p \equiv p(\alpha,\theta) \equiv (p_{00}, p_{01}, p_{10}, p_{11})'$, where

$$\begin{aligned} p_{00} &\equiv \Pr[y_t = 0 | y_{t-1} = 0; \alpha, \theta] = \Lambda(-\alpha), & p_{01} &\equiv 1 - p_{00} = \Lambda(\alpha), \\ p_{10} &\equiv \Pr[y_t = 0 | y_{t-1} = 1; \alpha, \theta] = \Lambda(-\alpha-\theta), & p_{11} &\equiv 1 - p_{10} = \Lambda(\alpha+\theta). \end{aligned} \tag{S.6.1}$$

The contribution of any observed sequence y to the profile log-likelihood depends on y only via $w = w(y)$ and is given by

$$\widehat{l}_w(\theta) = w' \log p(\alpha_w(\theta), \theta),$$

where

$$\alpha_w(\theta) \equiv \arg\max_{\alpha \in A} w' \log p(\alpha, \theta)$$

(we write $\alpha_w(\theta)$ instead of $\widehat{\alpha}(\theta)$). The profile log-likelihood for $N = \infty$ is

$$l_T(\theta) = \mathbb{E}[w' \log p(\alpha_w(\theta), \theta)], \tag{S.6.2}$$

where the expectation is over w. Let $\varpi \equiv \mathbb{E}[w]$. Note that ϖ depends on θ_0 and $\mathcal{G}$ but not on T. We obtain $l_0(\theta)$ from $l_T(\theta)$ by replacing $\alpha_w(\theta)$ with

$$\alpha_\varpi(\theta) \equiv \arg\max_{\alpha \in A} \mathbb{E}[w' \log p(\alpha, \theta)],$$

which gives

$$l_0(\theta) = \mathbb{E}[w' \log p(\alpha_\varpi(\theta), \theta)] = \varpi' \log p(\alpha_\varpi(\theta), \theta). \tag{S.6.3}$$

We can write (S.6.2) and (S.6.3) more compactly as

$$l_T(\theta) = \mathbb{E}[g(w,\theta)], \qquad l_0(\theta) = g(\varpi, \theta),$$

where we define $g(v,\theta) \equiv v' \log p(\alpha_v(\theta), \theta)$ for every $v \equiv (v_{00}, v_{01}, v_{10}, v_{11})' \in \Delta^3$. A closed-form expression for $\alpha_v(\theta)$ is given below. Hence, given that

$$l_T(\theta) = \sum_w \Pr(w) g(w, \theta),$$

where $\Pr(w)$ is the probability of obtaining w, we can calculate $l_T(\theta)$ and $l_0(\theta)$ exactly.

We now examine the expansion of $l_T(\theta)$. Because w is a sample mean converging to ϖ and because all moments of w exist, we can expand $l_T(\theta) = \mathbb{E}[g(w,\theta)]$ around the point $w = \varpi$ provided that, roughly speaking, $g(v,\theta)$ is sufficiently smooth in v for every θ. We use the following notation. For $q \equiv (q_{00}, q_{01}, q_{10}, q_{11})' \in \mathbb{N}^4$, let $\nabla_q g(v,\theta) \equiv \partial^{\iota' q} g(v,\theta) / \partial v_{00}^{q_{00}} \ldots \partial v_{11}^{q_{11}}$, where $\iota' q$ is the sum of the elements of q, and $(w - \varpi)^q \equiv \Pi_{ij=0,1} (w_{ij} - \varpi_{ij})^{q_{ij}}$. We show below that $g(v,\theta)$ is bounded in v on Δ^3 and that, for all q, $\nabla_q g(v,\theta)$ is bounded in v on a neighborhood of ϖ. Therefore, by Theorem 2 of Hurst (1976), it follows that, as $T \to \infty$,

$$\mathbb{E}[g(w,\theta)] = g(\varpi,\theta) + \sum_{j=1}^{2k} \sum_{\iota' q = j} \frac{\nabla_q g(\varpi,\theta)}{q_{00}! \ldots q_{11}!} \mathbb{E}[(w-\varpi)^q] + O(T^{-k-1/2})$$

for every k. Furthermore, for every q,

$$\mathbb{E}[(w-\varpi)^q] = \sum_{r = \lfloor (\iota' q + 1)/2 \rfloor}^{\iota' q - 1} \frac{c_r(q)}{T^r}$$

with coefficients $c_r(q)$ depending on r and q but not on T. On collecting terms of the same order, we conclude that $l_T(\theta)$ satisfies the expansion in (2.5) for every θ and any k.

So far we have assumed that $\mathcal{G}$ is degenerate, i.e., $\alpha_{i0} = \alpha_0$ for all i. The functions $l_T(\theta)$ and $l_0(\theta)$ and the functions $C_j(\theta)$ that appear in the expansion all implicitly depend on α_0. When $\mathcal{G}$ is non-degenerate, these functions have to be replaced with their integrals with respect to $\mathcal{G}$ provided that the integrals exist. A sufficient condition for the existence of the integrals is that $\mathcal{G}$ has bounded support.

Technical details After some algebra, we obtain $\alpha_v(\theta)$ in closed form:

$$\alpha_v(\theta) = \arg\max_{\alpha \in A} v' \log p(\alpha, \theta) = -\log x_v, \tag{S.6.4}$$

where

$$x_v = \frac{1}{2}\left(a_v + b_v e^\theta + \sqrt{\left(a_v + b_v e^\theta\right)^2 + 4\left(a_v + b_v + 1\right)e^\theta}\right),$$
$$a_v = \frac{v_{00} - v_{11}}{v_{01} + v_{11}}, \qquad b_v = \frac{v_{10} - v_{01}}{v_{01} + v_{11}},$$

and with the understanding that limits are taken as v approaches the boundary of Δ^3. In particular, as $v_{01}, v_{11} \to 0$, we have $a_v \to \infty$ or $b_v \to \infty$, implying $x_v \to \infty$ and $\alpha_v(\theta) \to -\infty$; and, as $v_{00}, v_{10} \to 0$, we have $a_v + b_v + 1 \to 0$, $x_v \to 0$, and $\alpha_v(\theta) \to \infty$. Note also that $a_v + b_v + 1 \geq 0$, ensuring $x_v \geq 0$. Using (S.6.1) and (S.6.4), we have

$$\begin{aligned} g(v,\theta) &= v' \log p(\alpha_v(\theta), \theta) \\ &= v_{00} \log \frac{x_v}{1 + x_v} + v_{01} \log \frac{1}{1 + x_v} + v_{10} \log \frac{x_v}{e^\theta + x_v} + v_{11} \log \frac{e^\theta}{e^\theta + x_v} \\ &= A + B + C + D \qquad \text{(say)}. \end{aligned}$$

To show that $g(v,\theta)$ is bounded in v on Δ^3, it suffices to examine A and C as $x_v \to 0$ and B and D as $x_v \to \infty$. When $x_v \to 0$, we have $a_v + b_v + 1 \to 0$, $v_{00} \to 0$, $v_{10} \to 0$, $v_{00} \log x_v \to 0$, and $v_{10} \log x_v \to 0$, so A and C are bounded. When $x_v \to \infty$, we have $a_v + b_v e^\rho \to \infty$, $v_{01} \to 0$, $v_{11} \to 0$, $v_{01} \log(1 + x_v) \to 0$, and $v_{11} \log(1 + x_v) \to 0$, so B and D are bounded. We now examine the partial derivatives of g with respect to v in a neighborhood of ϖ. Consider the term A. Because ϖ is an interior point of Δ^3, x_v is bounded away from zero and infinity in a neighborhood of ϖ. Therefore, the function $\log \frac{x}{1+x}$ is bounded in a neighborhood of x_ϖ, and so are its derivatives of all orders. Further, x_v, viewed as a function of a_v and b_v, is bounded in a neighborhood of a_ϖ and b_ϖ, and so are its partial derivatives of all orders with respect to a_v and b_v because the square-root term that appears in x_v is bounded away from zero. Finally, the partial derivatives of a_v and b_v of all orders are bounded in a neighborhood of ϖ because the denominator of a_v and b_v is bounded away from zero. By the chain rule, the partial derivatives of A of all orders are bounded as required. By the same reasoning, the same is true for B, C, and D. □

7. UNINFORMATIVENESS AND SEPARATION IN BINARY PANEL DATA

The maximum-likelihood estimator of a dynamic binary panel model with fixed effects may be indeterminate or infinite. This occurs when the profile likelihood is flat or when its maximum is reached at infinity, respectively. We characterize these situations for binary AR(1) data without covariates or with one covariate.[1] This leads to an explicit derivation of $T_{\min}$ and $T'_{\min}$.

[1] The problem of data separation in binary and multinomial data is well known in the cross-sectional setting. Albert and Anderson (1984) give a complete taxonomy for multinomial data.

7.1. Binary choice without covariates

The model is $y_{it} = 1(\alpha_i + \rho y_{it-1} - \varepsilon_{it} \geq 0)$, where the cdf of ε_{it}, say F, is continuous and strictly increasing on the real line. The contribution of unit i to the profile log-likelihood is

$$\widehat{l}_i(\rho) = \max_{\alpha_i} \left\{ A_i \log\left(1 - F(\alpha_i)\right) + B_i \log F(\alpha_i) \; + C_i \log\left(1 - F(\alpha_i + \rho)\right) + D_i \log F(\alpha_i + \rho) \right\}, \quad \text{(S.7.1)}$$

where

$$A_i \equiv T^{-1} \textstyle\sum_{t=1}^{T} 1(y_{it-1} = 0, y_{it} = 0), \qquad B_i \equiv T^{-1} \textstyle\sum_{t=1}^{T} 1(y_{it-1} = 0, y_{it} = 1),$$
$$C_i \equiv T^{-1} \textstyle\sum_{t=1}^{T} 1(y_{it-1} = 1, y_{it} = 0), \qquad D_i \equiv T^{-1} \textstyle\sum_{t=1}^{T} 1(y_{it-1} = 1, y_{it} = 1),$$

are transition frequencies.

Let $\widehat{\rho}_i \equiv \arg\max_\rho \widehat{l}_i(\rho)$ and $\widehat{\rho} \equiv \arg\max_\rho N^{-1} \sum_{i=1}^{N} \widehat{l}_i(\rho)$. A sequence $y_i = (y_{i0}, \ldots, y_{iT})$ is *uninformative* if $\widehat{l}_i(\rho)$ is constant. This occurs if and only if

$$A_i = B_i = 0 \text{ or } C_i = D_i = 0 \text{ or } A_i = C_i = 0 \text{ or } B_i = D_i = 0. \quad \text{(S.7.2)}$$

The "if" part follows from noting that $\widehat{l}_i(\rho)$ corresponding to the four cases in (S.7.2) is

$$\widehat{l}_i(\rho) = \max_{\alpha_i} \left\{ C_i \log\left(1 - F(\alpha_i + \rho)\right) + D_i \log F(\alpha_i + \rho) \right\} = \max_{\alpha_i} \left\{ C_i \log\left(1 - F(\alpha_i)\right) + D_i \log F(\alpha_i) \right\},$$
$$\widehat{l}_i(\rho) = \max_{\alpha_i} \left\{ A_i \log\left(1 - F(\alpha_i)\right) + B_i \log F(\alpha_i) \right\},$$
$$\widehat{l}_i(\rho) = \max_{\alpha_i} \left\{ B_i \log F(\alpha_i) + D_i \log F(\alpha_i + \rho) \right\} B_i \log F(+\infty) + D_i \log F(+\infty + \rho) = 0,$$
$$\widehat{l}_i(\rho) = \max_{\alpha_i} \left\{ A_i \log\left(1 - F(\alpha_i)\right) + C_i \log\left(1 - F(\alpha_i + \rho)\right) \right\} = 0,$$

respectively. In each case $\widehat{l}_i(\rho)$ is constant. For the "only if" part, if (S.7.2) does not hold, then either $B_i \neq 0$ and $C_i \neq 0$, or at most one of A_i, B_i, C_i, D_i is zero (note that $B_i = C_i = 0$, $A_i \neq 0$, and $D_i \neq 0$ cannot jointly occur). In all of these cases, $\widehat{l}_i(\rho)$ can be taken to $-\infty$ by taking ρ to $-\infty$ or $+\infty$, while $\widehat{l}_i(\rho)$ is finite if ρ is finite; hence $\widehat{l}_i(\rho)$ is non-constant. Because $\widehat{\rho}_i$ is indeterminate if and only y_i is uninformative, $\widehat{\rho}$ is indeterminate if and only if all y_i are uninformative. Uninformative sequences are removed, as they do not affect $\widehat{\rho}$. A sequence y_i is *monotone* if $B_i = 0$ or $C_i = 0$. A sequence y_i is *semi-alternating* if $A_i = 0$ or $D_i = 0$. We have $\widehat{\rho}_i = +\infty$ if and only if y_i is informative and monotone, and $\widehat{\rho}_i = -\infty$ if and only if y_i is informative and semi-alternating. Suppose there is at least one informative sequence. Then, $\widehat{\rho} = +\infty$ if and only if all informative sequences are monotone, and $\widehat{\rho} = -\infty$ if and only if all informative sequences are semi-alternating. When $T = 2$, there are eight possible sequences y_i. Only two of these, $(0, 1, 0)$ and $(1, 0, 1)$, are informative. Both are semi-alternating, so either $\widehat{\rho}$ is indeterminate or $\widehat{\rho} = -\infty$, implying $\rho_2 = -\infty$ and $T_{\min} > 2$. Further, both have $A_i = D_i = 0$, $B_i = C_i = \frac{1}{2}$, and

$$\widehat{l}_i(\rho) = \max_{\alpha_i} \frac{1}{2} \left\{ \log F(\alpha_i) + \log(1 - F(\alpha_i + \rho)) \right\} \equiv \lambda(\rho) \quad \text{(say)}.$$

It follows that $l_2(\rho) = \pi\lambda(\rho) + c$, where π is the probability that y_i is informative and c is an inessential constant. Hence $T'_{\min} = 2$. When $T = 3$, some informative sequences are monotone, for example, $(0, 0, 1, 1)$, and others are semi-alternating, for example, $(0, 1, 1, 0)$. Hence $T_{\min} = 3$.

7.2. Binary choice with a covariate

The contribution of unit i to the profile log-likelihood is

$$\widehat{l}_i(\rho,\beta) = \max_{\alpha_i} T^{-1} \sum_{t=1}^{T} \{A_i \log(1-F(\alpha_i+\beta x_{it})) + B_i \log F(\alpha_i+\beta x_{it}) \\ +C_i \log(1-F(\alpha_i+\rho+\beta x_{it})) + D_i \log F(\alpha_i+\rho+\beta x_{it})\},$$

with $A_i,\dots,D_i$ as before.

Assume that $x_{it} \neq x_{it'}$ for $t \neq t'$. Let $(\widehat{\rho}_i,\widehat{\beta}_i) \equiv \arg\max_{\rho,\beta} \widehat{l}_i(\rho,\beta)$ and $(\widehat{\rho},\widehat{\beta}) \equiv \arg\max_{\rho,\beta} N^{-1}\sum_{i=1}^{N} \widehat{l}_i(\rho,\beta)$. A sequence y_i is *uninformative about* β_0 if $\widehat{l}_i(\rho,\beta)$ is constant in β, which occurs if and only if $A_i = C_i = 0$ or $B_i = D_i = 0$. A sequence y_i is *uninformative about* ρ_0 if $\widehat{l}_i(\rho,\beta)$ is constant in ρ, which occurs if and only if (S.7.2) holds. The maximum-likelihood estimator ($\widehat{\rho}$ or $\widehat{\beta}$) is indeterminate if and only if all y_i are uninformative about the corresponding parameter. Indeterminacy of $\widehat{\beta}$ implies indeterminacy of $\widehat{\rho}$. Remove the sequences that are uninformative about ρ_0 so that any remaining y_i is informative about ρ_0 and β_0. A sequence y_i is *separable* if there exists $(\rho,\beta)\neq(0,0)$ such that

$$\rho y_{it-1} + \beta x_{it} \geq \rho y_{it'-1} + \beta x_{it'} \quad \text{for all } t,t' \geq 1 : y_{it}=1 \text{ and } y_{it'}=0. \tag{S.7.3}$$

$\widehat{\rho}_i = \pm\infty$ or $\widehat{\beta}_i = \pm\infty$ if and only if y_i is *separable*, and $\widehat{\rho} = \pm\infty$ or $\widehat{\beta} = \pm\infty$ if and only if all y_i are jointly separable, i.e., there exists $(\rho,\beta)\neq(0,0)$ such that (S.7.3) holds for all i. To check for joint separability, let $T_i^{a,b} = \{t : y_{it-1}=a, y_{it}=b\}$ for $a,b \in \{0,1\}$, define the intervals

$$X_i^{a,b} = \begin{cases} [\min_{t\in T_i^{a,b}} x_{it},\ \max_{t\in T_i^{a,b}} x_{it}] & \text{if } T_i^{a,b} \neq \varnothing, \\ \varnothing & \text{if } T_i^{a,b} = \varnothing, \end{cases}$$

and note that (S.7.3) holds for all i if and only if

$$\begin{array}{lll} \beta X_i^{0,1} \geq \beta X_i^{0,0}, & \beta X_i^{1,1} \geq \beta X_i^{1,0}, & \text{for all } i, \\ \beta X_i^{0,1} \geq \rho + \beta X_i^{1,0}, & \beta X_i^{1,1} \geq -\rho + \beta X_i^{0,0}, & \text{for all } i, \end{array} \tag{S.7.4}$$

where $S_1 \geq S_2$ means $s_1 \geq s_2$ for all $s_1 \in S_1$ and $s_2 \in S_2$. It suffices to check whether (S.7.4) has a non-zero solution with $\beta \in \{-1,0,1\}$. For $\beta = 0$, there is a solution with $\rho \neq 0$ if and only if all y_i are monotone (because then either $X_i^{0,1}$ or $X_i^{1,0}$ is empty) or all y_i are semi-alternating (because then either $X_i^{1,1}$ or $X_i^{0,0}$ is empty). For $\beta \in \{-1,1\}$, define

$$\rho_{\max}(\beta) = \max\{\rho : \beta X_i^{0,1} \geq \rho + \beta X_i^{1,0} \text{ for all } i\}, \qquad \rho_{\min}(\beta) = \min\{\rho : \beta X_i^{1,1} \geq -\rho + \beta X_i^{0,0} \text{ for all } i\}.$$

Thus, if and only if

$$X_i^{0,1} \geq X_i^{0,0}, \quad X_i^{1,1} \geq X_i^{1,0}, \quad \text{for all } i; \quad \rho_{\max}(1) \geq \rho_{\min}(1);$$

or

$$X_i^{0,1} \leq X_i^{0,0}, \quad X_i^{1,1} \leq X_i^{1,0}, \quad \text{for all } i; \quad \rho_{\max}(-1) \geq \rho_{\min}(-1),$$

is there a solution with $\beta \in \{-1,1\}$.

8. EMPIRICAL APPLICATION

Table S.3 presents descriptive statistics of the data used in the empirical application. Table S.4 provides the estimation results of the model

$$y_{it} = 1\{\alpha_{i0} + \gamma_0 y_{it-1} + x_{it}'\delta_0 + d_{it}'\beta_0 \geq \varepsilon_{it}\},$$

where ε_{it} are independent standard-normal innovations, x_{it} is the same vector of time-varying covariates as in the main text, and d_{it} is a set of year dummies.

Table S.3. Descriptive statistics

	mean and standard deviation (in parentheses) over all units (1461 observations)								
year	1980	1981	1982	1983	1984	1985	1986	1987	1988
lagged participation	.722	.707	.695	.692	.711	.746	.740	.741	.735
	(.448)	(.455)	(.461)	(.462)	(.453)	(.435)	(.439)	(.438)	(.441)
# children 0–2	.323	.331	.318	.261	.211	.190	.168	.137	.103
	(.528)	(.537)	(.533)	(.494)	(.455)	(.440)	(.408)	(.363)	(.325)
# children 3–5	.307	.313	.321	.330	.337	.320	.268	.215	.185
	(.524)	(.525)	(.535)	(.524)	(.540)	(.541)	(.505)	(.463)	(.430)
# children 6–17	.934	.960	.973	1.015	1.034	1.077	1.124	1.164	1.165
	(.138)	(1.124)	(1.102)	(1.081)	(1.068)	(1.062)	(1.064)	(1.081)	(1.109)
husband income	39.200	39.041	39.115	40.541	43.039	43.572	44.485	45.580	46.038
	(23.514)	(23.598)	(30.601)	(34.375)	(41.915)	(39.798)	(42.621)	(53.411)	(55.784)
age	33.310	34.251	35.300	36.279	37.288	38.358	39.272	40.309	41.336
	(8.841)	(8.848)	(8.829)	(8.861)	(8.863)	(8.847)	(8.845)	(8.837)	(8.857)
	mean and standard deviation (in parentheses) over units who change participation status (664 observations)								
year	1980	1981	1982	1983	1984	1985	1986	1987	1988
lagged participation	.580	.538	.511	.505	.547	.623	.610	.613	.599
	(.494)	(.499)	(.500)	(.500)	(.498)	(.485)	(.488)	(.487)	(.490)
# children 0–2	.411	.434	.408	.336	.262	.238	.199	.157	.108
	(.572)	(.595)	(.577)	(.547)	(.498)	(.483)	(.429)	(.388)	(.330)
# children 3–5	.372	.381	.407	.423	.444	.408	.339	.271	.229
	(.559)	(.564)	(.582)	(.563)	(.586)	(.582)	(.556)	(.514)	(.462)
# children 6–17	.840	.908	.955	1.035	1.098	1.181	1.264	1.355	1.378
	(1.109)	(1.116)	(1.095)	(1.075)	(1.078)	(1.079)	(1.070)	(1.105)	(1.139)
husband income	40.340	40.693	40.402	42.714	45.742	45.925	46.921	47.749	48.438
	(23.838)	(24.445)	(33.351)	(41.737)	(49.612)	(44.334)	(50.015)	(49.305)	(54.358)
age	31.574	32.526	33.565	34.559	35.592	36.622	37.544	38.581	39.611
	(8.320)	(8.318)	(8.300)	(8.342)	(8.352)	(8.317)	(8.320)	(8.309)	(8.333)

Data source: PSID 1979–1988.

Table S.4. Female labor-force participation: Estimation results including time dummies

	model parameters						
	$\widehat{\theta}$	$\widetilde{\theta}_{1/2}$	$\widetilde{\theta}_{\mathrm{HK}}$	$\widetilde{\theta}_{\mathrm{F}}$	$\dot{\theta}_{1/2}$	$\dot{\theta}_{\mathrm{AH}}$	$\dot{\theta}_{\mathrm{C}}$
lagged participation	.757	1.351	.994	1.034	1.057	.982	1.096
	(.043)	(.053)	(.043)	(.043)	(.053)	(.043)	(.043)
# kids 0–2	−.553	−.639	−.478	−.436	−.534	−.473	−.409
	(.058)	(.087)	(.058)	(.058)	(.087)	(.058)	(.058)
# kids 3–5	−.290	−.360	−.226	−.203	−.256	−.188	−.188
	(.053)	(.091)	(.054)	(.054)	(.091)	(.054)	(.054)
# kids 6–17	−.074	−.145	−.055	−.049	−.063	.032	−.038
	(.043)	(.078)	(.043)	(.043)	(.078	(.043)	(.043)
log husband income	−.252	−.313	−.233	−.213	−.257	−.010	−.214
	(.055)	(.075)	(.056)	(.056)	(.075)	(.055)	(.056)
age	2.333	1.762	2.094	1.832	2.170	−.144	1.814
	(.627)	(1.082)	(.636)	(.636)	(1.082)	(.631)	(.639)
age squared	−.244	−.151	−.212	−.188	−.222	.014	−.187
	(.052)	(.118)	(.053)	(.053)	(.118)	(.052)	(.053)

Coefficients for age and age squared are multiplied by 10 and 100, respectively. Standard errors in parentheses.
Data source: PSID 1979–1988.

REFERENCES

Albert, A. and J. Anderson (1984). On the existence of maximum likelihood estimators in logistic regression models. *Biometrika 71*, 1–10.

Bun, M. J. and M. A. Carree (2005). Bias-corrected estimation in dynamic panel data models. *Journal of Business and Economic Statistics 23*, 200–210.

Dhaene, G. and K. Jochmans (2013). Likelihood inference in an autoregression with fixed effects. Discussion Paper No 2013-07, Department of Economics, Sciences Po.

Hahn, J. and G. Kuersteiner (2002). Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and T are large. *Econometrica 70*, 1639–1657.

Hurst, J. (1976). Asymptotic expansions of functions of statistics. *Applications of Mathematics 21*, 444—456.

Littlewood, D. E. (1958). *The Theory of Group Characters and Matrix Representations of Groups* (Second ed.). Oxford, Clarendon.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica 49*, 1417–1426.