Kaddoura, Yousef; Westerlund, Joakim

Working Paper

Estimation of Panel Data Models with Interactive Effects and Multiple Structural Breaks When T Is Fixed

Working Paper, No. 2021:15

Provided in Cooperation with:
Department of Economics, School of Economics and Management, Lund University

Suggested Citation: Kaddoura, Yousef; Westerlund, Joakim (2021) : Estimation of Panel Data Models with Interactive Effects and Multiple Structural Breaks When T Is Fixed, Working Paper, No. 2021:15, Lund University, School of Economics and Management, Department of Economics, Lund

This Version is available at:
https://hdl.handle.net/10419/260335

Working Paper 2021:15

Department of Economics
School of Economics and Management

Estimation of Panel Data Models with Interactive Effects and Multiple Structural Breaks When T Is Fixed

Yousef Kaddoura
Joakim Westerlund

November 2021

LUND UNIVERSITY

Estimation of Panel Data Models with Interactive Effects and Multiple Structural Breaks when T is Fixed *

Yousef Kaddoura
Lund University

Joakim Westerlund†
Lund University
and
Deakin University

September 17, 2021

Abstract

In this article, we propose a new estimator of panel data models with interactive fixed effects and multiple structural breaks that is suitable when the number of time periods, T, is fixed and only the number of cross-sectional units, N, is large. This is done by viewing the determination of the breaks as a shrinkage problem, and to estimate both the regression coefficients, and the number of breaks and their locations by applying a version of the Lasso approach. We show that with probability approaching one the approach can correctly determine the number of breaks and the dates of these breaks, and that the estimator of the regime-specific regression coefficients is consistent and asymptotically normal. We also provide Monte Carlo results suggesting that the approach performs very well in small samples, and empirical results suggesting that the coefficients of the deterrence model of crime are not constant as typically assumed but subject to structural change.

JEL Classification: C13; C23; C33; K42.

Keywords: Panel data; Interactive effects; Common factors; Structural change; Lasso.

*Westerlund would like to thank the Knut and Alice Wallenberg Foundation for financial support through a Wallenberg Academy Fellowship.

†Department of Economics, Lund University, Box 7082, 220 07 Lund, Sweden. Telephone: +46 46 222 8997. Fax: +46 46 222 4613. E-mail address: joakim.westerlund@nek.lu.se.

1 Introduction

Dealing with structural breaks is an important step in most, if not all, empirical economic research. This is particularly true in panel data comprised of many cross-sectional units, such as individuals, firms or countries, which are all affected by major economic events. The worry is that if left unattended, existing breaks will manifest themselves as omitted variables, leading to inconsistent estimates of the slope coefficients of the model. It is therefore important to know if and when structural breaks have occurred. Of course, such knowledge is rarely available in practice, which means that it has to be inferred from the data. We need to be able to check if there are any breaks present and, if there are, to infer both the break dates and the regime-specific slope coefficients. This should be possible even if the number of time periods, T, is fixed and only the number of cross-sectional units, N, is large, as many economic data sets have this "short" form. The procedure should also be easy to implement, it should not require data to be stationary, and it should be robust to unobserved heterogeneity. This last demand is potentially very important because unattended heterogeneity can be mistaken for structural breaks. The current paper contributes by developing a procedure that meets the above list of demands.

While the literature concerned with structural breaks in time series is huge, the literature concerned with such breaks in panel data is much smaller (see Boldea et al., 2020, for a recent overview). Yet, panel data are particularly susceptible to structural change. One reason for this is that the sample frequency is usually much lower than in pure series data. Panel data sets therefore tend to have long time spans, which means that the assumption of constant coefficients is likely to be violated because of major economic events. Another reason is that while T is usually quite small, because the number of cross-sectional units for which time series data is readily available is ever-increasing, N is potentially very large. This is important because the larger is N, the higher the risk that at least some of the cross-sectional units are subject to structural change. A related issue is how many breaks there are. If the literature on structural breaks in panel data is sparse, the part of the literature that deals with an unknown number of breaks is almost nonexistent.[1] The only exceptions known to us are Boldea et al. (2020), Li et al. (2016),

[1] An incomplete list of studies dealing with a single structural break in panel data include Antoch et al. (2019), Baltagi et al. (2016, 2017), Hidalgo and Schafgans (2017), Karavias et al. (2021), and Zhu et al. (2020).

and Qian and Su (2016), where the last two studies assume that both N and T are large, which is again something that we would like to avoid in the present paper.[2]

Another prominent feature of the type of disaggregated "micro" panel data that we have in mind, where typically the regressors explain only a small fraction of the variation in the dependent variable, is the presence of unobserved heterogeneity. Studies such as Ahn et al. (2013), Bai (2009), Moon and Weidner (2015), Pesaran (2006), Robertson and Sarafidis (2015), and Westerlund et al. (2019) allow for unobserved heterogeneity in the form of interactive effects that are dealt with by using either some kind of "de-factoring" or generalized method of moments (GMM); however, they do not allow for breaks and many assume that T is large. Li et al. (2016) allow for both multiple breaks and interactive effects, but then again in their paper T is large. Boldea et al. (2020) allow for interactive effects without for that matter requiring any correction thereof. This makes their approach very simple, although at a cost in terms of additional restrictive conditions. In particular, it is assumed that the omitted variables bias caused by the omitted interactive effects is time-invariant, up to the breakpoints, which limits the type of effects and regressors that can be permitted.

The proposed methodology builds upon the so-called "adaptive group fused" Lasso approach of Li et al. (2016), and Qian and Su (2016), which is suitable when the variation in the slopes has a natural ordering, as when time-stamped like in the current paper. However, because in our setup T is fixed, we cannot use principal components as a means to purge the interactive effects as in Li et al. (2016). In fact, a major complication when T is fixed is that we cannot easily separate the breaks from the effects. Qian and Su (2016) transform their data by taking first-differences before applying Lasso, which is expected to work also when T is fixed. However, differencing can only handle time-invariant effects. Moreover, while differencing solves the separation problem, it does so in an awkward way, since the (time-varying) slope coefficients in the model for the data in differences are not the same as for the data in levels.

The approach used in the present paper can be seen as a reaction to the discussion of the last paragraph. The idea is to apply Lasso to cross-sectionally demeaned data. The demeaning

[2]Qian and Su (2016) recognize the importance of allowing T to be finite and discuss likely implications for theory, but they do not provide any formal results for the fixed-T case. Similarly, while in Baltagi et al. (2016) there is a discussion of how to proceed in the presence of multiple breaks, their theory supposes that there is just one break.

does not affect the slopes and it makes the resulting estimator, henceforth referred to as "post-demeaned Lasso least squares (LS)", "PDL2S" for short, robust to interactive effects, provided that they satisfy a certain random coefficient condition. Another advantage of the new procedure is that it puts almost no assumptions on the structure of the breaks. In fact, there can be no breaks at all, and if there are breaks present the procedure does not make any assumptions about their number. The procedure is therefore valid even if some, or indeed all, regimes have a single observation, which is very useful when wanting to detect a break as quickly as possible. Yet another advantage is that the procedure does not place any conditions on the serial correlation properties of the data. Hence, the data can be stationary, as required in the bulk of the previous literature (see Baltagi et al., 2017, for a discussion), but it does not have to be.

The rest of the paper is organized as follows. Section 2 describes the model and the PDL2S approach that we will use to estimate it. Section 3 reports our main asymptotic results, whose accuracy in small samples is evaluated by means of Monte Carlo simulation in Section 4. Section 5 presents the results of a small empirical illustration using as an example the economics of crime. Section 6 concludes. All proofs and theoretical results of secondary nature are provided in the online appendix.

2 Model and estimator

Consider a scalar panel data variable $y_{i,t}$, observable across $t = 1, \ldots, T$ time periods and $i = 1, \ldots, N$ cross-section units. The data generating process of this variable is given by

$$y_{i,t} = \mathbf{x}_{i,t}'\boldsymbol{\beta}_t + u_{i,t}, \tag{2.1}$$

$$u_{i,t} = \boldsymbol{\lambda}_i'\mathbf{f}_t + \varepsilon_{i,t}, \tag{2.2}$$

where $\mathbf{x}_{i,t}$ is a $p \times 1$ vector of known regressors with $\boldsymbol{\beta}_t$ being a conformable vector of unknown slope coefficients that we allow to change over time, and $u_{i,t}$ is a composite error term that can be both serially and cross-sectionally correlated in a very general fashion. The assumption we make is that $u_{i,t}$ admits to a common factor structure in which $\mathbf{f}_t$ and $\boldsymbol{\lambda}_i$ are $r \times 1$ vectors of unobserved factors and loadings, respectively, and $\varepsilon_{i,t}$ is a mean zero error term.[3] The interactive effects are

[3] As we explain later in Section 3, the type of factors that can be permitted under our assumptions is very broad. This suggests that there is no need to discriminate between known and unknown factors, but that one can just as

here given by $\lambda_i'\mathbf{f}_t$. Any cross-sectional dependence in $u_{i,t}$ are assumed to be captured by these effects, so that the remainder, $\varepsilon_{i,t}$, is completely idiosyncratic. As usual, r and p are assumed to be fixed numbers.

We will assume that $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_T$ takes on $m+1$ distinct vectors $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{m+1}$, such that

$$\boldsymbol{\beta}_t = \boldsymbol{\alpha}_j \tag{2.3}$$

for $t = T_{j-1}, \ldots, T_j - 1$, $j = 1, \ldots, m+1$, $m \in [0, T-1]$, $T_0 = 1$ and $T_{m+1} = T+1$. Hence, in this model, $\boldsymbol{\beta}_t$ has $m+1$ distinct regimes, or m breaks, that occur at time $T_1, \ldots, T_m$. At the one end of the scale, we have $m = 0$, in which case there is only one regime and $\boldsymbol{\beta}_1 = \ldots = \boldsymbol{\beta}_T = \boldsymbol{\alpha}_1$, whereas, at the other end, $m = T-1$, which means that there are as many regimes as time periods, and hence $\boldsymbol{\beta}_t = \boldsymbol{\alpha}_t$ for all $t = 1, \ldots, T$. It is useful to stack $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{m+1}$ and $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_T$ into the $(m+1)p \times 1$ and $Tp \times 1$ vectors $\mathbf{A}_m = [\boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_{m+1}']'$ and $\mathbf{B}_T = [\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_T']'$, respectively, and to denote by $\mathcal{T}_m = \{T_1, \ldots, T_m\}$ the set of breakpoints when $m > 0$. If $m = 0$, then we define $\mathcal{T}_m = \mathcal{T}_0 = \varnothing$ as the empty set. It is also useful to note that if $m+1 = T$, so that each regime contains only one observation, then the set of breakpoints is given by $\mathcal{T}_{T-1} = \{1, \ldots, T\}$. In what follows, we will therefore use $\mathcal{T}_{T-1}$ to denote the full set of time series observations.

Remark 1. The fact that the number of time periods within each regime is completely unrestricted is noteworthy because in the existing literature it is standard to assume that the break regimes are expanding with T (see, for example, Baltagi et al., 2016). There is also no need to truncate the sample endpoints, and in this way restrict the breakpoint to the middle of the sample, which is again standard in the literature. This means that breaks can be detected very quickly.

The goal of this paper is to infer $\mathbf{A}_m$ and $\mathcal{T}_m$. Let us therefore denote by $\mathbf{A}_{m^0}^0 = [\boldsymbol{\alpha}_1^{0\prime}, \ldots, \boldsymbol{\alpha}_{m^0+1}^{0\prime}]'$ the true value of $\mathbf{A}_m$, where m^0 is the true value of m. The set of true breakpoints is henceforth denoted $\mathcal{T}_{m^0}^0 = \{T_1^0, \ldots, T_{m^0}^0\}$. It is also useful to introduce $\mathbf{B}_T^0 = [\boldsymbol{\beta}_1^{0\prime}, \ldots, \boldsymbol{\beta}_T^{0\prime}]'$ as the true value of $\mathbf{B}_T$.

well treat them all as unknown. This is the main rationale for writing (2.2) in terms of (the unknown) $\mathbf{f}_t$ only.

Denote by $\bar{\mathbf{a}}_t = N^{-1}\sum_{i=1}^N \mathbf{a}_{i,t}$ the cross-sectional average of any variable $\mathbf{a}_{i,t}$, and let $\tilde{\mathbf{a}}_{i,t} = \mathbf{a}_{i,t} - \bar{\mathbf{a}}_t$ be the cross-sectionally demeaned version of $\mathbf{a}_{i,t}$. In this notation, (2.1) can be written as

$$\tilde{y}_{i,t} = \tilde{\mathbf{x}}_{i,t}'\boldsymbol{\beta}_t + \tilde{u}_{i,t}. \tag{2.4}$$

Cross-sectional demeaning is tantamount to demeaning with respect to common time effects. It is important to point out, however, that while we do allow for common time effects, our model does not necessarily include such effects. Hence, unless $\boldsymbol{\lambda}_i = \boldsymbol{\lambda}$ for all i, so that $\boldsymbol{\lambda}_i'\mathbf{f}_t = \boldsymbol{\lambda}'\mathbf{f}_t$, the model is misspecified. Demeaning is still key, though, as it enables us to eliminate the mean of $\boldsymbol{\lambda}_i$ from the regression error in (2.4), which is enough to ensure consistency and asymptotic (mixed) normality as long as the remaining part is uncorrelated with $\tilde{\mathbf{x}}_{i,t}$.

To estimate $\mathbf{B}_T^0$, we propose minimizing the following objective function:

$$\ell_\gamma(\mathbf{B}_T) = \frac{1}{N}\sum_{i=1}^N\sum_{t=1}^T(\tilde{y}_{i,t} - \tilde{\mathbf{x}}_{i,t}'\boldsymbol{\beta}_t)^2 + \gamma\cdot\sum_{t=2}^T w_t\|\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1}\|, \tag{2.5}$$

where $\gamma = \gamma(N) > 0$ is a tuning parameter, w_t is a data-driven weight defined by $w_t = \|\dot{\boldsymbol{\beta}}_t - \dot{\boldsymbol{\beta}}_{t-1}\|^{-\kappa}$, $\kappa > 0$ is a user-specified constant, and $\dot{\boldsymbol{\beta}}_t$ is a preliminary estimator of $\boldsymbol{\beta}_t$, which is obtained by minimizing the first term in $\ell_\gamma(\mathbf{B}_T)$. That is, $\dot{\boldsymbol{\beta}}_t$ is simply the period-by-period LS estimator;

$$\dot{\boldsymbol{\beta}}_t = \left(\sum_{i=1}^N \tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,t}'\right)^{-1}\sum_{i=1}^N \tilde{\mathbf{x}}_{i,t}\tilde{y}_{i,t}. \tag{2.6}$$

Simple as it may be, it is useful to be able to write this estimator in a more general notation. Let us therefore introduce

$$\mathbf{Q}_N(\mathcal{T}_m) = \operatorname{diag}\left(\frac{1}{N}\sum_{t=T_0}^{T_1-1}\sum_{i=1}^N \tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,t}', \ldots, \frac{1}{N}\sum_{t=T_m}^{T_{m+1}-1}\sum_{i=1}^N \tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,t}'\right), \tag{2.7}$$

$$\mathbf{R}_N(\mathcal{T}_m) = \begin{bmatrix} \frac{1}{N}\sum_{t=T_0}^{T_1-1}\sum_{i=1}^N \tilde{\mathbf{x}}_{i,t}\tilde{y}_{i,t} \\ \vdots \\ \frac{1}{N}\sum_{t=T_m}^{T_{m+1}-1}\sum_{i=1}^N \tilde{\mathbf{x}}_{i,t}\tilde{y}_{i,t} \end{bmatrix}, \tag{2.8}$$

whose dimensions are given by $(m+1)p \times (m+1)p$ and $(m+1)p \times 1$, respectively. These quantities are well defined not only when $m > 0$ but also when $m = 0$, in which case $T_{m+1} = T_1 = T+1$, and hence $\mathbf{Q}_N(\mathcal{T}_0) = N^{-1}\sum_{t=1}^T\sum_{i=1}^N \tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,t}'$ and $\mathbf{R}_N(\mathcal{T}_0) = N^{-1}\sum_{t=1}^T\sum_{i=1}^N \tilde{\mathbf{x}}_{i,t}\tilde{y}_{i,t}$.

We also note how $\mathbf{Q}_N(\mathcal{T}_{T-1}) = \text{diag}(N^{-1}\sum_{i=1}^N \tilde{\mathbf{x}}_{i,1}\tilde{\mathbf{x}}_{i,1}', \ldots, N^{-1}\sum_{i=1}^N \tilde{\mathbf{x}}_{i,T}\tilde{\mathbf{x}}_{i,T}')$ and $\mathbf{R}_N(\mathcal{T}_{T-1}) = [N^{-1}\sum_{i=1}^N \tilde{\mathbf{x}}_{i,1}'\tilde{y}_{i,1}, \ldots, N^{-1}\sum_{i=1}^N \tilde{\mathbf{x}}_{i,T}'\tilde{y}_{i,T}]'$. In this notation,

$$\dot{\mathbf{B}}_T = \begin{bmatrix} \dot{\boldsymbol{\beta}}_1 \\ \vdots \\ \dot{\boldsymbol{\beta}}_T \end{bmatrix} = \begin{bmatrix} \dot{\boldsymbol{\beta}}_1(\mathcal{T}_{T-1}) \\ \vdots \\ \dot{\boldsymbol{\beta}}_T(\mathcal{T}_{T-1}) \end{bmatrix} = \mathbf{Q}_N(\mathcal{T}_{T-1})^{-1}\mathbf{R}_N(\mathcal{T}_{T-1}). \tag{2.9}$$

The proposed PDL2S estimator of $\mathbf{B}_T^0$ is given by

$$\hat{\mathbf{B}}_T = \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \vdots \\ \hat{\boldsymbol{\beta}}_T \end{bmatrix} = \arg\min_{\mathbf{B}_T} \ell_\gamma(\mathbf{B}_T), \tag{2.10}$$

where the dependence on γ here is suppressed for notational simplicity. For a given $\hat{\mathbf{B}}_T$, the set of estimated breaks is given by $\hat{\mathcal{T}}_{\hat{m}} = \{\hat{T}_1, \ldots, \hat{T}_{\hat{m}}\}$, where $\hat{T}_1 < \ldots < \hat{T}_{\hat{m}}$ for $\hat{m} > 0$ are such that $\|\hat{\boldsymbol{\beta}}_t - \hat{\boldsymbol{\beta}}_{t-1}\| \neq 0$ for $t = \hat{T}_1, \ldots, \hat{T}_{\hat{m}}$. If $\|\hat{\boldsymbol{\beta}}_t - \hat{\boldsymbol{\beta}}_{t-1}\| = 0$ for all $t = 1, \ldots, T$, then $\hat{m} = 0$ and $\hat{\mathcal{T}}_{\hat{m}} = \hat{\mathcal{T}}_0 = \varnothing$. We also define $\hat{T}_0 = 1$ and $\hat{T}_{\hat{m}+1} = T + 1$. The set $\hat{\mathcal{T}}_{\hat{m}}$ divides the sample into $\hat{m} + 1$ regimes such that the parameter estimates remain constant within each regime. The proposed estimator $\hat{\mathbf{A}}_{\hat{m}}$ of $\mathbf{A}_{m^0}^0$ is obtained by PDL2S, which is regime-by-regime LS conditional on $\hat{\mathcal{T}}_m$.[4] In terms of the notation introduced earlier,

$$\hat{\mathbf{A}}_{\hat{m}} = \begin{bmatrix} \hat{\boldsymbol{\alpha}}_1 \\ \vdots \\ \hat{\boldsymbol{\alpha}}_{\hat{m}+1} \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{\alpha}}_1(\hat{\mathcal{T}}_{\hat{m}}) \\ \vdots \\ \hat{\boldsymbol{\alpha}}_{\hat{m}+1}(\hat{\mathcal{T}}_{\hat{m}}) \end{bmatrix} = \mathbf{Q}_N(\hat{\mathcal{T}}_{\hat{m}})^{-1}\mathbf{R}_N(\hat{\mathcal{T}}_{\hat{m}}), \tag{2.11}$$

where the dependence on $\hat{\mathcal{T}}_{\hat{m}}$ and γ is again suppressed.

Remark 2. Tibshirani et al. (2005) propose the fused Lasso, which penalizes the ℓ_1 norm of both the individual slope coefficients themselves and their differences. Our objective function, which is similar to the one in Li et al. (2016), and Qian and Su (2016), differs from the one used in fused Lasso. The main differences are; (i) the penalization is done by using the Frobenius norm, as opposed to the ℓ_1 norm, (ii) only the coefficient differences are penalized, and (iii) different weights w_t are assigned to different coefficient differences. The use of the Frobenius norm allows us to induce sparsity for the entire vector of differences $\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1}$, there is no reason to shrink the coefficients themselves to zero, and the weighting is necessary to achieve consistency. Because

[4] One can also use the regular Lasso estimator of $\mathbf{A}_{m^0}^0$, as given by $[\hat{\boldsymbol{\beta}}_{\hat{T}_0}', \ldots, \hat{\boldsymbol{\beta}}_{\hat{T}_{\hat{m}}}']'$. However, as is well known in the literature, post-Lasso typically outperforms regular Lasso, and our (unreported) Monte Carlo results confirm this. In this paper, we therefore focus on post-Lasso LS.

of these differences, (2.5) can be seen as an adaptive group fused Lasso objective function (see Li et al., 2016, and Qian and Su, 2016, for discussions).

3 Assumptions and asymptotic results

3.1 Assumptions

The conditions that we will be working under are given in Assumptions EPS, LAM, Q, MOM and J. However, before we state these assumptions, we introduce some notation. Specifically, if $\mathbf{A}$ is a matrix, $\lambda_{min}(\mathbf{A})$ and $\lambda_{max}(\mathbf{A})$ signify its smallest and largest eigenvalues, respectively, tr $\mathbf{A}$ signifies its trace, and $\|\mathbf{A}\| = \sqrt{\text{tr}\,\mathbf{A}'\mathbf{A}}$ signifies its Frobenius norm. If $\mathbf{B}$ is also a matrix, then $\text{diag}(\mathbf{A}, \mathbf{B})$ denotes the block-diagonal matrix that takes $\mathbf{A}$ ($\mathbf{B}$) as the upper left (lower right) block. The symbols $\to_d$, $\to_p$ and $MN(\cdot, \cdot)$ signify convergence in distribution, convergence in probability and a mixed normal distribution, respectively. We use w.p.1 (w.p.a.1) to denote with probability (approaching) one. $\mathcal{C}$ denotes the sigma-field generated by $(\mathbf{f}_1', \ldots, \mathbf{f}_T')'$.

Assumption EPS.

(a) $\varepsilon_{i,t}$ is conditionally independent across i given $\mathcal{C}$ with $\mathbb{E}(\varepsilon_{i,t}|\mathcal{C}) = 0$ w.p.1;

(b) $\varepsilon_{i,t}$ is independent of $\mathbf{x}_{j,s}$ for all i, j, t and s.

Assumption LAM.

(a) $\boldsymbol{\lambda}_i = \boldsymbol{\lambda} + \boldsymbol{\nu}_i$, where $\boldsymbol{\nu}_i$ is conditionally independent across i given $\mathcal{C}$ with $\mathbb{E}(\boldsymbol{\nu}_i|\mathcal{C}) = \mathbf{0}_{r\times 1}$ w.p.1;

(b) $\boldsymbol{\nu}_i$ is independent of $(\mathbf{x}_{j,t}, \varepsilon_{j,t})$ for all i, j and t.

Assumption Q.

(a) $\inf_{\mathcal{T}_m} \lambda_{min}[\mathbf{Q}_N(\mathcal{T}_m)] > 0$ w.p.1;

(b) $\mathbf{Q}_N(\mathcal{T}_m) \to_p \mathbf{Q}_0(\mathcal{T}_m) = \lim_{N\to\infty} \mathbb{E}[\mathbf{Q}_N(\mathcal{T}_m)|\mathcal{C}]$ as $N \to \infty$, where $\infty > \inf_{\mathcal{T}_m} \lambda_{min}[\mathbf{Q}_0(\mathcal{T}_m)] > 0$ w.p.1.

Assumption MOM. $\mathbb{E}\|\tilde{\mathbf{x}}_{i,t}\|^4 < \infty$, $\|\mathbf{f}_t\| < \infty$ w.p.1, $\mathbb{E}\|\mathbf{f}_t\|^4 < \infty$, $\mathbb{E}\|\boldsymbol{\nu}_i\|^4 < \infty$ and $\mathbb{E}\varepsilon_{i,t}^4 < \infty$ for all i and t.

Assumption J.

(a) $J_{max} = \max_{1\leq j\leq m^0+1} \|\boldsymbol{\alpha}_{j+1}^0 - \boldsymbol{\alpha}_j^0\| = O(1)$;

(b) $\sqrt{N}\gamma J_{min}^{-\kappa} \to c_1 \in [0,\infty)$ and $\sqrt{N}J_{min} \to c_2 \in (0,\infty]$ as $N \to \infty$, where $J_{min} = \min_{1\leq j\leq m^0+1} \|\boldsymbol{\alpha}_{j+1}^0 - \boldsymbol{\alpha}_j^0\|$;

(c) $N^{(\kappa+1)/2}\gamma \to \infty$ as $N \to \infty$.

Some comments are in order. Consider Assumption EPS. Many papers in the literature assume that $\varepsilon_{i,t}$ is (conditionally) independent over i (see, for example, Ahn et al., 2013, Moon and Weidner, 2015, Pesaran, 2006, Robertson and Sarafidis, 2015, and Westerlund et al., 2019), and so do we. Independence is not necessary, though, and can be relaxed to allow for weak cross-sectional dependence at the expense of additional high-level moment conditions (as in Bai, 2009). This is demonstrated in Section 4, where we use Monte Carlo simulations to investigate the effect of error cross-section dependence. The assumption that $\varepsilon_{i,t}$ is independent of $\mathbf{x}_{i,t}$, which is the same as in, for example, Bai (2009), Moon and Weidner (2015), Pesaran (2006), and Westerlund et al. (2019), is necessary and cannot be easily dispensed with. Qian and Su (2016) allow for endogenous regressors by using GMM (see also Ahn et al., 2013, and Robertson and Sarafidis, 2015, in absence of breaks), and in the empirical illustration of Section 5 we consider a version of this estimator. However, GMM requires that valid external instruments are available, which is not always the case in practice. Moreover, the condition that $\varepsilon_{i,t}$ is independent of $\mathbf{x}_{i,t}$ does not rule out endogeneity, as $\mathbf{x}_{i,t}$ can still be correlated with $\mathbf{f}_t$. The heteroskedasticity and serial correlation properties of $\varepsilon_{i,t}$ are not restricted in any way.

Assumption LAM is a random coefficient condition. It demands that $\boldsymbol{\lambda}_i$ is randomly distributed with constant mean, and that it is independent of $\mathbf{x}_{i,t}$ and $\varepsilon_{i,t}$, which are standard requirements in the common correlated effects (CCE) strand of the literature (see Westerlund et al., 2019, for an overview). As mentioned in Section 2, because of the demeaning, the PDL2S

estimator is exactly invariant with respect to $\boldsymbol{\lambda}_i'\mathbf{f}_t$ when $\boldsymbol{\lambda}_i = \boldsymbol{\lambda}$ for all i, so that $\boldsymbol{\lambda}_i'\mathbf{f}_t = \boldsymbol{\lambda}'\mathbf{f}_t$ reduces to a common time effect. Assumption LAM ensures that the PDL2S estimator is consistent and asymptotically mixed normal even in cases when $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_N$ are not all equal.[5] To put this into perspective, Boldea et al. (2020) assume that $N^{-1}\sum_{i=1}^{N}\mathbf{x}_{i,t}\boldsymbol{\lambda}_i'\mathbf{f}_t \to_p \mathbf{a}_j$ as $N \to \infty$ for all $t = T_{j-1}, \ldots, T_j - 1$ and $j = 1, \ldots, m+1$, so that asymptotically the sample cross-moment of the regressors and the interactive effects is constant within break regimes.[6] By contrast, under Assumption LAM, $N^{-1}\sum_{i=1}^{N}\mathbf{x}_{i,t}\boldsymbol{\lambda}_i'\mathbf{f}_t = N^{-1}\sum_{i=1}^{N}\mathbf{x}_{i,t}\boldsymbol{\lambda}'\mathbf{f}_t + o_p(1)$, which may vary freely over t.

Assumption Q is a non-collinearity condition that rules out cross-section-invariant regressors in $\mathbf{x}_{i,t}$. This is the same as the usual time fixed effects-only condition. The simplicity and transparency of this condition is a great advantage when compared to studies such as Bai (2009), and Moon and Weidner (2015), where the factors are estimated and the regressors are de-factored, as opposed to just demeaned. As a result, general "low-rank" regressors have to be ruled out in order to ensure that the de-factored regressors have enough variation.[7] The problem is that the ruled out low-rank regressors depend on $\boldsymbol{\lambda}_i$ and $\mathbf{f}_t$, which are unknown to the researcher. There is therefore a risk that the defactoring exhausts too much variation, causing the signal matrix to become (near) singular. This is particularly true in the type of small-T (microeconomic) panels that we have in mind where many regressors have low variation.

Assumption MOM supposes that $\tilde{\mathbf{x}}_{i,t}$, $\mathbf{f}_t$, $\boldsymbol{\nu}_i$ and $\varepsilon_{i,t}$ have a certain number of finite moments. Four finite moments are required for $\tilde{\mathbf{x}}_{i,t}$, which is a standard condition. This condition together with the non-collinearity condition in Assumption Q, and the independence of $\varepsilon_{i,t}$ and $\boldsymbol{\nu}_i$ in Assumptions EPS and LAM are the only conditions placed on the regressors. This is different from the CCE strand of the literature where it is standard to assume that $\mathbf{x}_{i,t}$ has a common factor structure that loads on the same factors as $u_{i,t}$ (see Pesaran, 2006, and Westerlund et al., 2019), which is restrictive in itself but also because it rules out models involving, for example,

[5] The condition that $\boldsymbol{\lambda}_i$ and $\mathbf{x}_{i,t}$ are uncorrelated is testable and has been subject to some scrutiny in the recent empirical literature (see, for example, Kapetanios et al., 2019, and Petrova and Westerlund, 2020). The evidence is favourable.

[6] The need for this condition is partly expected given the discussion in Section 1 on the difficulty of separating the breaks from the interactive effects. Boldea et al. (2020) do not do anything to control for the interactive effects but apply LS as if there were no effects present at all. This means that they have to put enough structure on the effects so as to ensure that they do not interfere with their break estimation procedure. One of the terms in the resulting omitted interactive effects bias of the LS estimator is given by $N^{-1}\sum_{i=1}^{N}\mathbf{x}_{i,t}\boldsymbol{\lambda}_i'\mathbf{f}_t$. If this is not constant within break regimes, the interactive effects will be mistaken for structural breaks.

[7] Certain low-rank regressors can be permitted but they then require special treatment (see Bai, 2009).

powers or products of the regressors. Boldea et al. (2020), and Qian and Su (2016) assume that the regressors are independent across the cross-section, which is even more restrictive.[8] In the present paper, $\mathbf{x}_{i,t}$ does not have a factor structure, nor does it have to be independent. In fact, $\mathbf{x}_{i,t}$ does not even have to be stochastic, but can also contain deterministic terms such as dummy variables. And those regressors that are stochastic can be arbitrarily correlated across both time and cross-section. The same is true for $\mathbf{f}_t$, which is almost completely without restriction. Note in particular that there are no conditions on the number of factors, r, provided that it is fixed. This is different from most CCE studies where r is bounded from above by the number of observables, $p+1$ (see, for example, Westerlund et al., 2019). Moreover, unlike in most GMM- and principal components-based studies, the proposed PDL2S estimator does not depend on the availability of a consistent estimator of r (see, for example, Ahn et al., 2013, Bai, 2009, and Robertson and Sarafidis, 2015).

Assumption J imposes some conditions on the tuning parameter γ and the size of the breaks, and are easy to justify. For example, if we assume that all the breaks are bounded away from zero and infinity, then Assumption J requires that $\gamma = O(N^{-(1+\delta)/2})$ with $\delta \in [0, \kappa)$. One way to satisfy Assumption J is therefore to set γ proportional to $N^{-1/2}$ (as in, for example, Belloni et al., 2016, and Hansen and Liao, 2019). We also note that the breaks do not have to be bounded away from zero and hence that some, or indeed all, breaks may be shrinking to zero. The breaks therefore do not have to be "large" for our procedure to be able to detect them, which is reassuring.

3.2 Asymptotic results

Our first main result characterizes the limit of $\hat{\boldsymbol{\beta}}_t$.

Theorem 1. *Suppose that Assumptions EPS, LAM, Q, MOM and J hold. Then, uniformly in* $t \in \mathcal{T}_{T-1}$,

$$\|\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^0\| = O_p(N^{-1/2}).$$

Theorem 1 establishes that the PDL2S estimator is consistent and that the rate of convergence

[8] The condition that the regressors are identically distributed can be relaxed (see Boldea et al., 2020). However, it is still necessary that the sample second moment matrix of the regressors is asymptotically time-invariant (within break regimes).

is given by $N^{-1/2}$, which is the highest possible rate for the type of parametric fixed-T panel data models that we consider. In Lemma A.1 of the online appendix, we show that the preliminary period-by-period LS estimator, $\dot{\boldsymbol{\beta}}_t$, is consistent at the same rate, which is just as expected because T is fixed. Hence, from a rate of convergence point of view, nothing is gained by using PDL2S. However, the preliminary estimator does not account for the fact that the slopes are constant within break regimes. It is therefore not as efficient as PDL2S. It is also completely uninformative regarding the number of breaks and their location. This brings us to our second main result.

Theorem 2. *Suppose that Assumptions EPS, LAM, Q, MOM, and J hold. Then, as $N \to \infty$,*

$$\mathbb{P}(\|\hat{\boldsymbol{\beta}}_t - \hat{\boldsymbol{\beta}}_{t-1}\| = 0 \text{ for all } t \in \mathcal{T}_{m^0}^{0c} = \mathcal{T}_{T-1} \setminus \mathcal{T}_{m^0}^{0}) \to 1.$$

The set $\mathcal{T}_{m^0}^{0c}$ is the complement of $\mathcal{T}_{m^0}^{0}$. Hence, since $\boldsymbol{\beta}_t^0$ is constant within break regimes, we have that $\boldsymbol{\beta}_t^0 = \boldsymbol{\beta}_{t-1}^0$ for all $t \in \mathcal{T}_{m^0}^{0c}$. Theorem 2 states that $\hat{\boldsymbol{\beta}}_t - \hat{\boldsymbol{\beta}}_{t-1}$ is strongly consistent for $\boldsymbol{\beta}_t^0 - \boldsymbol{\beta}_{t-1}^0$ when $t \in \mathcal{T}_{m^0}^{0c}$, which is a reflection of the usual sparsity result in the variable selection literature (see, for example, Fan and Li, 2006). But from Theorem 1, we know that $\hat{\boldsymbol{\beta}}_t - \hat{\boldsymbol{\beta}}_{t-1}$ is consistent for all t, including $t \in \mathcal{T}_{m^0}^{0}$. This means that PDL2S is able to identify the true model in (2.1) with the correct number of breaks and break dates. The following corollary to Theorems 1 and 2 formalizes this.

Corollary 1. *Suppose that Assumptions EPS, LAM, Q, MOM and J hold. Then, as $N \to \infty$,*

(a) $\mathbb{P}(\hat{m} = m^0) \to 1$;

(b) $\mathbb{P}(\hat{\mathcal{T}}_{\hat{m}} = \mathcal{T}_{m^0}^{0} | \hat{m} = m^0) \to 1$.

Theorem 3 reports the asymptotic distribution of the PDL2S estimator, and it does so conditional on the high probability even that $\hat{m} = m^0$.

Theorem 3. *Suppose that Assumptions EPS, LAM, Q, MOM and J hold, and that $\hat{m} = m^0$. Then, as $N \to \infty$,*

$$\sqrt{N}(\hat{\mathbf{A}}_{\hat{m}} - \mathbf{A}_{m^0}^{0}) \to_d MN(\mathbf{0}_{(m^0+1)p \times 1}, \mathbf{Q}_0^{-1} \boldsymbol{\Omega}_0 \mathbf{Q}_0^{-1}),$$

where

$$\mathbf{Q}_0 = \mathbf{Q}_0(\mathcal{T}^0_{m^0}),$$
$$\mathbf{\Omega}_0 = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}(\mathbf{u}_i\mathbf{u}_i'|\mathcal{C}),$$
$$\mathbf{u}_i = \mathbf{u}_i(\mathcal{T}^0_{m^0}) = \begin{bmatrix} \sum_{t=T^0_0}^{T^0_1-1} \tilde{\mathbf{x}}_{i,t}\tilde{u}_{i,t} \\ \vdots \\ \sum_{t=T^0_{m^0}}^{T^0_{m^0+1}-1} \tilde{\mathbf{x}}_{i,t}\tilde{u}_{i,t} \end{bmatrix}.$$

The definitions of $\mathbf{Q}_0$ and $\mathbf{\Omega}_0$ in Theorem 3 reveal that the PDL2S estimator is asymptotically equivalent to the infeasible LS estimator of (2.4) that takes all the breaks as known. In this sense, PDL2S is "oracle efficient". That being said, $\mathbf{\Omega}_0$ does depend on $\tilde{u}_{i,t}$, which is a function of $\tilde{\boldsymbol{\nu}}_i'\mathbf{f}_t$. Hence, while oracle efficient in the sense that it is asymptotically equivalent to the known break LS estimator, the PDL2S estimator is not asymptotically equivalent to the LS estimator that takes both the breaks and the factors as known. As pointed out in Section 2, the demeaning removes the mean of $\boldsymbol{\lambda}_i$, and this is enough to ensure $\sqrt{N}$-consistency and asymptotic (mixed) normality as long as $\varepsilon_{i,t}$ and $\boldsymbol{\nu}_i$ are uncorrelated with $\tilde{\mathbf{x}}_{i,t}$. However, this does not mean that the PDL2S estimator is asymptotically invariant with respect to $\boldsymbol{\lambda}_i'\mathbf{f}_t$, and Theorem 3 confirms this. In Section 4, we use Monte Carlo simulations as a means to investigate how the variance of the PDL2S estimator is affected by the interactive effects.

The asymptotic distribution of $\sqrt{N}(\hat{\mathbf{A}}_{\hat{m}} - \mathbf{A}^0_{m^0})$ is normal conditional on $\mathcal{C}$, which means that unconditionally it is mixed normal (see Andrews, 2005, for a discussion). The asymptotic distribution therefore supports standard normal and chi-squared inference. Of course, for such standard inference to be possible, we need a consistent estimator of $\mathbf{Q}_0^{-1}\mathbf{\Omega}_0\mathbf{Q}_0^{-1}$. Let us therefore define $\hat{u}_{i,t} = \tilde{y}_{i,t} - \tilde{\mathbf{x}}_{i,t}'\hat{\boldsymbol{\alpha}}_j$, where $t = \hat{T}_{j-1}, \ldots, \hat{T}_j - 1$ with $j = 1, \ldots, \hat{m}+1$. A natural estimator of $\mathbf{Q}_0^{-1}\mathbf{\Omega}_0\mathbf{Q}_0^{-1}$ given by

$$\mathbf{Q}_N(\hat{\mathcal{T}}_{\hat{m}})^{-1}\hat{\mathbf{\Omega}}\mathbf{Q}_N(\hat{\mathcal{T}}_{\hat{m}})^{-1} \tag{3.1}$$

where $\mathbf{Q}_N(\hat{\mathcal{T}}_{\hat{m}})$ is as before and the $(\hat{m}+1)p \times (\hat{m}+1)p$ matrix $\hat{\mathbf{\Omega}}$ is given by

$$\hat{\mathbf{\Omega}} = \frac{1}{N}\sum_{i=1}^{N} \hat{\mathbf{u}}_i\hat{\mathbf{u}}_i', \tag{3.2}$$

where

$$\hat{\mathbf{u}}_i = \hat{\mathbf{u}}_i(\hat{\mathcal{T}}_{\hat{m}}) = \begin{bmatrix} \sum_{t=\hat{T}_0}^{\hat{T}_1-1} \tilde{\mathbf{x}}_{i,t}\hat{u}_{i,t} \\ \vdots \\ \sum_{t=\hat{T}_m}^{\hat{T}_{m+1}-1} \tilde{\mathbf{x}}_{i,t}\hat{u}_{i,t} \end{bmatrix}. \tag{3.3}$$

The consistency of this estimator is a direct consequence of the consistency of $\hat{\mathbf{A}}_{\hat{m}}$, $\hat{m}$ and $\hat{\mathcal{T}}_{\hat{m}}$.

Corollary 2. *Suppose that Assumptions EPS, LAM, Q, MOM and J hold. Then, as $N \to \infty$,*

$$\mathbf{Q}_N(\hat{\mathcal{T}}_{\hat{m}})^{-1}\hat{\mathbf{\Omega}}\mathbf{Q}_N(\hat{\mathcal{T}}_{\hat{m}})^{-1} \to_p \mathbf{Q}_0^{-1}\mathbf{\Omega}_0\mathbf{Q}_0^{-1}.$$

Remark 3. A major point about Corollary 2 is that the asymptotic covariance matrix of the PDL2S estimator is very easily estimable. This stands in sharp contrast to the large-T framework that typically involves some kind of heteroskedasticity and autocorrelation consistent (HAC) correction (see, for example, Bai, 2009, and Pesaran, 2006), which is not only difficult to implement but also known to lead to poor small-sample properties.

Consider testing the null hypothesis of $H_0 : \mathbf{R}\mathbf{A}_{m^0}^0 = \mathbf{r}$, where $\mathbf{R}$ is a $q \times (m^0+1)p$ matrix of rank $q \leq (m^0+1)p$ and $\mathbf{r}$ is a $q \times 1$ vector. Again, conditional on the high probability even that $\hat{m} = m^0$, the relevant Wald test statistic is given by

$$W = N(\mathbf{R}\hat{\mathbf{A}}_{\hat{m}} - \mathbf{r})'[\mathbf{R}\mathbf{Q}_N(\hat{\mathcal{T}}_{\hat{m}})^{-1}\hat{\mathbf{\Omega}}\mathbf{Q}_N(\hat{\mathcal{T}}_{\hat{m}})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\mathbf{A}}_{\hat{m}} - \mathbf{r}). \tag{3.4}$$

Suppose that H_0 is true. Then, because of Theorem 3 and Corollary 2,

$$W = \sqrt{N}(\mathbf{R}\hat{\mathbf{A}}_{\hat{m}} - \mathbf{r})'(\mathbf{R}\mathbf{Q}_0^{-1}\mathbf{\Omega}_0\mathbf{Q}_0^{-1}\mathbf{R}')^{-1}\sqrt{N}(\mathbf{R}\hat{\mathbf{A}}_{\hat{m}} - \mathbf{r}) + o_p(1) \to_d \chi^2(q) \tag{3.5}$$

as $N \to \infty$. Similarly, if $q = 1$, then the t-statistic

$$t = \frac{\sqrt{N}(\mathbf{R}\hat{\mathbf{A}}_{\hat{m}} - \mathbf{r})}{\sqrt{\mathbf{R}\mathbf{Q}_N(\hat{\mathcal{T}}_{\hat{m}})^{-1}\hat{\mathbf{\Omega}}\mathbf{Q}_N(\hat{\mathcal{T}}_{\hat{m}})^{-1}\mathbf{R}'}} \tag{3.6}$$

has a limiting $N(0,1)$ distribution under H_0.

All the estimates considered so far are conditional on the tuning parameter γ. While in theory any choice satisfying Assumption J will do, as with most other tuning parameters, in practice the results can be sensitive to different specifications of γ. It might therefore be preferable to set

this parameter in a data-driven fashion. In this paper, we follow Li et al. (2016), and Qian and Su (2016), and set γ by minimizing an information criterion;

$$\hat{\gamma} = \arg\min_{\gamma} \mathrm{IC}(\gamma), \tag{3.7}$$

with

$$\mathrm{IC}(\gamma) = \hat{\sigma}^2(\hat{\mathcal{T}}_{\hat{m}}(\gamma)) + \phi \cdot p[\hat{m}(\gamma) + 1], \tag{3.8}$$

where $\hat{\mathcal{T}}_{\hat{m}}(\gamma)$ and $\hat{m}(\gamma)$ are $\hat{\mathcal{T}}_{\hat{m}}$ and $\hat{m}$, respectively, when treated as functions of γ, $\phi = \phi(N) > 0$ is a penalty, and

$$\hat{\sigma}^2(\mathcal{T}_m) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{m+1} \sum_{t=T_{j-1}}^{T_j - 1} (\tilde{y}_{i,t} - \tilde{\mathbf{x}}_{i,t}' \hat{\boldsymbol{\alpha}}_j)^2. \tag{3.9}$$

Theorem 4. *Suppose that Assumptions EPS, LAM, Q, MOM and J hold, that $\phi \to 0$ and that $N\phi \to \infty$. Then, as $N \to \infty$,*

$$\mathbb{P}[\hat{m}(\hat{\gamma}) = m^0] \to 0.$$

As usual, the penalty ϕ is not unique and has to be set by the researcher. Analogous to Qian and Su (2016), in this paper we set $\phi = (\ln N)/N$, which makes $\mathrm{IC}(\gamma)$ similar to the conventional Schwarz Bayesian information criterion (BIC).

4 Monte Carlo simulations

4.1 Setup

In this section, we use Monte Carlo simulations as a means to evaluate the finite sample properties of the proposed PDL2S approach. The data generating process used for this purpose is given by a restricted version of (2.1) and (2.2) that sets $p = 4$ and $r = 5$. Similarly to Qian and Su (2016), we consider $m^0 \in \{0, 1, 2\}$ with $\boldsymbol{\beta}_t = \mathbf{0}_{p\times 1}$ when $m^0 = 0$, $\boldsymbol{\beta}_t = \mathbf{1}_{p\times 1} \cdot 1(T/2 \leq t \leq T)$ when $m^0 = 1$ and $\boldsymbol{\beta}_t = \mathbf{1}_{p\times 1} \cdot [1(\lfloor T/3 \rfloor \leq t < \lfloor 2T/3 \rfloor) + 2 \cdot 1(\lfloor 2T/3 \rfloor \leq t \leq T)]$ when $m^0 = 2$, where $1(\cdot)$ and $\lfloor \cdot \rfloor$ are the indicator and integer part functions, respectively, and $\mathbf{0}_{p\times 1}$ ($\mathbf{1}_{p\times 1}$) is a $p \times 1$ vector of zeroes (ones).

If $u_{i,t}$ and $\mathbf{x}_{i,t}$ are independent, the interactive effects in (2.1) can be ignored without risking the consistency of the regular post-Lasso LS estimator based on raw (non-demeaned) data. Hence, in order to make $u_{i,t}$ and $\mathbf{x}_{i,t}$ dependent, we allow $\mathbf{x}_{i,t}$ to load on the same set of factors as $u_{i,t}$, which as pointed out in Section 3.1 is a requirement in CCE. Specifically, $\mathbf{x}_{i,t}$ is generated according to the following factor model:

$$\mathbf{x}_{i,t} = \mathbf{\Gamma}_i \mathbf{f}_t + \boldsymbol{\nu}_{i,t}, \tag{4.1}$$

where

$$\mathbf{f}_t = (1-\varphi) + \varphi \mathbf{f}_{t-1} + \boldsymbol{\eta}_t, \tag{4.2}$$

with $\mathbf{f}_0 = \mathbf{0}_{r\times 1}$, $\varphi \in \{0.8, 1\}$ and $\boldsymbol{\eta}_t \sim N(\mathbf{0}_{r\times 1}, \mathbf{I}_r)$. Hence, while stationary (although highly persistent) when $\varphi = 0.8$, when $\varphi = 1$, $\mathbf{f}_t$ is unit root non-stationary. The data generating process considered for $\boldsymbol{\nu}_{i,t}$ is also very general and is the same as in, for example, Petrova and Westerlund (2020). It is given by

$$\boldsymbol{\nu}_{i,t} = \pi \boldsymbol{\nu}_{i,t-1} + \mathbf{e}_{i,t} + \sum_{j=1}^{K} \pi(\mathbf{e}_{i-j,t} + \mathbf{e}_{i+j,t}), \tag{4.3}$$

where $\boldsymbol{\nu}_{1,0} = ... = \boldsymbol{\nu}_{N,0} = \mathbf{0}_{p\times 1}$, $\pi \in \{0.4, 0.8\}$, $K = 10$ and $\mathbf{e}_{i,t} \sim N(\mathbf{0}_{p\times 1}, \mathbf{I}_p)$. This means that $\boldsymbol{\nu}_{i,t}$ is weakly correlated over time as well as with $2K$ of its neighbouring cross-sectional units.[9] If $\pi = 0.4$, we say that the error dependence is "low", whereas if $\pi = 0.8$ the error dependence is said to be "high". A similar process is used for generating $\varepsilon_{i,t}$;

$$\varepsilon_{i,t} = \pi \varepsilon_{i,t-1} + \xi_{i,t} + \sum_{j=1}^{K} \pi(\xi_{i-j,t} + \xi_{i+j,t}), \tag{4.4}$$

where $\varepsilon_{1,0} = ... = \varepsilon_{N,0} = 0$ and $\xi_{i,t} \sim N(0, \sigma_i^2)$ with $\sigma_i^2 \sim U(0.5, 1)$. Hence, $\varepsilon_{i,t}$ is not only weakly serially and cross-sectionally correlated but also heteroskedastic. Finally, to ensure that Assumption LAM is met the loadings in $\mathbf{\Gamma}_i$ and $\boldsymbol{\lambda}_i$ are drawn independently from $N(2, 1)$.

As for the sample size, we consider all combinations of $N \in \{25, 50, 100, 200, 600\}$ and $T \in \{5, 10, 20\}$, where the values considered for T are intentionally smaller than those considered for N.

[9] The cross-sectional sum in $\boldsymbol{\nu}_{i,t}$ is truncated at beginning and end when not enough cross-sections are available. For example, when generating $\nu_{1,t}$, the sum only includes $\mathbf{e}_{2,t}, ..., \mathbf{e}_{11,t}$.

We report four performance measures, the frequency of false detection of the estimated number of breaks, the frequency of false detection of the estimated breakpoints given that the number of breaks is selected correctly, the average number of estimated breaks, and the mean squared error (MSE) of the PDL2S estimator (times 100), computed as the average $\|\hat{\mathbf{A}}_{\hat{m}} - \mathbf{A}^0_{m^0}\| / (\hat{m}+1)p$ across the Monte Carlo replications, whose number is here set to 1,000.

The estimation code was written in Python, which is one of the most common programming language in applications of the Lasso. Following the previous literature on the adaptive Lasso (see Qian and Su, 2016), we set $\kappa = 2$. For a given value γ, we optimize (2.5) using the convex optimization package CVXPY. We then determine the most appropriate value of γ by minimizing the information criterion in (3.8). To accomplish this, we need to choose a suitable grid containing values of γ that yield the true breaks. One way to do so is to first select an interval $[\gamma_{max}, \gamma_{min}]$, where γ_{min} (γ_{max}) is chosen so that the number of estimated breaks is zero ("many") (see Qian and Su, 2016). We then slice $[\gamma_{max}, \gamma_{min}]$ into 50 evenly sized intervals on a log-scale, optimize (2.5) at each value and select as $\hat{\gamma}$ the value that minimizes the information criterion in (3.8).

The simulations are too time consuming for a personal computer. We used the UPPMAX (Uppsala Multidisciplinary Center for Advanced Computational Science) cluster Rackham, which is accessible via the SNIC (Swedish National Infrastructure for Computing). Rackham consists of 486 nodes, each containing two 10-core Intel Xeon V4 central processing units.

4.2 Results

Tables 1–5 contain the results, which are reported for different constellations of φ and π. We consider four cases; (i) stationary factors ($\varphi = 0.8$) and low error dependence ($\pi = 0.4$), (ii) non-stationary factors ($\varphi = 1$) and low error dependence ($\pi = 0.4$), (iii) stationary factors ($\varphi = 0.8$) and high error dependence ($\pi = 0.8$), and (iv) non-stationary factors ($\varphi = 1$) and high error dependence ($\pi = 0.8$).

INSERT TABLES 1–5 ABOUT HERE

We begin by considering the results reported in Table 1 for the case with stationary factors and low error dependence. The first thing to note is that PDL2S does very well even when N

and T are as small as $N = 25$ and $T = 5$. In fact, in this case the breaks are estimated perfectly. The only exception is when $m = 2$, in which case the number of breaks is not always estimated correctly. However, this is only a small-sample effect that goes away with increasing values of N and T. The breakpoints are always estimated correctly. Hence, as expected given Corollary 1, the PDL2S approach is robust to the number of breaks, and works very well even if there are no breaks at all. The MSE decreases with increasing values of N and T. The effect of N is anticipated and is a reflection of the consistency of the PDL2S estimator (Theorem 1). The improvement that comes from increasing T cannot be explained by our theoretical results, which are silent about the effect of T. It suggests that T does not have to be "small" but that the estimator works well also when T is relatively large.

As expected given the unrestricted specification of the factors, increasing their persistence from $\varphi = 0.8$ to $\varphi = 1$ has no effect on the results. This is clear from comparing the results reported in Table 1 with those reported in Table 2. In the literature it is common to assume that the factors are stationary (see, for example, Bai, 2009, and Pesaran, 2006), which rules out factors that are, for example, breaking or trending. This may be justified in some applications, but certainly not in general. The fact that the performance of the PDL2S estimator is unaffected by the specification of the factors is therefore a great advantage.

High error dependence generally leads to worse performance than if the dependence is low, as is evident by comparing the results reported in Tables 1 and 3. Note in particular how the gain in performance that comes from increasing N is relatively slow when the dependence is high, which is partly expected given the high level of error cross-section correlation in this case. The effect is not detrimental, though, and so performance is still acceptable. Hence, as discussed in Section 3, error cross-section independence is not necessary. This is true not only when the factors are stationary, but also when they are non-stationary, as they are in Table 4.

The results reported so far are for the proposed PDL2S estimator. In order to investigate the importance of the demeaning as a means to account for the interactive effects, in Table 5 we report some results for the post-Lasso LS estimator when applied to raw (non-demeaned) data. As in Table 4, the factors are non-stationary and the error dependence is high. As expected given the presence of $\mathbf{f}_t$ in the equations for both $\mathbf{x}_{i,t}$ and $u_{i,t}$, failure to demean leads to a massive loss of

performance. Not only are the results reported in Table 5 uniformly worse than those reported in Table 4 but there is also no improvement as the sample size increases, suggesting that the estimator is inconsistent, as expected.

All-in-all, we find that the PDL2S estimator performs very well in the type of small-T panels considered, and that it does so under a wide range of empirically relevant scenarios. It should therefore be an attractive alternative to the already existing menu of estimators of panel regression models with possible interactive effects and breaks.

5 Empirical illustration

5.1 Motivation

There is a large and growing empirical literature concerned with the socioeconomic determinants of crime. Usual suspects include deterrence variables capturing the probabilities of apprehension and punishment, and variables that control for the relative rate of return of legal opportunities. One of the main conclusions from this literature is that aggregate data do not provide much support of the deterrence idea that policy can reduce crime by raising expected costs (see, for example, Dills et al., 2010).

One of the most widely held explanations for this lack of empirical support is the presence of unobserved heterogeneity, which, unless appropriately accounted for, may well render the LS estimator biased and inconsistent. Cornwell and Trumbull (1994) were among the first to make this point. According to them, the issue of unobserved heterogeneity cannot be ignored, and there are by now plenty of research that confirms this (see, for example, Bushway et al., 1999, Cherry and List, 2002, and Worrall and Pratt, 2004). The following quotation, taken from Nagin and Paternoster (2000, page 131), illustrates the issue: "There is also a critical substantive reason for employing models that control for unobserved heterogeneity. If there is unobserved heterogeneity that accounts for offending over time, the failure to explicitly consider this will lead to biased estimates of observed time-varying factors [...] Unobserved heterogeneity is like any other omitted variable, it will result in biased estimates of other parameters in the model, such as prior offending or delinquent peers (see Bushway et al., 1999). The practical consequence is that the estimated effect of time-varying variables that reflect state dependence will be inflated."

Another explanation is that while most theories of crime are about the behaviour of individuals, many studies use aggregated data, usually at the state or country level, even though there is by now plenty of evidence to suggest that individual (crime) behaviour is not well preserved under aggregation. For instance, Lott and Mustard (1997) use both state- and county-level data in their study of the effect of concealed handgun laws on violent crime in the US. One of their main conclusions is that "the very different results between state- and county-level data should make us very cautious in aggregating crime data and would imply that the data should remain as disaggregated as possible" (page 39).

Yet another explanation for the lack of support of the deterrence idea is the presence of structural breaks. According to McDowall and Loftin (2005, page 359), "[c]onventional explanations of crime rate trends assume that changes in the rates follow a process that is linear and constant [...] Questioning the conventional assumptions, an emerging class of historical contingency theories stresses variation in the crime-generating mechanism. According to contingency explanations, the process underlying the rates [...] has a structure that shifts over time." The concern is that failure to control for such shifts is likely to result in inconsistent estimates of the model parameters.

Although the presence of unobserved heterogeneity and structural breaks have been more or less ignored in most studies, some attempts have been made to obtain at least a partial solution. Cornwell and Trumbull (1994) use data on 90 counties in North Carolina between 1981 and 1987. The fact that their data set has a panel structure makes it possible to control for certain types of unobserved heterogeneity while at the same time maintaining a relatively low level of aggregation. The main conclusion is that the estimated deterrence effect is highly sensitive to the treatment of unobserved heterogeneity, and hence that researchers "should no longer disregard this important source of specification error" (page 366). By contrast, studies such as Batton and Jensen (2002), and Carlson and Michalowski (1997) employ aggregate time series data that they split into subperiods based on major events in order to account for structural change.

Of course, while potentially quite useful by themselves, these solutions are bound to be inadequate in any application that is characterized by both unobserved heterogeneity and structural breaks. One possibility is to use panel data to account for unobserved heterogeneity and to slice

up the sample period to account for breaks. But then this means that the breaks are treated as known, which is risky, as misplaced breaks are just as problematic as omitted breaks. This is important, as there is usually great uncertainty over both the number of breaks and their location. As an example, Batton and Jensen (2002) used the Chow test to test for the presence of breaks at given dates. They urged caution in interpreting their test results, since almost every potential breakpoint was found to be significant.

The discussion of the last paragraph suggests that there is a need for an approach that is general enough to accommodate not only unobserved heterogeneity but also structural breaks. The PDL2S estimator fits this bill and we will therefore use it in this empirical illustration.

5.2 Main results

The data that we will use are the same as in Cornwell and Trumbull (1994) (see also Baltagi, 2006, and Baltagi and Liu, 2009).[10] Hence, in this illustration, $N = 90$ and $T = 7$, which means that it is important to use techniques that do not require T to be large. This is another reason for considering PDL2S.

The included deterrence variables, which are standard in the literature, are the probability of arrest (PRBARR), the probability of conviction given arrest (PRBCONV), the probability of a prison sentence given a conviction (PRBPRIS), the average prison sentence in days (AVGSEN), and the number of police per capita (POLPC). In addition to the deterrence variables, the data set contains a number of wage variables that are intended to capture opportunities in the legal sector. These are the average weekly wages in construction (WCON), transportation, utilities and communication (WTUC), wholesale and retail trade (WTRD), finance, insurance and real estate (WFIR), services (WSER), manufacturing (WMFG), federal government (WFED), state government (WSTA) and local government (WLOC). Population density (DENSITY), and percent young male (PCTYMLE) are also included, as crime tends to depend on these. All-in-all there are $p = 16$ regressors, which are again similar to those considered previously in the literature (see, for example, Ghasemi, 2017, and the references provided therein). All regressors are trans-

[10]The data can be downloaded on-line from the Journal of Applied Econometrics data archive, available at http://qed.econ.queensu.ca/jae/.

formed into logs, as is the crime rate (CRMRTE).[11]

INSERT TABLE 6 ABOUT HERE

The PDL2S estimator is implemented as described in Section 4. We estimate $\hat{m}(\hat{\gamma}) = 5$ breaks in 1982, 1983, 1984, 1985 and 1986, which means that according to our break detection procedure the only year where there is no break is 1981. The resulting PDL2S results are reported in Table 6, which also contains the results for a model without breaks. The first thing to note is that PRBARR and PRBCONV are significant. The estimated effects of the former (latter) regressor vary quite substantially, from -0.681 (-0.569) to -0.417 (-0.271), but they are all negative. This is important because PRBARR and PRBCONV are two of the deterrence variables that have attracted most interest in the previous literature (see, for example, Cherry and List, 2002, Cornwell and Trumbull, 1994, and Ghasemi, 2017). The results reported here for PRBARR and PRBCONV are therefore supportive of the deterrence idea. They also suggest that the deterrence effect is not constant, as usually assumed, but time-varying. The estimated effects of PRBPRIS vary even more and even change sign on several occasions. Most of these estimates are, however, insignificant, which means that the differences in the results need not be due to structural shifts but that they also reflect estimation uncertainty. The same is true for AVGSEN, PCTYMLE and the legal sector wages (with possible exceptions for WFED). Not all estimates are insignificant, though, and there are some marked jumps in the results over time. POLPC enters significantly but with an unexpected negative sign, a finding that is consistent with the results of Cornwell and Trumbull (1994), Baltagi (2006), and Baltagi and Liu (2009). The estimated effects vary quite substantially, and this is true also for DENSITY, which enters significantly positive, as expected.

The fact that PRBPRIS and AVGSEN are generally insignificant is consistent with studies such as Baltagi (2006), Bun et al. (2020), and Ghasemi (2017). This fact, together with the significantly negative effect of PRBARR and PRBCONVI, suggests that imprisoning more criminals, or imprisoning them for longer, is not as effective as increasing the risk of apprehension or conviction once arrested. This provides support to the idea that the consequences of being arrested and found guilty of a crime do not stop with the punishment of the criminal justice system but

[11] We refer to Cornwell and Trumbull (1994) for a more detailed description of the data.

that they also include indirect sanctions imposed by society. "A convicted individual may no longer enjoy the same opportunities in the labor market or the same treatment by their peers, and so the opportunity cost of lost income and the cost to the individual of social stigmatization is implied in the event of conviction" (Bun et al., 2020, page 2322).

The importance of the risk of apprehension or conviction once arrested is one of the main findings of this empirical illustration. Another finding is that the estimated coefficients are not stable but time-varying. This is important because the role of breaks in the crime generating process is still an unsettled issue. The following quotation from McDowall and Loftin (2005, page 361) captures the sentiment in the literature: "contingency theories do not disagree with the conventional approach about the variables that produce the rate changes. Instead, they add a new layer of complexity to allow for the context within which the rate-generating process operates. If these theories are correct, they could significantly improve knowledge about how crime rates change over time. If they are incorrect, they might needlessly complicate attempts to refine the standard approach." Hence, while there are theories that are suggestive of shifts, their empirical relevance has not yet been determined. For the 1981–1987 period that we are considering there were a number of major events that might have caused the crime process to change (see, for example, Carlson and Michalowski, 1997). Most importantly there was (i) the election of Ronald Reagan in 1980 and the political-economic reorganization that followed, (ii) a displacement of nonwhite inner-city males from the regular labour force to the criminogenic informal drug economy, and (iii) a steep increase in juvenile violent crime. Of course, societal change is rarely abrupt but trends to be gradual in nature (see, for example, Batton and Jensen, 2002, and McDowall and Loftin, 2005). It is therefore difficult say exactly what event caused which breaks. What we can say, however, is that breaks are important and that they cannot be ignored when estimating models of crime.

5.3 Robustness

In order to get a feeling for the validity of the interactive effects assumption, we computed the average correlation coefficient of the PDL2S residuals for all pairs of counties, and the CD test of Pesaran (2021), which tests the null hypothesis of no cross-sectional correlation. If the interactive

effects assumption is correct, the regression errors should be cross-county uncorrelated, whereas if the assumption is incorrect there should be some remaining cross-county correlation. Hence, only if the residuals are cross-county uncorrelated can we conclude in favor of the interactive effects assumption. The average correlation coefficient is -0.011 and the CD statistic is -1.806, which is significant at the 10% but not at the 5% level or better. We take these results to suggest that there are no major violations of the interactive effects assumption.

As mentioned in Section 3, our assumptions allow for virtually any dynamics in the idiosyncratic regression errors. In order to shed some light on the persistence of these errors, we estimated the largest autoregressive root of the PDL2S residuals. The estimated autoregressive root is 0.752 and it is highly significant. Hence, as expected given the results of Bun et al. (2020), and Ghasemi (2017), crime is highly persistent and it is therefore important to use methods that are robust in this regard.

Cornwell and Trumbull (1994) argue that PRBARR and POLPC may be endogenous. As pointed out in Section 2, we do not require strict exogeneity but only exogeneity conditional on the factors. In order to assess the validity of this condition, we employed a post-demeaned version of the Lasso GMM of Qian and Su (2016). The instruments used are the same as in Cornwell and Trumbull (1994). They are the fraction of crimes that involve face-to-face contact, and per capita tax revenue. While the first instrument is likely to be correlated with PRBARR, as face-to-face contact makes it possible for victim to identify the offender, the second is likely to correlate with POLPC, as counties with preferences for law enforcement will vote for higher taxes to fund a larger police force.[12] The results, available upon request, are very similar to those reported in Table 6. The main difference is that the standard errors are much larger in the GMM specification, which is consistent with the results of Baltagi (2006), Baltagi and Liu (2009), Bun et al. (2020), and Cornwell and Trumbull (1994). We interpret these results as providing evidence in favor of the PDL2S results reported in Table 6. The logic goes as follows: While LS and GMM are

[12]Hence, there are two instruments, one for each of the two endogenous regressors. This means that the model is just identified. We experimented with using the one-year lagged values of PRBARR and POLPC as additional instruments. Because the resulting model is overidentified, we can apply the overidentifying restrictions J-statistic to assess the validity of the instruments. The instruments passed the test. The problem is that the lags do not appear to be very relevant, in that PRBARR and POLPC are basically serially uncorrelated, which casts doubt on the results based on the larger instrument set. For this reason, we follow the previous literature and focus on the just-identified model specification. All other regressors are treated as exogenous and are therefore included in the set of instruments.

both consistent under exogeneity (conditional on the factors), the GMM instruments are not as informative as the regressors that they replace, leading to variance inflation. Hence, as Murray (2006, page 115) points out, "even valid instruments that are correlated with the troublesome variable might still prove too inefficient to be informative". This is consistent with Bun et al. (2020), and Cornwell and Trumbull (1994), who on efficiency grounds prefer LS over GMM.

6 Conclusion

The present paper considers what we believe to be an empirically very relevant scenario, namely, a researcher faced with the task of estimating a panel data model with unobserved heterogeneity and slope coefficients that may be subject to multiple structural breaks. The researcher wants to be able to estimate not only the slope coefficients within each regime, but also the unknown breakpoints and their number. Moreover, because the panel data set is short, estimation must be possible even if the number of time periods, T, is fixed and only the number of cross-sectional units, N, is large. The current paper contributes by developing a Lasso-based approach that meets this list of demands.

Our asymptotic results show that with probability approaching one the new approach correctly determines the number of breaks and their locations, and that the estimator of the regime-specific regression coefficients is consistent and asymptotically normal. Simulation results are also provided to suggest that the asymptotic predictions are borne out well in small samples.

References

Ahn, S. C., Y. H. Lee, and P. Schmidt (2013). Panel Data Models with Multiple Time-Varying Individual Effects. *Journal of Econometrics* **174**, 1–14.

Antoch, J., J. Hanousek, L. Horváth, M. Hušková and S. Wang (2019). Structural Breaks in Panel Data: Large Number of Panels and Short Length Time Series. *Econometric Reviews* **38**, 828–855.

Andrews, D. W. K. (2005). Cross-Section Regression with Common Shocks. *Econometrica* **73**, 1551–1585.

Bai, J. (2009). Panel Data Models with Interactive Fixed Effects. *Econometrica* **77**, 1229–1279.

Baltagi, B. H. (2006). Estimating an Economic Model of Crime using Panel Data from North Carolina. *Journal of Applied Econometrics* **21**, 543–547.

Baltagi, B. H., and L. Liu (2009). A Note on the Application of EC2SLS and EC3SLS Estimators in Panel Data Models. *Statistics and Probability Letters* **79**, 2189–2192.

Baltagi, B. H., Q. Feng and C. Kao (2016). Estimation of Heterogeneous Panels with Structural Breaks. *Journal of Econometrics* **191**, 176–195.

Baltagi, B. H., C. Kao and L. Liu (2017). Estimation and Identification of Change Points in Panel Models with Nonstationary or Stationary Regressors and Error Term. *Econometric Reviews* **36**, 85–102.

Belloni, A., V. Chernozhukov, C. Hansen and D. Kozbur (2016) Inference in High-Dimensional Panel Models with an Application to Gun Control. *Journal of Business & Economic Statistics* **34**, 590–605.

Boldea, O., B. Drepper and Z. Gan (2020). Change Point Estimation in Panel Data with Time-Varying Individual Effects. *Journal of Applied Econometrics* **35**, 712–727.

Bun, M. J. G., R. Kelaher, V. Sarafidis and D. Weatherburn (2020) Crime, Deterrence and Punishment Revisited. *Empirical Economics* **59**, 2303–2333.

Bushway, S., R. Brame and R. Paternoster (1999). Assessing Stability and Change in Criminal Offending: A Comparison of Random Effects, Semiparametric, and Fixed Effects Modeling Strategies. *Journal of Quantitative Criminology* **15**, 23–61.

Cherry, T., and J. List (2002). Aggregation Bias in the Economic Model of Crime. *Economics Letters* **75**, 81–86.

Cornwell, C., and W. N. Trumbull (1994). Estimating the Economic Model of Crime with Panel Data. *Review of Economics and Statistics* **76**, 360–366.

Dills, A. K., J. A. Miron and G. Summers (2010). What Do Economists Know about Crime? In Di Tella, R., S. Edwards and E. Schargrodsky (Eds), *The Economics of Crime: Lessons for and from Latin America*, 269–302. National Bureau of Economic Research, University of Chicago Press.

Hansen, C., and Y. Liao (2019). The Factor-Lasso and *K*-Step Bootstrap Approach for Inference in High-Dimensional Economic Applications. *Econometric Theory* **35**, 465–509.

Hidalgo, J., and M. Schafgans (2017). Inference and Testing Breaks in Large Dynamic Panels with Strong Cross-Sectional Dependence. *Journal of Econometrics* **196**, 259–274.

Kapetanios, G., L. Serlenga and Y. Shin (2019). Testing for Correlated Factor Loadings in Cross Sectionally Dependent Panels. SERIES Working papers N. 02/2019.

Karavias, Y., P. Narayan and J. Westerlund (2021). Structural Breaks in Interactive Effects Panels and the Stock Market Reaction to COVID–19. Unpublished manuscript.

Li, D., J. Qian and L. Su (2016). Panel Data Models With Interactive Fixed Effects and Multiple Structural Break. *Journal of the American Statistical Association* **111**, 1804–1819.

Lott, J. R. Jr., and D. B. Mustard (1997). The Right-to-Carry Concealed Guns and the Importance of Deterrence. *Journal of Legal Studies* **26**, 1–64.

Moon, H. R., and M. Weidner (2015). Linear Regression for Panel with Unknown Number of Factors as Interactive Fixed Effects. *Econometrica* **83**, 1543–1579.

Murray, M. P. (2006). Avoiding Invalid Instruments and Coping with Weak Instruments. *Journal of Economic Perspectives* **20**, 111–132.

Nagin, D., and R. Paternoster (2000). Population Heterogeneity and State Dependence: State of the Evidence and Directions for Future Research. *Journal of Quantitative Criminology* **16**, 117–144.

Pesaran, M. H. (2006). Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure. *Econometrica* **74**, 967–1012.

Pesaran, M. H. (2021). General Diagnostic Tests for Cross-Sectional Dependence in Panels. *Empirical Economics* **60**, 13–50.

Petrova, Y., and J. Westerlund (2020). Fixed Effects Demeaning in the Presence of Interactive Effects in Treatment Effects Regressions and Elsewhere. *Journal of Applied Econometrics* **35**, 960–964.

Qian, J., and L. Su (2016). Shrinkage Estimation of Common Breaks in Panel Data Models via Adaptive Group Fused Lasso. *Journal of Econometrics* **191**, 86–109.

Robertson, D., and V. Sarafidis (2015). IV Estimation of Panels with Factor Residuals. *Journal of Econometrics* **185**, 526–541.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu and K. Knight (2005). Sparsity and Smoothness via the Fused Lasso. *Journal of the Royal Statistical Society, Series B* **67**, 91–108.

Westerlund, J., Y. Petrova and M. Norkute (2019). CCE in Fixed-T Panels. *Journal of Applied Econometrics* **34**, 746–761.

Worrall, J. L., and T. C. Pratt (2004). On the Consequences of ignoring Unobserved Heterogeneity when Estimating Macro-Level Models of Crime. *Social Science Research* **33**, 79–105.

Zhu, H., V. Sarafidis and M. J. Silvapulle (2020). A New Structural Break Test for Panels with Common Factors. *Econometrics Journal* **23**, 137–155.

Table 1: Simulation results for the case with stationary factors ($\phi = 0.8$) and low error dependence ($\pi = 0.4$).

	$T = 5$					$T = 10$					$T = 20$				
N	25	50	100	200	600	25	50	100	200	600	25	50	100	200	600
							$m = 0$								
NB	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000
BP	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AVE	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000
MSE	0.086	0.077	0.041	0.044	0.018	0.083	0.038	0.033	0.011	0.010	0.035	0.026	0.009	0.012	0.001
							$m = 1$								
NB	0.019	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
BP	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AVE	0.983	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
MSE	0.387	0.120	0.106	0.062	0.041	0.142	0.090	0.061	0.019	0.018	0.105	0.040	0.025	0.020	0.011
							$m = 2$								
NB	0.151	0.045	0.005	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
BP	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AVE	1.840	1.955	1.995	2.000	2.000	1.997	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000
MSE	1.571	0.519	0.144	0.053	0.036	0.188	0.075	0.048	0.044	0.016	0.106	0.068	0.031	0.016	0.011

Notes: "NB", "BP", "AVE" and "MSE" refer to the frequency of false detection of the estimated number of breaks, the frequency of false detection of the estimated breakpoints given that the number of breaks is selected correctly, the average number of estimated breaks, and the MSE of PDL2S estimator times 100, computed as the average $100 \cdot \|\hat{\mathbf{A}}_{\hat{m}} - \mathbf{A}^0_{m^0}\|/(\hat{m}+1)p$ across the Monte Carlo replications, respectively. m, ϕ and π refer to the number of breaks, the persistence of the factors, and the serial and cross-sectional correlation of the errors in the equations for $y_{i,t}$ and $\mathbf{x}_{i,t}$, respectively.

Table 2: Simulation results for the case with non-stationary factors ($\phi = 1$) and low error dependence ($\pi = 0.4$).

	$T = 5$					$T = 10$					$T = 20$				
N	25	50	100	200	600	25	50	100	200	600	25	50	100	200	600
							$m = 0$								
NB	0.001	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.000	0.038	0.002	0.000	0.000	0.000
BP	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AVE	0.001	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.000	0.038	0.002	0.000	0.000	0.000
MSE	0.119	0.108	0.035	0.020	0.022	0.056	0.060	0.039	0.027	0.016	0.077	0.052	0.031	0.012	0.011
							$m = 1$								
NB	0.015	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	3.000	0.000	0.000	0.000	0.000
BP	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AVE	0.985	0.990	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.003	1.000	1.000	1.000	1.000
MSE	0.351	0.096	0.076	0.081	0.030	0.131	0.043	0.039	0.041	0.021	0.050	0.053	0.0139	0.024	0.014
							$m = 2$								
NB	0.153	0.043	0.012	0.000	0.000	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
BP	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AVE	1.847	1.957	1.988	2.000	2.000	1.998	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000
MSE	1.596	0.457	0.191	0.039	0.029	0.124	0.073	0.068	0.024	0.016	0.099	0.058	0.027	0.026	0.000

Notes: See Table 1 for an explanation.

Table 3: Simulation results for the case with stationary factors ($\phi = 0.8$) and high error dependence ($\pi = 0.8$).

	$T = 5$					$T = 10$					$T = 20$				
N	25	50	100	200	600	25	50	100	200	600	25	50	100	200	600
							$m = 0$								
NB	0.011	0.050	0.062	0.000	0.001	0.017	0.047	0.060	0.000	0.001	0.014	0.057	0.043	0.014	0.001
BP	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AVE	0.011	0.001	0.068	0.000	0.001	0.018	0.051	0.000	0.000	0.001	0.014	0.069	0.047	0.015	0.001
MSE	0.128	0.102	0.132	0.023	0.044	0.076	0.074	0.071	0.035	0.025	0.049	0.038	0.011	0.028	0.000
							$m = 1$								
NB	0.011	0.024	0.030	0.007	0.001	0.003	0.018	0.010	0.002	0.000	0.015	0.019	0.004	0.003	0.000
BP	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AVE	1.005	1.024	1.033	1.007	1.001	1.003	1.020	1.010	1.002	1.000	1.015	1.023	1.005	1.003	1.000
MSE	0.313	0.329	0.177	0.123	0.043	0.173	0.144	0.133	0.051	0.041	0.054	0.120	0.019	0.031	0.015
							$m = 2$								
NB	0.098	0.037	0.008	0.003	0.000	0.003	0.013	0.004	0.002	0.000	0.005	0.013	0.006	0.000	0.000
BP	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AVE	1.909	1.992	2.007	2.003	2.000	2.003	2.012	2.004	2.002	2.000	2.005	2.013	2.006	2.000	2.000
MSE	1.182	0.485	0.238	0.113	0.057	0.208	0.181	0.143	0.062	0.023	0.148	0.127	0.076	0.050	0.020

Notes: See Table 1 for an explanation.

Table 4: Simulation results for the case with non-stationary factors ($\phi = 1$) and high error dependence ($\pi = 0.8$).

	$T = 5$					$T = 10$					$T = 20$				
N	25	50	100	200	600	25	50	100	200	600	25	50	100	200	600
							$m = 0$								
NB	0.015	0.049	0.048	0.019	0.000	0.028	0.057	0.051	0.020	0.001	0.057	0.068	0.034	0.007	0.003
BP	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AVE	0.017	0.051	0.052	0.02	0.000	0.029	0.068	0.056	0.02	0.001	0.061	0.075	0.037	0.007	0.005
MSE	0.179	0.069	0.074	0.073	0.040	0.104	0.049	0.071	0.041	0.025	0.038	0.034	0.032	0.014	0.012
							$m = 1$								
NB	0.003	0.021	0.021	0.006	0.000	0.002	0.012	0.009	0.001	0.000	0.005	0.016	0.007	0.002	0.000
BP	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AVE	1.003	1.024	1.023	1.006	1.000	1.002	1.015	1.009	1.001	1.000	1.005	1.016	1.007	1.002	1.000
MSE	0.270	0.238	0.150	0.061	0.056	0.170	0.151	0.051	0.063	0.053	0.102	0.059	0.049	0.035	0.013
							$m = 2$								
NB	0.103	0.026	0.008	0.001	0.000	0.002	0.005	0.004	0.000	0.000	0.003	0.003	0.006	0.000	0.000
BP	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AVE	1.898	1.998	2.007	2.001	2.000	2.002	2.005	2.004	2.003	2.000	2.003	2.004	2.006	2.000	2.000
MSE	1.208	0.412	0.238	0.130	0.048	0.115	0.121	0.143	0.062	0.032	0.100	0.065	0.076	0.024	0.015

Notes: See Table 1 for an explanation.

Table 5: Simulation results for the post-Lasso LS estimator based on non-demeaned data when factors are non-stationary ($\phi = 1$) and error dependence is high ($\pi = 0.8$).

	$T = 5$					$T = 10$					$T = 20$				
N	25	50	100	200	600	25	50	100	200	600	25	50	100	200	600
								$m = 0$							
NB	0.505	0.621	0.659	0.587	0.526	0.685	0.757	0.710	0.658	0.581	0.771	0.816	0.812	0.708	0.591
BP	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AVE	0.953	1.326	1.478	1.125	0.871	1.876	2.412	2.305	1.849	1.383	2.889	3.529	3.305	2.63	2.018
MSE	3.990	3.863	3.897	3.902	3.948	2.990	3.009	2.970	2.959	3.027	2.321	2.250	2.261	2.258	2.270
								$m = 1$							
NB	0.407	0.491	0.503	0.440	0.376	0.443	0.544	0.543	0.472	0.427	0.622	0.652	0.619	0.519	0.424
BP	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000
AVE	1.620	1.843	1.837	1.704	1.524	2.033	2.463	2.399	2.113	1.870	2.882	3.328	3.157	2.595	2.158
MSE	3.870	2.380	3.810	3.834	3.866	3.110	3.012	2.920	2.975	2.968	2.282	2.259	2.259	2.241	2.243
								$m = 2$							
NB	0.275	0.371	0.385	0.310	0.278	0.383	0.457	0.435	0.360	0.371	0.495	0.520	0.503	0.400	0.371
BP	0.004	0.000	0.000	0.000	0.000	0.001	0.001	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.000
AVE	2.341	2.505	2.521	2.411	2.341	2.705	3.013	2.973	2.696	2.637	3.268	3.619	3.449	3.052	2.876
MSE	3.700	3.750	3.612	3.706	3.689	3.032	2.947	2.934	2.936	2.929	2.284	2.257	2.248	2.249	2.251

Notes: See Table 1 for an explanation.

Table 6: Empirical estimation results.

	PDL2S						
Regressor	1981–1982	1983	1984	1985	1986	1987	NBR
PRBARR	−0.417***	−0.681***	−0.532***	−0.663***	−0.634***	−0.457***	−0.521***
PRBCONV	−0.360***	−0.379***	−0.405***	−0.394***	−0.569***	−0.271***	−0.398***
PRBPRIS	0.055	0.419***	0.067	−0.116	0.214	−0.052	0.090
AVGSEN	−0.113	−0.253*	0.094	−0.044	−0.123	−0.258	−0.116
POLPC	0.175**	0.377***	0.338***	0.309***	0.474***	0.271**	0.290***
DENSITY	0.218***	0.012	0.246***	0.198**	0.038	0.248***	0.179***
WCON	−0.126	−0.049	0.040	−0.128	0.333	0.231	−0.021
WTUC	−0.070	−0.007	−0.502**	0.264	−0.226	−0.049	−0.046
WTRD	0.124	1.019***	0.088	0.476	0.043	0.189	0.153
WFIR	0.011	−0.088	0.155	−0.259	0.072	−0.506	0.029
WSER	0.002	−0.463	0.068	−0.578	0.027	−0.293	−0.032
WMGF	−0.133	−0.129	−0.476***	−0.152	−0.144	0.020	−0.217
WFED	0.687**	0.547	0.541	0.480	0.524	1.005**	0.626**
WSTA	−0.266	−0.213	−0.389	−0.320	−0.449	−0.085	−0.279
WLOC	0.257	0.481	0.121	0.850	0.510	−0.091	0.251
PCTYMLE	0.292	0.243	−0.022	0.200	−0.076	0.200	0.175

Notes: The dependent variable is the crime rate (CRMRTE). The deterrence regressors are cases the probability of arrest (PRBARR), the probability of conviction given arrest (PRBCONV), and the probability of a prison sentence given a conviction (PRBPRIS). The included control variables are average prison sentence in days (AVGSEN), the number of police per capita (POLPC), population density (DENSITY), percent young male (PCTYMLE), and the average weekly wage by industry, where the included industries are construction (WCON), transportation, utilities and communication (WTUC), wholesale and retail trade (WTRD), finance, insurance and real estate (WFIR), services (WSER), manufacturing (WMFG), federal government (WFED), state government (WSTA), and local government (WLOC). All variables are expressed in logs. The column labelled "BNR" contains the estimation results when no breaks are allowed. The symbols *, ** and *** denote statistical significance at the 10%, 5% and 1% levels, respectively.

ONLINE APPENDIX TO "ESTIMATION OF PANEL DATA MODELS WITH INTERACTIVE EFFECTS AND MULTIPLE STRUCTURAL BREAKS WHEN T IS FIXED"

Yousef Kaddoura
Lund University

Joakim Westerlund*
Lund University
and
Deakin University

September 17, 2021

Abstract

This online appendix supplement provides the proof of the asymptotic results provided in Section 3 of the main paper.

Lemma A.1. *Suppose that Assumptions EPS, LAM, Q, MOM and J hold. Then, uniformly in* $t \in \mathcal{T}_{T-1}$,

$$\|\dot{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^0\| = O_p(N^{-1/2}).$$

Proof: Clearly,

$$\sqrt{N}(\dot{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^0) = \left(\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,t}'\right)^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\tilde{\mathbf{x}}_{i,t}\tilde{u}_{i,t}. \tag{A.1}$$

Note how $\tilde{u}_{i,t} = \tilde{\boldsymbol{\lambda}}_i'\mathbf{f}_t + \tilde{\varepsilon}_{i,t} = \tilde{\boldsymbol{\nu}}_i'\mathbf{f}_t + \tilde{\varepsilon}_{i,t}$. Let $g_{i,t} = \boldsymbol{\nu}_i'\mathbf{f}_t + \varepsilon_{i,t}$. By using this and deviations from means,

$$\sum_{i=1}^{N}\tilde{\mathbf{x}}_{i,t}\tilde{u}_{i,t} = \sum_{i=1}^{N}\tilde{\mathbf{x}}_{i,t}g_{i,t}. \tag{A.2}$$

*Department of Economics, Lund University, Box 7082, 220 07 Lund, Sweden. Telephone: +46 46 222 8997. Fax: +46 46 222 4613. E-mail address: joakim.westerlund@nek.lu.se.

Let us further define $\mathbf{X}_i = [\mathbf{x}_{i,1}, ..., \mathbf{x}_{i,T}]'$ and denote by $\mathcal{X}$ be the sigma-field generated by $(\mathbf{X}_1, ..., \mathbf{X}_N)$. Since $\mathbb{E}(g_{i,t}g_{j,t}|\mathcal{C}) = 0$ for $j \neq i$ and $\mathbb{E}(g_{i,t}^2|\mathcal{C}) = \mathbf{f}_t'\mathbf{\Sigma}_i\mathbf{f}_t + \sigma_{i,t}^2$, where $\mathbf{\Sigma}_i = \mathbb{E}(\boldsymbol{\nu}_i\boldsymbol{\nu}_i'|\mathcal{C})$ and $\sigma_{i,t}^2 = \mathbb{E}(\varepsilon_{i,t}^2|\mathcal{C})$, we can show that

$$\begin{aligned}
\mathbb{E}\left(\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\tilde{\mathbf{x}}_{i,t}g_{i,t}\right\|^2|\mathcal{C}\right) &= \mathbb{E}\left(\text{tr}\left[\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\tilde{\mathbf{x}}_{i,t}g_{i,t}\right)\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\tilde{\mathbf{x}}_{i,t}g_{i,t}\right)'\right]|\mathcal{C}\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}[\text{tr}(\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{j,t}')g_{i,t}g_{j,t}|\mathcal{C}] \\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}[\text{tr}(\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{j,t}')\mathbb{E}(g_{i,t}g_{j,t}|\mathcal{X},\mathcal{C})|\mathcal{C}] \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\text{tr}(\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,s}')\mathbb{E}(g_{i,t}^2|\mathcal{X},\mathcal{C})|\mathcal{C}] \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\text{tr}(\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,t}')|\mathcal{C}](\mathbf{f}_t'\mathbf{\Sigma}_i\mathbf{f}_t + \sigma_{i,t}^2) \\
&\leq \left(\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(\|\tilde{\mathbf{x}}_{i,t}\|^4|\mathcal{C})\right)^{1/2}\left(\frac{1}{N}\sum_{i=1}^{N}(\mathbf{f}_t'\mathbf{\Sigma}_i\mathbf{f}_t + \sigma_{i,t}^2)^2\right)^{1/2} \\
&\leq \left(\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(\|\tilde{\mathbf{x}}_{i,t}\|^4|\mathcal{C})\right)^{1/2}\left(\frac{2}{N}\sum_{i=1}^{N}(\|\mathbf{f}_t\|^4\|\mathbf{\Sigma}_i\| + \sigma_{i,t}^4)\right)^{1/2} \\
&= O_p(1). \qquad \text{(A.3)}
\end{aligned}$$

Hence, since the variance is bounded,

$$\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\tilde{\mathbf{x}}_{i,t}g_{i,t}\right\| = O_p(1). \qquad \text{(A.4)}$$

It follows that since $\|(N^{-1}\sum_{i=1}^{N}\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,t}')^{-1}\| = O_p(1)$ by assumption,

$$\|\sqrt{N}(\dot{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^0)\| \leq \left\|\left(\frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,t}'\right)^{-1}\right\|\left\|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\tilde{\mathbf{x}}_{i,t}\tilde{u}_{i,t}\right\| = O_p(1), \qquad \text{(A.5)}$$

which is what we wanted to show. ■

Proof of Theorem 1.

Define $\mathbf{b}_t = \sqrt{N}(\boldsymbol{\beta}_t - \boldsymbol{\beta}_t^0)$ and $\hat{\mathbf{b}}_t = \sqrt{N}(\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^0)$. The first term in the objective function depends on $\tilde{y}_{i,t} - \tilde{\mathbf{x}}_{i,t}'\boldsymbol{\beta}_t$. Making use of the definition of $\mathbf{b}_t$ and $\tilde{y}_{i,t} = \tilde{\mathbf{x}}_{i,t}'\boldsymbol{\beta}_t^0 + \tilde{u}_{i,t}$, this term can

be written as

$$\tilde{y}_{i,t} - \tilde{\mathbf{x}}_{i,t}'\boldsymbol{\beta}_t = \tilde{y}_{i,t} - \tilde{\mathbf{x}}_{i,t}'(\boldsymbol{\beta}_t^0 + N^{-1/2}\mathbf{b}_t) = \tilde{u}_{i,t} - N^{-1/2}\tilde{\mathbf{x}}_{i,t}'\mathbf{b}_t. \tag{A.6}$$

Since N and $\ell_\gamma(\mathbf{B}_T^0)$ do not depend on $\mathbf{B}_T$, centering and scaling of the objective function by these quantities are inconsequential. We therefore proceed to minimize $N[\ell_\gamma(\mathbf{B}_T) - \ell_\gamma(\mathbf{B}_T^0)]$. Note first that

$$\begin{aligned}
N\ell_\gamma(\mathbf{B}_T) &= \sum_{i=1}^N \sum_{t=1}^T (\tilde{y}_{i,t} - \tilde{\mathbf{x}}_{i,t}'\boldsymbol{\beta}_t)^2 + N\gamma \sum_{t=2}^T w_t \|\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1}\| \\
&= \sum_{i=1}^N \sum_{t=1}^T (\tilde{u}_{i,t} - N^{-1/2}\tilde{\mathbf{x}}_{i,t}'\mathbf{b}_t)^2 \\
&+ N\gamma \sum_{t=2}^T w_t \|\boldsymbol{\beta}_t^0 + N^{-1/2}\mathbf{b}_t - (\boldsymbol{\beta}_{t-1}^0 + N^{-1/2}\mathbf{b}_{t-1})\| \\
&= \sum_{i=1}^N \sum_{t=1}^T \tilde{u}_{i,t}^2 - \frac{2}{\sqrt{N}} \sum_{i=1}^N \sum_{t=1}^T \tilde{u}_{i,t}\tilde{\mathbf{x}}_{i,t}'\mathbf{b}_t + \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T \mathbf{b}_t'\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,t}'\mathbf{b}_t \\
&+ N\gamma \sum_{t=2}^T w_t \|\boldsymbol{\beta}_t^0 - \boldsymbol{\beta}_{t-1}^0 + N^{-1/2}(\mathbf{b}_t - \mathbf{b}_{t-1})\|.
\end{aligned} \tag{A.7}$$

Hence, since

$$N\ell_\gamma(\mathbf{B}_T^0) = \sum_{i=1}^N \sum_{t=1}^T \tilde{u}_{i,t}^2 + N\gamma \sum_{t=2}^T w_t \|\boldsymbol{\beta}_t^0 - \boldsymbol{\beta}_{t-1}^0\|, \tag{A.8}$$

we obtain

$$\begin{aligned}
&N[\ell_\gamma(\mathbf{B}_T) - \ell_\gamma(\mathbf{B}_T^0)] \\
&= \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T \mathbf{b}_t'\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,t}'\mathbf{b}_t - \frac{2}{\sqrt{N}} \sum_{i=1}^N \sum_{t=1}^T \tilde{u}_{i,t}\tilde{\mathbf{x}}_{i,t}'\mathbf{b}_t \\
&+ N\gamma \sum_{t=2}^T w_t [\|\boldsymbol{\beta}_t^0 - \boldsymbol{\beta}_{t-1}^0 + N^{-1/2}(\mathbf{b}_t - \mathbf{b}_{t-1})\| - \|\boldsymbol{\beta}_t^0 - \boldsymbol{\beta}_{t-1}^0\|] \\
&= \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T \mathbf{b}_t'\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,t}'\mathbf{b}_t - \frac{2}{\sqrt{N}} \sum_{i=1}^N \sum_{t=1}^T \tilde{u}_{i,t}\tilde{\mathbf{x}}_{i,t}'\mathbf{b}_t \\
&+ N\gamma \sum_{t \in \mathcal{T}_{m^0}^0} w_t [\|\boldsymbol{\beta}_t^0 - \boldsymbol{\beta}_{t-1}^0 + N^{-1/2}(\mathbf{b}_t - \mathbf{b}_{t-1})\| - \|\boldsymbol{\beta}_t^0 - \boldsymbol{\beta}_{t-1}^0\|] \\
&+ N\gamma \sum_{t \in \mathcal{T}_{m^0}^{0c}} w_t \|N^{-1/2}(\mathbf{b}_t - \mathbf{b}_{t-1})\|,
\end{aligned} \tag{A.9}$$

where $\mathcal{T}_{m^0}^{0c} = \{1, \ldots, T\} \setminus \mathcal{T}_{m^0}^0$ is the complement of $\mathcal{T}_{m^0}^0$ and the last equality holds because $\boldsymbol{\beta}_t^0 = \boldsymbol{\beta}_{t-1}^0$ for all $t \in \mathcal{T}_{m^0}^{0c}$ ($\boldsymbol{\beta}_t^0$ is constant within break regimes). Let us write the above equation

more compactly as

$$N[\ell_\gamma(\mathbf{B}_T) - \ell_\gamma(\mathbf{B}_T^0)] = H_1(\mathbf{b}) - 2H_2(\mathbf{b}) + H_3(\mathbf{b}) + H_4(\mathbf{b}), \tag{A.10}$$

where $\mathbf{b} = [\mathbf{b}_1', \ldots, \mathbf{b}_T']'$ and implicit definitions of $H_1(\mathbf{b})$, $H_2(\mathbf{b})$, $H_3(\mathbf{b})$ and $H_4(\mathbf{b})$. For $\|\mathbf{b}\|$, there are two possibilities; it is either bounded or unbounded. We now show that if $\|\mathbf{b}\| > 0$ is unbounded, $N[\ell_\gamma(\mathbf{B}_T) - \ell_\gamma(\mathbf{B}_T^0)] > 0$, which means that $\ell_\gamma(\mathbf{B}_T)$ cannot be minimized in this case. This implies that $\hat{\mathbf{B}}_T$ is consistent with $\|\mathbf{b}\| = \|\hat{\mathbf{b}}\| = O_p(1)$.

Consider $H_1(\mathbf{b})$. This term is positive, as is clear from

$$\begin{aligned} H_1(\mathbf{b}) &= \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} \mathbf{b}_t'\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,t}'\mathbf{b}_t = \mathbf{b}'\mathbf{Q}_N(\mathcal{T}_{T-1})\mathbf{b} = |\mathbf{b}'\mathbf{Q}_N(\mathcal{T}_{T-1})\mathbf{b}| \\ &\geq \lambda_{min}[\mathbf{Q}_N(\mathcal{T}_{T-1})]\|\mathbf{b}\|^2 > 0, \end{aligned} \tag{A.11}$$

where $\mathbf{Q}_N(\mathcal{T}_{T-1})$ is as in the main text.

Define

$$\mathbf{U}_N(\mathcal{T}_m) = \begin{bmatrix} \frac{1}{N}\sum_{t=T_0}^{T_1-1}\sum_{i=1}^{N}\tilde{\mathbf{x}}_{i,t}\tilde{u}_{i,t} \\ \vdots \\ \frac{1}{N}\sum_{t=T_m}^{T_{m+1}-1}\sum_{i=1}^{N}\tilde{\mathbf{x}}_{i,t}\tilde{u}_{i,t} \end{bmatrix}, \tag{A.12}$$

a $(m+1)p \times 1$ vector. Note that $\mathbf{U}_N(\mathcal{T}_{T-1}) = [N^{-1}\sum_{i=1}^{N}\tilde{\mathbf{x}}_{i,1}'\tilde{u}_{i,1}, \ldots, N^{-1}\sum_{i=1}^{N}\tilde{\mathbf{x}}_{i,T}'\tilde{u}_{i,T}]'$. Hence, in this notation, $H_2(\mathbf{b}) = \mathbf{b}'\sqrt{N}\mathbf{U}_N(\mathcal{T}_{T-1})$. This is not positive but we know from the proof of Lemma A.1 that $\|N^{-1/2}\sum_{i=1}^{N}\tilde{\mathbf{x}}_{i,t}'\tilde{u}_{i,t}\| = O_p(1)$ uniformly in t, and hence $\|\sqrt{N}\mathbf{U}_N(\mathcal{T}_{T-1})\| = O_p(1)$. It follows that

$$|H_2(\mathbf{b})| \leq \|\mathbf{b}\|\|\sqrt{N}\mathbf{U}_N(\mathcal{T}_{T-1})\| = O_p(\|\mathbf{b}\|). \tag{A.13}$$

Hence, since $H_1(\mathbf{b}) = O_p(\|\mathbf{b}\|^2)$, if $\|\mathbf{b}\|$ is unbounded, $H_2(\mathbf{b})$ will be dominated by $H_1(\mathbf{b})$.

Next up is $H_3(\mathbf{b})$. Consider

$$\begin{aligned} \sum_{t\in\mathcal{T}_{m^0}^0}\|b_t - b_{t-1}\| &\leq \left(\sum_{t\in\mathcal{T}_{m^0}^0}\|b_t - b_{t-1}\|^2\right)^{1/2} \leq \left(\sum_{t\in\mathcal{T}_{m^0}^0}\|b_t + b_{t-1}\|^2\right)^{1/2} \\ &\leq \left(2\sum_{t\in\mathcal{T}_{m^0}^0}(\|b_t\|^2 + \|b_{t-1}\|^2)\right)^{1/2} \leq \left(2\sum_{t\in\mathcal{T}_{m^0}^0}\|b_t\|^2\right)^{1/2} \\ &\leq \sqrt{2}\|\mathbf{b}\|, \end{aligned} \tag{A.14}$$

which in turn implies

$$\begin{aligned}
H_3(\mathbf{b}) &= N\gamma \sum_{t\in\mathcal{T}^0_{m^0}} w_t[\|\beta^0_t - \beta^0_{t-1} + N^{-1/2}(b_t - b_{t-1})\| - \|\beta^0_t - \beta^0_{t-1}\|] \\
&\leq N \sum_{t\in\mathcal{T}^0_{m^0}} w_t[\|\beta^0_t - \beta^0_{t-1}\| + \|N^{-1/2}(b_t - b_{t-1})\| - \|\beta^0_t - \beta^0_{t-1}\|] \\
&= \sqrt{N}\gamma \sum_{t\in\mathcal{T}^0_{m^0}} w_t\|b_t - b_{t-1}\| \\
&\leq \sqrt{N}\gamma \max_{s\in\mathcal{T}^0_{m^0}} w_s \sum_{t\in\mathcal{T}^0_{m^0}} \|b_t - b_{t-1}\| \\
&\leq \sqrt{2N}\gamma \max_{s\in\mathcal{T}^0_{m^0}} w_s\|\mathbf{b}\|. \qquad \text{(A.15)}
\end{aligned}$$

Consider $\max_{s\in\mathcal{T}^0_{m^0}} w_s$. By Lemma A.1, $\|\dot{\boldsymbol{\beta}}_t - \boldsymbol{\beta}^0_t\| = O_p(N^{-1/2})$. By using this, the fact that $\|\boldsymbol{\beta}^0_t - \boldsymbol{\beta}^0_{t-1}\| > 0$ for $t \in \mathcal{T}^0_{m^0}$, and assumption J (b) we get

$$\begin{aligned}
w_t &= \|\dot{\boldsymbol{\beta}}_t - \dot{\boldsymbol{\beta}}_{t-1}\|^{-\kappa} = \left(\|\boldsymbol{\beta}^0_t - \boldsymbol{\beta}^0_{t-1}\| + O_p(N^{-1/2})\right)^{-\kappa} \\
&\leq \left(\min_{t\in\mathcal{T}^0_{m^0}} \|\boldsymbol{\beta}^0_t - \boldsymbol{\beta}^0_{t-1}\| + O_p(N^{-1/2})\right)^{-\kappa} \\
&= \left(\min_{1\leq j\leq m^0+1} \|\boldsymbol{\alpha}^0_{j+1} - \boldsymbol{\alpha}^0_j\| + O_p(N^{-1/2})\right)^{-\kappa} \\
&= O_p(J_{min}^{-\kappa}) \qquad \text{(A.16)}
\end{aligned}$$

for all $t \in \mathcal{T}^0_{m^0}$, which means that $\max_{s\in\mathcal{T}^0_{m^0}} w_s$ is of the same order. By using this and the condition that $\sqrt{N}\gamma J_{min}^{-\kappa} = O_p(1)$ (Assumption J), we obtain

$$H_3(\mathbf{b}) \leq \sqrt{2N}\gamma \max_{s\in\mathcal{T}^0_{m^0}} w_s\|\mathbf{b}\| = O_p(\sqrt{N}\gamma J_{min}^{-\kappa}\|\mathbf{b}\|) = O_p(\|\mathbf{b}\|). \qquad \text{(A.17)}$$

The above results imply

$$\begin{aligned}
H_1(\mathbf{b}) - 2H_2(\mathbf{b}) + H_3(\mathbf{b}) &\geq H_1(\mathbf{b}) - 2|H_2(\mathbf{b})| - |H_3(\mathbf{b})| \\
&= O_p(\|\mathbf{b}\|^2) - O_p(\|\mathbf{b}\|) > 0. \qquad \text{(A.18)}
\end{aligned}$$

We also see that $H_4(\mathbf{b}) \geq 0$. Hence, if $\|\mathbf{b}\|$ is unbounded, then $N[\ell_\gamma(\mathbf{B}_T) - \ell_\gamma(\mathbf{B}^0_T)] > 0$. But we also know that $N[\ell_\gamma(\hat{\mathbf{B}}_T) - \ell_\gamma(\mathbf{B}^0_T)] \leq 0$, which means that $\|\mathbf{b}\| = \|\hat{\mathbf{b}}\|$ must be bounded. The required result is implied by this. ■

Proof of Theorem 2.

Let $\hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\beta}}_t - \hat{\boldsymbol{\beta}}_{t-1}$ and $\boldsymbol{\theta}_t = \boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1}$. We want to show that $\|\hat{\boldsymbol{\theta}}_t\| = 0$ for all $t \in \mathcal{T}_{m^0}^{0c}$ w.p.a.1. This is done through a contradiction argument. Let us therefore assume that $\|\hat{\boldsymbol{\theta}}_t\| > 0$ for some $t \in \mathcal{T}_{m^0}^{0c}$ and all N, including $N \to \infty$. We now show that if this is the case, $\hat{\boldsymbol{\theta}}_t$ cannot satisfy the first-order condition from which it was derived. This implies that $\|\hat{\boldsymbol{\theta}}_t\| > 0$ cannot be true and that $\|\hat{\boldsymbol{\theta}}_t\| = 0$ w.p.a.1.

We now consider the first-order partial derivative of $\ell_\gamma(\mathbf{B}_T)$ with respect to $\beta_{t,p}$, the p-th element of $\boldsymbol{\beta}_t$. In so doing, we make use of

$$\begin{aligned}\frac{\partial}{\partial \beta_{t,p}}\|\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1}\| &= \frac{\partial}{\partial \beta_{t,p}}[(\beta_{t,1} - \beta_{t-1,1})^2 + \ldots + (\beta_{t,p} - \beta_{t-1,p})^2]^{1/2} \\ &= \frac{\beta_{t,p} - \beta_{t-1,p}}{[(\beta_{t,1} - \beta_{t-1,1})^2 + \ldots + (\beta_{t,p} - \beta_{t-1,p})^2]^{1/2}} = \frac{\theta_{t,p}}{\|\boldsymbol{\theta}_t\|} \end{aligned} \tag{A.19}$$

for $\|\boldsymbol{\theta}_t\| > 0$, and similarly

$$\frac{\partial}{\partial \beta_{t,p}}\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\| = -\frac{\theta_{t+1,p}}{\|\boldsymbol{\theta}_{t+1}\|} \tag{A.20}$$

for $\|\boldsymbol{\theta}_{t+1}\| > 0$. Also, from the proof of Theorem 1, $\tilde{y}_{i,t} - \tilde{\mathbf{x}}_{i,t}'\boldsymbol{\beta}_t = \tilde{u}_{i,t} - \tilde{\mathbf{x}}_{i,t}'(\boldsymbol{\beta}_t - \boldsymbol{\beta}_t^0)$. By using these results and the definition of $\ell_\gamma(\mathbf{B}_T)$,

$$\begin{aligned}\frac{\partial \ell_\gamma(\mathbf{B}_T)}{\partial \beta_{t,p}} &= \frac{\partial}{\partial \beta_{t,p}} \sum_{i=1}^N \frac{1}{N} \sum_{t=1}^T (\tilde{y}_{i,t} - \tilde{\mathbf{x}}_{i,t}'\boldsymbol{\beta}_t)^2 + \frac{\partial}{\partial \beta_{t,p}} \gamma \sum_{t=2}^T w_t \|\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1}\| \\ &= -\frac{2}{N} \sum_{i=1}^N (\tilde{y}_{i,t} - \tilde{\mathbf{x}}_{i,t}'\boldsymbol{\beta}_t)\tilde{x}_{i,t,p} + \gamma \frac{\partial}{\partial \beta_{t,p}} (w_t \|\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1}\| + w_{t+1}\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t\|) \\ &= -\frac{2}{N} \sum_{i=1}^N (\tilde{y}_{i,t} - \tilde{\mathbf{x}}_{i,t}'\boldsymbol{\beta}_t)\tilde{x}_{i,t,p} + \gamma w_t \frac{\theta_{t,p}}{\|\boldsymbol{\theta}_t\|} - \gamma w_{t+1} \frac{\theta_{t+1,p}}{\|\boldsymbol{\theta}_{t+1}\|} \\ &= \frac{2}{N} \sum_{i=1}^N \tilde{\mathbf{x}}_{i,t}'(\boldsymbol{\beta}_t - \boldsymbol{\beta}_t^0)\tilde{x}_{i,t,p} - \frac{2}{N} \sum_{i=1}^N \tilde{u}_{i,t}\tilde{x}_{i,t,p} + \gamma w_t \frac{\theta_{t,p}}{\|\boldsymbol{\theta}_t\|} - \gamma w_{t+1} \frac{\theta_{t+1,p}}{\|\boldsymbol{\theta}_{t+1}\|}, \end{aligned} \tag{A.21}$$

where $\tilde{x}_{i,t,p}$ is the p-th row of $\tilde{\mathbf{x}}_{i,t}$. Assume without loss of generality that $|\hat{\theta}_{t,1}| \leq \ldots \leq |\hat{\theta}_{t,p}|$, where $\hat{\theta}_{t,j}$ is the j-th row of $\hat{\boldsymbol{\theta}}_t$. Moreover, while for the purpose of this proof we can assume that $\|\hat{\boldsymbol{\theta}}_t\| > 0$, we do not want to impose $\|\hat{\boldsymbol{\theta}}_{t+1}\| > 0$. Therefore, in order to ensure that $\theta_{t+1,p}/\|\boldsymbol{\theta}_{t+1}\|$ is well defined when evaluated at $\boldsymbol{\theta}_{t+1} = \hat{\boldsymbol{\theta}}_{t+1}$, we introduce the $p \times 1$ vector $\mathbf{g}_t$, whose p-th element is denoted $g_{t,p}$. This vector is such that $g_{t,p} = \hat{\theta}_{t,p}/\|\hat{\boldsymbol{\theta}}_t\|$ if $\|\hat{\boldsymbol{\theta}}_t\| > 0$ and $\|\mathbf{g}_t\| \leq 1$

if $\|\hat{\boldsymbol{\theta}}_t\| = 0$, where the latter result is due to the theory on sub-differential calculus. In this notation, the sought first-order condition (multiplied by $\sqrt{N}$) is given by

$$\begin{aligned}\sqrt{N}\frac{\partial \ell_\gamma(\hat{\mathbf{B}}_T)}{\partial \beta_{t,p}} &= \frac{2}{\sqrt{N}}\sum_{i=1}^{N}\tilde{\mathbf{x}}_{i,t}'(\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^0)\tilde{x}_{i,t,p} - \frac{2}{\sqrt{N}}\sum_{i=1}^{N}\tilde{u}_{i,t}\tilde{x}_{i,t,p} + \sqrt{N}\gamma w_t g_{t,p} \\ &\quad - \sqrt{N}\gamma w_{t+1} g_{t+1,p} \\ &= M_{1,t} - M_{2,t} + M_{3,t} - M_{3,t+1} = 0, \end{aligned} \tag{A.22}$$

with $M_{1,t}$, $M_{2,t}$ and $M_{3,t}$ implicitly defined.

Consider $M_{1,t}$. By Theorem 1, $\|\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^0\| = O_p(N^{-1/2})$, implying that

$$\begin{aligned}|M_{1,t}| &= \left|\frac{2}{N}\sum_{i=1}^{N}\tilde{x}_{i,t,p}\tilde{\mathbf{x}}_{i,t}'\sqrt{N}(\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^0)\right| \leq 2\left\|\frac{1}{N}\sum_{i=1}^{N}\tilde{x}_{i,t,p}\tilde{\mathbf{x}}_{i,t}\right\| \|\sqrt{N}(\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^0)\| \\ &\leq 2\left(\frac{1}{N}\sum_{i=1}^{N}\tilde{x}_{i,t,p}^2\right)^{1/2}\left(\frac{1}{N}\sum_{i=1}^{N}\|\tilde{\mathbf{x}}_{i,t}\|^2\right)^{1/2}\|\sqrt{N}(\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^0)\| = O_p(1), \end{aligned} \tag{A.23}$$

and by the proof of Lemma A.1, $|M_{2,t}|$ is of the same order. For $|M_{3,t}|$, we use the fact that $\boldsymbol{\beta}_t^0 - \boldsymbol{\beta}_{t-1}^0 = \mathbf{0}_{p\times 1}$ for $t \in \mathcal{T}_{m^0}^{0c}$. This implies

$$w_t = \|\dot{\boldsymbol{\beta}}_t - \dot{\boldsymbol{\beta}}_{t-1}\|^{-\kappa} = \|\boldsymbol{\beta}_t^0 - \boldsymbol{\beta}_{t-1}^0 + O_p(N^{-1/2})\|^{-\kappa} = O_p(N^{\kappa/2}). \tag{A.24}$$

Also, since $\|\hat{\boldsymbol{\theta}}_t\| > 0$ by assumption, $g_{t,p} = \hat{\theta}_{t,p}/\|\hat{\boldsymbol{\theta}}_t\|$. It follows that

$$|M_{3,t}| = \sqrt{N}\gamma w_t|g_{t,p}| = \sqrt{N}\gamma w_t\frac{|\hat{\theta}_{t,p}|}{\|\hat{\boldsymbol{\theta}}_t\|} \geq \sqrt{N}\gamma\frac{w_t}{\sqrt{p}} = O_p(N^{(\kappa+1)/2}\gamma) \to \infty, \tag{A.25}$$

where the inequality is due to

$$\frac{1}{\sqrt{p}} \leq \frac{|\hat{\theta}_{t,p}|}{\|\hat{\boldsymbol{\theta}}_t\|} \leq 1, \tag{A.26}$$

and the divergence follows from $N^{(\kappa+1)/2}\gamma \to \infty$ (Assumption J).

The order of $|M_{3,t+1}|$ depends on whether (A) $t+1 \in \mathcal{T}_{m^0}^0$ for $j \in \{1,\ldots,m^0\}$, or (B) $t+1 \in \mathcal{T}_{m^0}^{0c}$. Suppose that (A) is true, such that $\|\boldsymbol{\beta}_{t+1}^0 - \boldsymbol{\beta}_t^0\| > 0$. Then, by Theorem 1,

$$w_{t+1} = \|\dot{\boldsymbol{\beta}}_{t+1} - \dot{\boldsymbol{\beta}}_t\|^{-\kappa} = \|\boldsymbol{\beta}_{t+1}^0 - \boldsymbol{\beta}_t^0 + O_p(N^{-1/2})\|^{-\kappa} = O_p(J_{min}^{-\kappa}). \tag{A.27}$$

Since $\|\mathbf{g}_{t+1}\| \leq 1$, we know that $|g_{t+1,p}| = O_p(1)$. By using this, $w_{t+1} = O_p(J_{min}^{-\kappa})$ and $\sqrt{N}\gamma J_{min}^{-\kappa} = O(1)$ (Assumption J), we can show that

$$|M_{3,t+1}| = \sqrt{N}\gamma w_{t+1}|g_{t+1,p}| = O_p(\sqrt{N}\gamma J_{min}^{-\kappa}) = O_p(1). \tag{A.28}$$

Hence, if (A) holds, then $M_{1,t}$, $M_{2,t}$ and $M_{3,t+1}$ are all $O_p(1)$, while $|M_{3,t}| \to \infty$, which means that the first-order condition is violated. It must therefore be that $\|\hat{\boldsymbol{\theta}}_t\| = 0$ w.p.a.1. But if $\|\hat{\boldsymbol{\theta}}_t\| = 0$, $M_{3,t} = \sqrt{N}\gamma w_t g_{t,p}$ and this term must be $O_p(1)$ for the first-order condition to hold. The fact that $M_{3,t} = O_p(1)$ is partly unexpected, because $w_t = O_p(N^{\kappa/2}) \to \infty$, which means that $g_{t,p}$ must be going to zero at the same rate. We will use this result again now when we continue onto case (B).

If (B) holds, then $t+1 \in \mathcal{T}_{m^0}^{0c}$, and so we again have $w_{t+1} = O_p(N^{\kappa/2}) \to \infty$. We will now argue why, as in the proof of (A), this divergence does not spill over to $M_{3,t+1}$. Let us without loss of generality consider the case when $t+1 = T_j^0$ under (A). This means that $t = T_j^0 - 1$ and, from the analysis of (A), we have $M_{3,t} = M_{3,T_j^0-1} = O_p(1)$. We also know that the first-order condition is violated and hence that $\|\hat{\boldsymbol{\theta}}_t\| = \|\hat{\boldsymbol{\theta}}_{T_j^0-1}\| = 0$ w.p.a.1. Suppose now instead that $t = T_j^0 - 2$, or $t+1 = T_j^0 - 1$, so that (B) holds. In this case, $|M_{3,t}| = |M_{3,T_j^0-2}| = O_p(N^{(\kappa+1)/2}\gamma) \to \infty$ just as before and $M_{3,t+1} = M_{3,T_j^0-1} = O_p(1)$, since we know from (A) that $M_{3,T_j^0-1} = O_p(1)$. This means that the first-order condition is again violated, and therefore $\|\hat{\boldsymbol{\theta}}_t\| = \|\hat{\boldsymbol{\theta}}_{T_j^0-2}\| = 0$ w.p.a.1. This recursive argument can be used repeatedly to show that $\|\hat{\boldsymbol{\theta}}_t\| = 0$ for all $t = T_j^0 - 2, \ldots, T_{j-1}^0 + 1$ w.p.a.1. ■

Proof of Corollary 1.

By Theorem 2, $\hat{m} \leq m^0$ w.p.a.1, since asymptotically no time periods in $\mathcal{T}_{m^0}^{0c}$ can be misclassified as a breakpoint. In what remains, we show that all time periods in $\mathcal{T}_{m^0}^0$ are correctly classified as breakpoints. This implies that $\hat{m} = m^0$ w.p.a.1, and hence both (a) and (b) are proved. We begin by noting that by Theorem 1,

$$\|\hat{\boldsymbol{\theta}}_t\| = \|\hat{\boldsymbol{\beta}}_t - \hat{\boldsymbol{\beta}}_{t-1}\| = \|\boldsymbol{\beta}_t^0 - \boldsymbol{\beta}_{t-1}^0\| + O_p(N^{-1/2}) \tag{A.29}$$

for all $t \in \mathcal{T}_{m^0}^0$. Hence, if $\|\hat{\boldsymbol{\theta}}_t\| = 0$, so that $t \in \mathcal{T}_{m^0}^0$ is not classified as a breakpoint, then

$$\|\boldsymbol{\beta}_t^0 - \boldsymbol{\beta}_{t-1}^0\| = O_p(N^{-1/2}). \tag{A.30}$$

However, since $\|\boldsymbol{\beta}_t^0 - \boldsymbol{\beta}_{t-1}^0\| \geq J_{min}$ for all $t \in \mathcal{T}_{m^0}^0$, this violates the condition that $\sqrt{N}J_{min} \to \infty$ (Assumption J). We therefore conclude that $\|\hat{\boldsymbol{\theta}}_t\| > 0$ for all $t \in \mathcal{T}_{m^0}^0$ w.p.a.1, which means that all time periods in $\mathcal{T}_{m^0}^0$ are correctly classified. ■

Proof of Theorem 3.

By Corollary 1, we know that $\hat{\mathbf{A}}_{\hat{m}} = \hat{\mathbf{A}}_{m^0} = \hat{\mathbf{A}}_{m^0}(\mathcal{T}^0_{m^0}) = [\hat{\boldsymbol{\alpha}}_1(\mathcal{T}^0_{m^0})', \ldots, \hat{\boldsymbol{\alpha}}_{m+1}(\mathcal{T}^0_{m^0})']'$ w.p.a.1. The asymptotic distribution of $\hat{\mathbf{A}}_{\hat{m}}$ is therefore equal to that of $\hat{\mathbf{A}}_{m^0}$, which we now derive. From

$$\tilde{y}_{i,t} = \tilde{\mathbf{x}}'_{i,t}\boldsymbol{\alpha}^0_j + \tilde{u}_{i,t} \tag{A.31}$$

for $t = T^0_{j-1}, \ldots, T^0_j - 1$ with $j = 1, \ldots, m^0 + 1$, we get

$$\hat{\mathbf{A}}_m = \mathbf{A}^0_{m^0} + \mathbf{Q}_N(\mathcal{T}^0_{m^0})^{-1}\mathbf{U}_N(\mathcal{T}^0_{m^0}), \tag{A.32}$$

where $\mathbf{Q}_N(\mathcal{T}^0_{m^0})$ is defined in the main text and $\mathbf{U}_N(\mathcal{T}^0_{m^0})$ is defined in Proof of Theorem 1. Let $\mathbf{Q}_0 = \mathbf{Q}_0(\mathcal{T}^0_{m^0})$. In this notation,

$$\sqrt{N}(\hat{\mathbf{A}}_m - \mathbf{A}^0_{m^0}) = \mathbf{Q}_0^{-1}\sqrt{N}\mathbf{U}_N(\mathcal{T}^0_{m^0}) + [\mathbf{Q}_N(\mathcal{T}^0_{m^0})^{-1} - \mathbf{Q}_0^{-1}]\sqrt{N}\mathbf{U}_N(\mathcal{T}^0_{m^0}). \tag{A.33}$$

By Assumption Q, $\|\mathbf{Q}_N(\mathcal{T}^0_{m^0}) - \mathbf{Q}_0\| = o_p(1)$, where $\mathbf{Q}_N(\mathcal{T}^0_{m^0})$ and $\mathbf{Q}_0$ are both positive definite (w.p.1). Analogously to the proof of Lemma A.1, $\|\sqrt{N}\mathbf{U}_N(\mathcal{T}^0_{m^0})\| = O_p(1)$, $\|\mathbf{Q}_0^{-1}\| = O(1)$ and

$$\|\mathbf{Q}_N(\mathcal{T}^0_{m^0})^{-1} - \mathbf{Q}_0^{-1}\| = o_p(1), \tag{A.34}$$

which in turn implies

$$\begin{aligned}\|[\mathbf{Q}_N(\mathcal{T}^0_{m^0})^{-1} - \mathbf{Q}_0^{-1}]\sqrt{N}\mathbf{U}_N(\mathcal{T}^0_{m^0})\| &\leq \|\mathbf{Q}_N(\mathcal{T}^0_{m^0})^{-1} - \mathbf{Q}_0^{-1}\|\|\sqrt{N}\mathbf{U}_N(\mathcal{T}^0_{m^0})\| \\ &= o_p(1).\end{aligned} \tag{A.35}$$

The second term on the right-hand side of the above expression for $\sqrt{N}(\hat{\mathbf{A}}_m - \mathbf{A}^0_{m^0})$ is therefore negligible.

We now show that $\sqrt{N}\mathbf{U}_N(\mathcal{T}^0_{m^0})$ is asymptotically normal conditional on the sigma-field generated by $(\mathbf{f}_1, \ldots, \mathbf{f}_T)$, $\mathcal{C}$. Let

$$\mathbf{u}_i = \mathbf{u}_i(\mathcal{T}^0_{m^0}) = \begin{bmatrix} \sum_{t=T^0_0}^{T^0_1-1} \tilde{\mathbf{x}}_{i,t}g_{i,t} \\ \vdots \\ \sum_{t=T^0_{m^0}}^{T^0_{m^0+1}-1} \tilde{\mathbf{x}}_{i,t}g_{i,t} \end{bmatrix}, \tag{A.36}$$

an $(m^0+1)p \times 1$ vector. Here $g_{i,t} = \boldsymbol{\nu}'_i\mathbf{f}_t + \varepsilon_{i,t}$, as in Proof of Lemma A.1. In this notation, $\sqrt{N}\mathbf{U}_N(\mathcal{T}^0_{m^0}) = N^{-1/2}\sum_{i=1}^N \mathbf{u}_i$. Let $\mathcal{F}_i$ be the sigma-field generated by $\mathcal{C}$ and $(\mathbf{u}_1, ..., \mathbf{u}_i)$. Then

$\{(\mathbf{u}_i, \mathcal{F}_i) : i \geq 1\}$ is a martingale difference sequence (MDS), because $\mathbf{u}_i$ is independent across i conditional on $\mathcal{C}$, and $\mathbb{E}(\mathbf{u}_i|\mathcal{F}_{i-1}) = \mathbb{E}(\mathbf{u}_i|\mathcal{C}) = \mathbf{0}_{(m^0+1)p\times 1}$ (see, for example, Andrews, 2005, for a similar MDS construction). A conditional Lindeberg condition holds because $\mathbf{u}_i$ have four finite moments. Therefore, by the MDS CLT given in Proposition A.1 of Magdalinos and Phillips (2009),

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{u}_i \to_d MN(\mathbf{0}_{(m^0+1)p\times 1}, \mathbf{\Omega}_0) \tag{A.37}$$

as $N \to \infty$, where $MN(\cdot,\cdot)$ signifies a mixed normal distribution and the $(m^0+1)p \times (m^0+1)p$ matrix $\mathbf{\Omega}_0$ is given by

$$\mathbf{\Omega}_0 = \lim_{N\to\infty}\mathbb{E}\left[\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{u}_i\right)\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{u}_i\right)'|\mathcal{C}\right] = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}(\mathbf{u}_i\mathbf{u}_i'|\mathcal{C}). \tag{A.38}$$

We now characterize this matrix by considering a typical block. In particular, by the same arguments used in Proof of Lemma A.1 to show that $\|N^{-1/2}\sum_{i=1}^{N}\tilde{\mathbf{x}}_{i,t}g_{i,t}\| = O_p(1)$, the (m,n)-th $p\times p$ block of $\mathbf{\Omega}_0$ with $(m,n) \in \{1,\ldots,m^0+1\}$ is given by

$$\begin{aligned}
&\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=T_m^0}^{T_{m+1}^0-1}\sum_{s=T_n^0}^{T_{n+1}^0-1}\mathbb{E}(\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{j,s}'g_{i,t}g_{j,s}|\mathcal{C})\\
&= \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=T_m^0}^{T_{m+1}^0-1}\sum_{s=T_n^0}^{T_{n+1}^0-1}\mathbb{E}[\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{j,s}'\mathbb{E}(g_{i,t}g_{j,s}|\mathcal{X},\mathcal{C})|\mathcal{C}]\\
&= \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=T_m^0}^{T_{m+1}^0-1}\sum_{s=T_n^0}^{T_{n+1}^0-1}\mathbb{E}[\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,s}'\mathbb{E}(g_{i,t}g_{i,s}|\mathcal{X},\mathcal{C})|\mathcal{C}]\\
&= \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=T_m^0}^{T_{m+1}^0-1}\sum_{s=T_n^0}^{T_{n+1}^0-1}\mathbb{E}(\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,s}'|\mathcal{C})(\mathbf{f}_t'\mathbf{\Sigma}_i\mathbf{f}_s + \sigma_{i,t,s}),
\end{aligned} \tag{A.39}$$

where $\sigma_{i,t,s} = \mathbb{E}(\varepsilon_{i,t}\varepsilon_{i,s}|\mathcal{C})$, and $\mathbf{\Sigma}_i$ and $\mathcal{X}$ are as in the proof of Lemma A.1.

The asymptotic mixed normality of $N^{-1/2}\sum_{i=1}^{N}\mathbf{u}_i$ can be used together with Slutsky's theorem to show that

$$\begin{aligned}
\sqrt{N}(\hat{\mathbf{A}}_m - \mathbf{A}_{m^0}^0) &= \mathbf{Q}_0^{-1}\sqrt{N}\mathbf{U}_N(\mathcal{T}_{m^0}^0) + [\mathbf{Q}_N(\mathcal{T}_{m^0}^0)^{-1} - \mathbf{Q}_0^{-1}]\sqrt{N}\mathbf{U}_N(\mathcal{T}_{m^0}^0)\\
&= \mathbf{Q}_0^{-1}\sqrt{N}\mathbf{U}_N(\mathcal{T}_{m^0}^0) + o_p(1)\\
&\to_d MN(\mathbf{0}_{(m^0+1)p\times 1}, \mathbf{Q}_0^{-1}\mathbf{\Omega}_0\mathbf{Q}_0^{-1})
\end{aligned} \tag{A.40}$$

as $N \to \infty$, which is what we wanted to show. ■

Proof of Theorem 4.

Define $\Gamma_- = \{\gamma : \hat{m}(\gamma) < m^0\}$, $\Gamma_+ = \{\gamma : \hat{m}(\gamma) > m^0\}$ and $\Gamma_0 = \{\gamma : \hat{m}(\gamma) = m^0\}$. The idea behind this proof is to show that $\mathbb{P}[\hat{m}(\gamma) = m^0] \to 0$ for all $\hat{m}(\gamma) \neq m^0$, which is equivalent to showing that $\mathbb{P}[\text{IC}(\gamma) > \text{IC}(\gamma^0)] \to 1$, or equivalently

$$\mathbb{P}[\text{IC}(\gamma) - \text{IC}(\gamma^0) > 0] \to 1 \tag{A.41}$$

for $\gamma \in \Gamma_- \cup \Gamma_+$ and $\gamma^0 \in \Gamma_0$. In so doing, it is convenient to split the proof in two cases; $\hat{m}(\hat{\gamma}) < m^0$ and $\hat{m}(\hat{\gamma}) \geq m^0$.

Consider the case when $\hat{m}(\gamma) < m^0$, so that the number of breaks is underspecified. Note first that since $\phi = o(1)$ by assumption, and $\hat{\mathcal{T}}_{\hat{m}}(\gamma^0) = \mathcal{T}^0_{m^0}$ w.p.a.1 by Corollary 1,

$$\begin{aligned} \text{IC}(\gamma) - \text{IC}(\gamma^0) &= \hat{\sigma}^2(\hat{\mathcal{T}}_{\hat{m}}(\gamma)) - \hat{\sigma}^2(\hat{\mathcal{T}}_{\hat{m}}(\gamma^0)) + \phi \cdot p[\hat{m}(\gamma) - \hat{m}(\gamma^0)] \\ &= \hat{\sigma}^2(\hat{\mathcal{T}}_{\hat{m}}(\gamma)) - \hat{\sigma}^2(\mathcal{T}^0_{m^0}) + o_p(1). \end{aligned} \tag{A.42}$$

The proof for the case when $\hat{m}(\gamma) < m^0$ consists of showing that $\hat{\sigma}^2(\mathcal{T}_m) - \hat{\sigma}^2(\mathcal{T}^0_{m^0}) \to_p c > 0$ as $N \to \infty$ for $\mathcal{T}_m = \mathcal{T}_m(\gamma)$ and $\gamma \in \Gamma_-$, which in turn implies

$$\mathbb{P}[\text{IC}(\gamma) - \text{IC}(\gamma^0) > 0] \to 1 \tag{A.43}$$

for $\gamma \in \Gamma_-$.

Let $\sigma_0^2 = (NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T g_{i,t}^2$, where $g_{i,t} = \boldsymbol{\nu}_i' \mathbf{f}_t + \varepsilon_{i,t}$. In this notation,

$$\hat{\sigma}^2(\mathcal{T}_m) - \hat{\sigma}^2(\mathcal{T}^0_{m^0}) = \hat{\sigma}^2(\mathcal{T}_m) - \sigma_0^2 - [\hat{\sigma}^2(\mathcal{T}^0_{m^0}) - \sigma_0^2], \tag{A.44}$$

where

$$\hat{\sigma}^2(\mathcal{T}_m) - \sigma_0^2 = \frac{1}{NT} \sum_{i=1}^N \sum_{j=1}^{m+1} \sum_{t=T_{j-1}}^{T_j - 1} [(\tilde{y}_{i,t} - \tilde{\mathbf{x}}_{i,t}' \hat{\boldsymbol{\alpha}}_j)^2 - g_{i,t}^2]. \tag{A.45}$$

Consider $\hat{\sigma}^2(\mathcal{T}^0_{m^0}) - \sigma_0^2$. From $\tilde{y}_{i,t} = \tilde{\mathbf{x}}'_{i,t}\boldsymbol{\alpha}^0_j + \tilde{u}_{i,t}$,

$$
\begin{aligned}
&\hat{\sigma}^2(\mathcal{T}^0_{m^0}) - \sigma_0^2 \\
&= \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{m^0+1}\sum_{t=T^0_{j-1}}^{T^0_j-1}[(\tilde{y}_{i,t} - \tilde{\mathbf{x}}'_{i,t}\hat{\boldsymbol{\alpha}}_j)^2 - g_{i,t}^2] \\
&= \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{m^0+1}\sum_{t=T^0_{j-1}}^{T^0_j-1}[(\tilde{u}_{i,t} - \tilde{\mathbf{x}}'_{i,t}(\hat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^0_j))^2 - g_{i,t}^2] \\
&= \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{m^0+1}\sum_{t=T^0_{j-1}}^{T^0_j-1}[\tilde{u}_{i,t}^2 - 2\tilde{u}_{i,t}\tilde{\mathbf{x}}'_{i,t}(\hat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^0_j) + (\hat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^0_j)'\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}'_{i,t}(\hat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^0_j) - g_{i,t}^2] \\
&= \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{m^0+1}\sum_{t=T^0_{j-1}}^{T^0_j-1}(\tilde{u}_{i,t}^2 - g_{i,t}^2) - \frac{2}{NT}\sum_{i=1}^{N}\sum_{j=1}^{m^0+1}\sum_{t=T^0_{j-1}}^{T^0_j-1}\tilde{u}_{i,t}\tilde{\mathbf{x}}'_{i,t}(\hat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^0_j) \\
&+ \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{m^0+1}\sum_{t=T^0_{j-1}}^{T^0_j-1}(\hat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^0_j)'\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}'_{i,t}(\hat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^0_j) \\
&= M_1 - 2M_2 + M_3, \qquad \text{(A.46)}
\end{aligned}
$$

with implicit definitions of M_1, M_2 and M_3. For M_1, we use $\tilde{u}_{i,t} = g_{i,t} - \overline{g}_t$, where $\overline{g}_t = \overline{\boldsymbol{\nu}}'\mathbf{f}_t + \overline{\varepsilon}_t$, giving

$$
\begin{aligned}
M_1 &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{m^0+1}\sum_{t=T^0_{j-1}}^{T^0_j-1}(\tilde{u}_{i,t}^2 - g_{i,t}^2) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{m^0+1}\sum_{t=T^0_{j-1}}^{T^0_j-1}[(g_{i,t} - \overline{g}_t)^2 - g_{i,t}^2] \\
&= \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{m^0+1}\sum_{t=T^0_{j-1}}^{T^0_j-1}(-2g_{i,t}\overline{g}_t + \overline{g}_t^2) = -\frac{1}{T}\sum_{j=1}^{m^0+1}\sum_{t=T^0_{j-1}}^{T^0_j-1}\overline{g}_t^2 = -\frac{1}{T}\sum_{t=1}^{T}\overline{g}_t^2. \qquad \text{(A.47)}
\end{aligned}
$$

Hence, since $\|\mathbf{f}_t\| < \infty$ w.p.1, $\|\overline{\boldsymbol{\nu}}\| = O_p(N^{-1/2})$ and $\|\overline{\varepsilon}_t\| = O_p(N^{-1/2})$ for all t,

$$
|M_1| = \frac{1}{T}\sum_{t=1}^{T}\overline{g}_t^2 \leq \frac{2}{T}\sum_{t=1}^{T}(\|\overline{\boldsymbol{\nu}}\|^2\|\mathbf{f}_t\|^2 + \overline{\varepsilon}_t^2) = O_p(N^{-1}). \qquad \text{(A.48)}
$$

For M_2,

$$
\begin{aligned}
|M_2| &= \left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{m^0+1}\sum_{t=T^0_{j-1}}^{T^0_j-1}\tilde{u}_{i,t}\tilde{\mathbf{x}}'_{i,t}(\hat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^0_j)\right| \\
&\leq \left(\sum_{j=1}^{m^0+1}\left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=T^0_{j-1}}^{T^0_j-1}\tilde{u}_{i,t}\tilde{\mathbf{x}}'_{i,t}\right\|^2\right)^{1/2}\left(\sum_{j=1}^{m^0+1}\|\hat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}^0_j\|^2\right)^{1/2} = O_p(N^{-1}), \qquad \text{(A.49)}
\end{aligned}
$$

which holds because $\|N^{-1}\sum_{i=1}^{N}\tilde{u}_{i,t}\tilde{\mathbf{x}}_{i,t}'\| = O_p(N^{-1/2})$ by Proof of Lemma A.1 and $\|\hat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}_j^0\| = O_p(N^{-1/2})$ by Theorem 1. The order of M_3 is the same, as is clear from

$$|M_3| \leq \left(\sum_{j=1}^{m^0+1} \left\| \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=T_{j-1}^0}^{T_j^0-1} \tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,t}' \right\|^2 \right)^{1/2} \left(\sum_{j=1}^{m^0+1} \|\hat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}_j^0\|^4 \right)^{1/2} = O_p(N^{-1}). \tag{A.50}$$

It follows that

$$|\hat{\sigma}^2(\mathcal{T}_{m^0}^0) - \sigma_0^2| \leq |M_1| + 2|M_2| + |M_3| = O_p(N^{-1}). \tag{A.51}$$

Next up is $\hat{\sigma}^2(\mathcal{T}_0) - \sigma_0^2$. Suppose for simplicity that $m = 0 < m^0 = 1$. By using $\tilde{y}_{i,t} = \tilde{\mathbf{x}}_{i,t}'\boldsymbol{\beta}_t^0 + \tilde{u}_{i,t}$ and the fact that $|M_1| = O_p(N^{-1})$,

$$\begin{aligned}
\hat{\sigma}^2(\mathcal{T}_0) - \sigma_0^2 &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}[(\tilde{y}_{i,t} - \tilde{\mathbf{x}}_{i,t}'\hat{\boldsymbol{\alpha}}_1)^2 - g_{i,t}^2] \\
&= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}[\tilde{u}_{i,t}^2 - 2\tilde{u}_{i,t}\tilde{\mathbf{x}}_{i,t}'(\hat{\boldsymbol{\alpha}}_1 - \boldsymbol{\beta}_t^0) + (\hat{\boldsymbol{\alpha}}_1 - \boldsymbol{\beta}_t^0)'\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,t}'(\hat{\boldsymbol{\alpha}}_1 - \boldsymbol{\beta}_t^0) - g_{i,t}^2] \\
&= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{u}_{i,t}^2 - g_{i,t}^2) - \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{i,t}\tilde{\mathbf{x}}_{i,t}'(\hat{\boldsymbol{\alpha}}_1 - \boldsymbol{\beta}_t^0) \\
&+ \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\hat{\boldsymbol{\alpha}}_1 - \boldsymbol{\beta}_t^0)'\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,t}'(\hat{\boldsymbol{\alpha}}_1 - \boldsymbol{\beta}_t^0) \\
&= -\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{i,t}\tilde{\mathbf{x}}_{i,t}'(\hat{\boldsymbol{\alpha}}_1 - \boldsymbol{\beta}_t^0) + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\hat{\boldsymbol{\alpha}}_1 - \boldsymbol{\beta}_t^0)'\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,t}'(\hat{\boldsymbol{\alpha}}_1 - \boldsymbol{\beta}_t^0) \\
&+ O_p(N^{-1}).
\end{aligned} \tag{A.52}$$

Note how

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\mathbf{x}}_{i,t}\tilde{y}_{i,t} &= \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_1^0}\tilde{\mathbf{x}}_{i,t}\tilde{y}_{i,t} + \frac{1}{N}\sum_{i=1}^{N}\sum_{t=T_1^0+1}^{T}\tilde{\mathbf{x}}_{i,t}\tilde{y}_{i,t} \\
&= \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_1^0}\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,t}'\boldsymbol{\alpha}_1^0 + \frac{1}{N}\sum_{i=1}^{N}\sum_{t=T_1^0+1}^{T}\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,t}'\boldsymbol{\alpha}_2^0 + \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\mathbf{x}}_{i,t}\tilde{u}_{i,t} \\
&= \mathbf{Q}_{0,1}\boldsymbol{\alpha}_1^0 + \mathbf{Q}_{0,2}\boldsymbol{\alpha}_2^0 + o_p(1),
\end{aligned} \tag{A.53}$$

where $\mathbf{Q}_{0,j} = \lim_{N\to\infty} N^{-1}\sum_{i=1}^{N}\sum_{t=T_{j-1}^0}^{T_j^0-1}\mathbb{E}(\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,t}'|\mathcal{C})$. Hence, letting $\mathbf{Q}_0 = \mathbf{Q}_0(\mathcal{T}_1^0)$ and $\boldsymbol{\alpha}_1^* = \mathbf{Q}_0^{-1}(\mathbf{Q}_{0,1}\boldsymbol{\alpha}_1^0 + \mathbf{Q}_{0,2}\boldsymbol{\alpha}_2^0)$,

$$\hat{\boldsymbol{\alpha}}_1 = \left(\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,t}'\right)^{-1}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\mathbf{x}}_{i,t}\tilde{y}_{i,t} = \boldsymbol{\alpha}_1^* + o_p(1) \tag{A.54}$$

It follows that

$$\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{u}_{i,t}\tilde{\mathbf{x}}_{i,t}'(\hat{\boldsymbol{\alpha}}_1-\boldsymbol{\beta}_t^0)\right| \leq \left(\frac{1}{T}\sum_{t=1}^{T}\left\|\frac{1}{N}\sum_{i=1}^{N}\tilde{u}_{i,t}\tilde{\mathbf{x}}_{i,t}'\right\|^2\right)^{1/2}\left(\frac{1}{T}\sum_{t=1}^{T}\|\hat{\boldsymbol{\alpha}}_1-\boldsymbol{\beta}_t^0\|^2\right)^{1/2}$$
$$= O_p(N^{-1/2}), \tag{A.55}$$

and

$$\begin{aligned}
&\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\hat{\boldsymbol{\alpha}}_1-\boldsymbol{\beta}_t^0)'\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,t}'(\hat{\boldsymbol{\alpha}}_1-\boldsymbol{\beta}_t^0)\\
&=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T_1^0}(\hat{\boldsymbol{\alpha}}_1-\boldsymbol{\alpha}_1^0)'\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,t}'(\hat{\boldsymbol{\alpha}}_1-\boldsymbol{\alpha}_1^0)+\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=T_1^0+1}^{T}(\hat{\boldsymbol{\alpha}}_1-\boldsymbol{\alpha}_2^0)'\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,t}'(\hat{\boldsymbol{\alpha}}_1-\boldsymbol{\alpha}_2^0)\\
&=\frac{1}{T}[(\boldsymbol{\alpha}_1^*-\boldsymbol{\alpha}_1^0)'\mathbf{Q}_{0,1}(\boldsymbol{\alpha}_1^*-\boldsymbol{\alpha}_1^0)+(\boldsymbol{\alpha}_1^*-\boldsymbol{\alpha}_2^0)'\mathbf{Q}_{0,2}(\boldsymbol{\alpha}_1^*-\boldsymbol{\alpha}_2^0)]+o_p(1)\\
&\to_p c>0
\end{aligned} \tag{A.56}$$

as $N\to\infty$, where c is simply the sum of the first two terms on the right. Direct insertion into the above expression for $\hat{\sigma}^2(\mathcal{T}_0)-\sigma_0^2$ gives

$$\hat{\sigma}^2(\mathcal{T}_0)-\sigma_0^2=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\hat{\boldsymbol{\alpha}}_1-\boldsymbol{\beta}_t^0)'\tilde{\mathbf{x}}_{i,t}\tilde{\mathbf{x}}_{i,t}'(\hat{\boldsymbol{\alpha}}_1-\boldsymbol{\beta}_t^0)+O_p(N^{-1/2})\to_p c>0. \tag{A.57}$$

This last result holds not only in the simple case considered here but in general when $m<m^0$, although the exact definition of c will vary. It follows that

$$\begin{aligned}
\hat{\sigma}^2(\mathcal{T}_m)-\hat{\sigma}^2(\mathcal{T}_{m^0}^0)&=\hat{\sigma}^2(\mathcal{T}_m)-\sigma_0^2-[\hat{\sigma}^2(\mathcal{T}_{m^0}^0)-\sigma_0^2]=\hat{\sigma}^2(\mathcal{T}_m)-\sigma_0^2+O_p(N^{-1})\\
&\to_p c>0,
\end{aligned} \tag{A.58}$$

which is what we set out to show. This establishes the required result for the case when $\hat{m}(\gamma)<m^0$.

Suppose now that $\hat{m}(\gamma)\geq m^0$, so that the number of breaks is not underspecified. We have already shown that $\hat{\sigma}^2(\mathcal{T}_{m^0}^0)-\sigma_0^2=O_p(N^{-1})$. If $\gamma\in\Gamma_+$, $\hat{\sigma}^2(\mathcal{T}_m)$ contain estimated regimes in between which there are no breaks. But since the slope estimates are consistent even if there are no breaks, $\hat{\sigma}^2(\mathcal{T}_m)-\sigma_0^2=O_p(N^{-1})$ even if $\gamma\in\Gamma_+$. It follows that

$$\hat{\sigma}^2(\mathcal{T}_m)-\hat{\sigma}^2(\mathcal{T}_{m^0}^0)=\hat{\sigma}^2(\mathcal{T}_m)-\sigma_0^2-[\hat{\sigma}^2(\mathcal{T}_{m^0}^0)-\sigma_0^2]=O_p(N^{-1}). \tag{A.59}$$

By using this,

$$\frac{1}{\phi}[\mathrm{IC}(\gamma) - \mathrm{IC}(\gamma^0)] = \frac{1}{\phi}[\hat{\sigma}^2(\hat{\mathcal{T}}_{\hat{m}}(\gamma)) - \hat{\sigma}^2(\hat{\mathcal{T}}_{\hat{m}}(\gamma^0))] + p[\hat{m}(\gamma) - \hat{m}(\gamma^0)], \tag{A.60}$$

and the fact that $\phi > 0$, we can show that

$$\begin{aligned}
&\lim_{N\to\infty} \mathbb{P}[\mathrm{IC}(\gamma) - \mathrm{IC}(\gamma^0) > 0] \\
&= \lim_{N\to\infty} \mathbb{P}\left(\frac{1}{\phi}[\mathrm{IC}(\gamma) - \mathrm{IC}(\gamma^0)] > 0\right) \\
&= \lim_{N\to\infty} \mathbb{P}\left(\frac{1}{N\phi} N[\hat{\sigma}^2(\hat{\mathcal{T}}_{\hat{m}}(\gamma)) - \hat{\sigma}^2(\hat{\mathcal{T}}_{\hat{m}}(\gamma^0))] + p[\hat{m}(\gamma) - \hat{m}(\gamma^0)] > 0\right) \\
&= \lim_{N\to\infty} \mathbb{P}(p[\hat{m}(\gamma) - \hat{m}(\gamma^0)] > 0) = 1
\end{aligned} \tag{A.61}$$

for all $\gamma \in \Gamma_+$, where the third equality is due to $N[\hat{\sigma}^2(\hat{\mathcal{T}}_{\hat{m}}(\gamma)) - \hat{\sigma}^2(\hat{\mathcal{T}}_{\hat{m}}(\gamma^0))] = O_p(1)$ and $N\phi \to \infty$ by assumption, while the fourth and last equality is due to the fact that $\hat{m}(\gamma) - \hat{m}(\gamma^0) > 0$ for all $\gamma \in \Gamma_+$. We have therefore shown that

$$\mathbb{P}[\mathrm{IC}(\gamma) - \mathrm{IC}(\gamma^0) > 0] \to 1 \tag{A.62}$$

for all $\gamma \in \Gamma_- \cup \Gamma_+$. In other words, the minimizer of $\mathrm{IC}(\gamma)$ cannot be given by $\gamma \in \Gamma_- \cup \Gamma_+$ w.p.a.1, but can only be given by $\gamma \in \Gamma_0$ w.p.a.1. ■

References

Andrews, D. W. K. (2005). Cross-Section Regression with Common Shocks. *Econometrica* **73**, 1551–1585.

Magdalinos, T., and P. C. Phillips (2009). Limit Theory for Cointegrated Systems with Moderately Integrated and Moderately Explosive Regressors. *Econometric Theory* **25**, 482–526.