3-2016

# Shrinkage estimation of common breaks in panel data models via adaptive group fused Lasso

Junhui QIAN
*Shanghai Jiaotong University*

Liangjun SU
*Singapore Management University*, ljsu@smu.edu.sg

## Citation

QIAN, Junhui and SU, Liangjun. Shrinkage estimation of common breaks in panel data models via adaptive group fused Lasso. (2016). *Journal of Econometrics*. 191, (1), 86-109.
**Available at:** https://ink.library.smu.edu.sg/soe_research/1745

Published in Journal of Econometrics, 2016 March
Volume 191, Issue 1, Pages 86–109
http://dx.doi.org/10.1016/j.jeconom.2015.09.004

# Shrinkage Estimation of Common Breaks in Panel Data Models via Adaptive Group Fused Lasso*

Junhui Qian[a] and Liangjun Su[b]
[a] Antai College of Economics and Management, Shanghai Jiao Tong University
[b] School of Economics, Singapore Management University

September 24, 2015

### Abstract

In this paper we consider estimation and inference of common breaks in panel data models via adaptive group fused lasso. We consider two approaches – penalized least squares (PLS) for first-differenced models without endogenous regressors, and penalized GMM (PGMM) for first-differenced models with endogeneity. We show that with probability tending to one both methods can correctly determine the unknown number of breaks and estimate the common break dates consistently. We establish the asymptotic distributions of the Lasso estimators of the regression coefficients and their post Lasso versions. We also propose and validate a data-driven method to determine the tuning parameter used in the Lasso procedure. Monte Carlo simulations demonstrate that both the PLS and PGMM estimation methods work well in finite samples. We apply our PGMM method to study the effect of foreign direct investment (FDI) on economic growth using a panel of 88 countries and regions from 1973 to 2012 and find multiple breaks in the model.

**JEL Classification:** C13, C23, C33, C51

**Key Words:** Adaptive Lasso; Change point; Group fused Lasso; Panel data; Penalized least squares; Penalized GMM; Structural change

*The authors would like to thank the co-editor Jianqing Fan, an associate editor, and three anonymous referees for their constructive comments and suggestions. They also express their sincere appreciation to Xiaohong Chen, Han Hong, and Chihwa Kao for discussions on the subject matter and participants in the 10th Symposium on Econometric Theory and Applications (SETA 2014, Taipei) and the 4th Shanghai Econometrics Workshop for helpful comments. Su gratefully acknowledges the Singapore Ministry of Education for Academic Research Fund under grant number MOE2012-T2-2-021. Address Correspondence to: Liangjun Su, School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903; E-mail: ljsu@smu.edu.sg, Phone: +65 6828 0386.

# 1 Introduction

Recently there has been a growing literature on the estimation and tests of common breaks in panel data models in which there are $N$ individual units and $T$ time series observations for each individual. Depending on whether $T$ is allowed to pass to infinity, the model is called "short" for fixed $T$ and "large" (or of large dimension) if $T$ passes to infinity. Implicitly, one usually allows $N$ to pass to infinity in panel data models.[1] Most of the literature falls into two categories depending on whether the parameters of interest are allowed to be heterogenous across individuals or not. The first category focuses on homogenous panel data models and includes De Watcher and Tzavalis (2005), Baltagi et al. (2015), and De Watcher and Tzavalis (2012). De Watcher and Tzavalis (2005) compare the relative performance of two model and moment selection methods in detecting breaks in short panels; Baltagi et al. (2015) consider the estimation and identification of change points in large dimensional panel models with either stationary or nonstationary regressors and error terms; De Watcher and Tzavalis (2012) develop a testing procedure for common breaks in short linear dynamic panel data models. The second category considers estimation and inference of common breaks in heterogenous panel data models; see Bai (2010), Kim (2011, 2014), Hsu and Lin (2012), Baltagi et al. (2014), among others. Bai (2010) establishes the asymptotic properties of the estimated break point in a location-scale heterogenous panel data model with either fixed or large $T$; Kim (2011) extends Bai's (2010) method and develops an estimation procedure for a common deterministic time trend break in large heterogenous panels with a multi-factor error structure; Kim (2014) continues the study by estimating the common break date and common factors jointly; Hsu and Lin (2012) extends Bai's (2010) theory to nonstationary panel data models where the error terms follow an I(1) process; Baltagi et al. (2014) study the estimation of large dimensional static heterogenous panels with a common break by extending Pesaran's (2006) common correlated effects (CCE) estimation procedure. In addition, Chan et al. (2008) extend the testing procedure of Andrews (2003) from time series to heterogenous panels where the breaks may occur at different time points across individuals; Liao and Wang (2012) study the estimation of individual-specific structural breaks that exhibit a common distribution in a location-scale panel data model; Yamazaki and Kurozumi (2014) develop an LM-type test for slope homogeneity along the time dimension in fixed-effects panel data models with fixed $N$ and large $T$.[2]

A common feature of all of the above works is that a one-time break, common or not, is assumed in the estimation procedure. Although the assumption of a single break greatly facilitates the estimation and inference procedure, inferences based on it could be misleading if the underlying model has an unknown number of multiple breaks. For this reason, a large literature on the estimation and inference of models with multiple structural changes has been developed in the single or multiple time series framework; see, e.g., Bai (1997a, 1997b), Bai and Perron (1998), Qu and Perron (2007), Su and White (2010), Kurozumi

[1] Bai (1997a), Bai et al. (1998) and Qu and Perron (2007) extend the estimation of single-time series models to multiple-ones with simultaneous structural breaks where the number of equations is fixed.

[2] Pesaran and Yamagata (2008) and Su and Chen (2013) propose LM-type tests for slope homogeneity along the cross section dimension in large dimensional linear panel data models with additive fixed effects and interactive fixed effects, respectively.

(2015), and Qian and Su (2014, 2015). In view of the fact that the conventional *avg*- and *exp*-type test statistics for multiple structural changes requires all permissible partitions of the sample which could be prohibitively large, Qian and Su (2015) propose shrinkage estimation of regression models with multiple structural changes by extending the fused Lasso of Tibshirani et al. (2005) to the time series regression framework.

In this paper we propose a shrinkage-based methodology for estimating panel data models with an unknown number of structural changes. The new methodology is most suitable for the vision that the regression coefficients in a panel data model may be time-varying but at the same time exhibit certain sparseness in abrupt changes or breaks. This vision seems pertinent in many applied studies using panel data that have a long time span measured in decades. During such a long time span, shocks to technologies, preferences, policies, and so on, may result in the change of a statistical relation applied economists seek to discover; but the shocks tend to be small over a relatively short time interval so that it does not alter the statistical relationship in short time. In this case, one has to allow the parameters in the model to change over time in an unknown way and recognize that parameters do not always alter from one time period to another one. Multiple structural breaks may occur during the whole time span but the number of breaks is generally small in comparison with the total number of time periods in the data, resulting in the sparseness of the breaks.

In terms of econometrics methodology, this paper extends the Lasso-type shrinkage approach in Qian and Su (2015) to panel data settings. To the best of our knowledge, this is the first in the literature to deal with panel data models with possibly multiple structural changes explicitly.[3] To stay focused, we consider homogenous linear panel data models with an unknown number of common breaks and we do not allow cross section dependence. The extension to heterogenous panel data models and to panel data models with cross section dependence will be discussed at the end of Section 7. For the advantage of the use of panel data to study common breaks, we refer the readers directly to Bai (2010) and De Watcher and Tzavalis (2012). Despite the fact that the Lasso-type shrinkage estimation has a long history and wide applications in statistics (see, e.g., Tibshirani 1996; Knight and Fu 2000; Fan and Li 2001), the application of Lasso-type shrinkage techniques in econometrics has a relatively short history. But the number of applications in econometrics has been increasing very fast in the last few years. For example, Caner (2009) and Fan and Liao (2014) consider covariate selection in GMM estimation; Belloni et al. (2012) and García (2011) consider selection of instruments in the GMM framework; Liao (2013) provides a shrinkage GMM method for moment selection and Cheng and Liao (2015) consider the selection of valid and relevant moments via penalized GMM; Liao and Phillips (2015) apply adaptive shrinkage techniques to cointegrated systems; Kock (2013) considers Bridge estimators of static linear panel data models with random or fixed effects; Caner and Knight (2013) apply Bridge estimators to differentiate a unit root from

[3] Bai (2010, Section 6) discusses the case of multiple breaks. As he remarks, if the number of breaks is given, the one-at-a-time approach of Bai (1997b) can be used to estimate the break dates, and if the number of breaks is unknown, a test for existence of break point can be applied to each subsample before estimating a break point. Alternatively, one can use information criteria to determine the number of breaks in the latter case, but further investigation is called for.

a stationary alternative; Caner and Han (2014) proposes a Bridge estimator for pure factor models and shows the selection consistency; Lu and Su (2015b) apply adaptive group Lasso to choose both regressors and the number of factors in panel data models with factor structures; Cheng et al. (2015) provide an adaptive group Lasso estimator for pure factor structures with a one-time structural break. This paper adds to the literature by applying the shrinkage idea to panel data models with an unknown number of breaks.

We propose two approaches, penalized least squares (PLS) and penalized general method of moments (PGMM), for the estimation of the panel data model with an unknown number of breaks. We apply first differencing to remove the fixed effects in the equation and focus on the first-differenced equation. When there is no endogeneity issue in the first-differenced equation, we propose to apply PLS to estimate the unknown number of break points and the regime-specific regression coefficients jointly where the penalty term is imposed through the adaptive group fused Lasso (AGFL) component. In the presence of endogeneity in the first-differenced equation, which may arise from endogenous regressors or lagged dependent variables in the original fixed-effects equation, we propose to apply PGMM to estimate the unknown number of break points and the regime-specific regression coefficients jointly where, again, the penalty term is imposed through the AGFL component. Unlike Qian and Su (2015) who can only establish the claim that the group fused Lasso can not under-estimate the number of breaks in a time series regression and that all the *break fractions* (but not the break dates) can be consistently estimated as in Bai and Perron (1998), we show that with probability approaching one (w.p.a.1) both of our PLS and PGMM methods can correctly determine the unknown number of breaks and estimate the common *break dates* consistently. We obtain estimates of the regression coefficients via both the Lasso and post Lasso procedures and establish their asymptotic distributions. We also propose and validate a data-driven method to determine the tuning parameter used in the Lasso procedure.

Both PLS and PGMM can be numerically solved using the fast block-coordinate descent algorithm. Monte Carlo simulations show that our methods perform well in finite samples. First, the probability of correctly estimating the number of breaks converges to one quickly as $N$ increases. Second, conditional on the correct estimation of the number of breaks, our methods accurately estimate the break dates in finite samples. Third, our method continues to perform well even if the number of breaks is allowed to increase with the time dimension.

As an empirical illustration, we employ our PGMM method to evaluate the effect of foreign direct investment (FDI) inflow on economic growth. We estimate a dynamic panel data model with possibly multiple breaks using the PGMM approach. We find that, with a tuning parameter selected via minimizing a BIC-type information criterion, there are three breaks (four regimes) in the span of seven five-year periods. In each regime, the post-Lasso estimation finds significant positive effect of FDI inflow on GDP growth. In contrast, if we estimate a usual dynamic panel data model with time-invariant parameters, we would find this effect to be statistically insignificant. This empirical example illustrates the perils of employing panel data models with restrictions on the number of breaks. Our contribution makes the restriction unnecessary.

It is worth mentioning that Ke et al. (2015) and Su et al. (2014) investigate similar problems to ours. Ke et al. (2015) proposes a new method called clustering algorithm in regression via data-driven segmentation (CARDS) to explore homogeneity of coefficients in high dimensional regression. Su et al. (2014) propose a procedure called classifier Lasso to estimate a latent panel structure where individuals belong to a number of homogenous groups within a broadly heterogenous population, regression parameters are the same within each group but differ across groups, and the individual's group membership is unknown. But neither paper requires the temporal ordering information.

The rest of the paper is organized as follows. Section 2 introduces our fixed-effect panel data model and PLS and PGMM estimation of the model depending on whether endogeneity is present in the first-differenced equation. Sections 3 and 4 analyze the asymptotic properties of PLS and PGMM estimators, respectively. Section 5 reports the Monte Carlo simulation results. Section 6 provides an empirical application and Section 7 concludes.

NOTATION. Throughout the paper we adopt the following notation. For an $m \times n$ real matrix $A$, we denote its transpose as $A'$, its Frobenius norm as $\|A\|$, and its spectral norm as $\|A\|_{\mathrm{sp}}$. When $A$ is symmetric, we use $\mu_{\max}(A)$ and $\mu_{\min}(A)$ to denote its largest and smallest eigenvalues, respectively. $\mathbb{I}_p$ denotes a $p \times p$ identity matrix and $\mathbf{0}_{a\times b}$ an $a \times b$ matrix of zeros. We use "p.d." and "p.s.d." abbreviate "positive definite" and "positive semi-definite", respectively. The operator $\xrightarrow{P}$ denotes convergence in probability, $\xrightarrow{D}$ convergence in distribution, and plim probability limit. Let $\Delta$ and $\Delta^2$ denote the difference operators of order 1 and 2, respectively. In addition, we use $\mathrm{TriD}(\cdot,\cdot)_T$ to denote a *symmetric block tridiagonal matrix* (SBTM):

$$\mathrm{TriD}(A,D)_T \equiv \begin{pmatrix} D_1 & -A_2' & & & & & \\ -A_2 & D_2 & -A_3' & & & & \\ & -A_3 & D_3 & -A_4' & & & \\ & \ddots & \ddots & \ddots & & & \\ & & & & -A_{T-1} & D_{T-1} & -A_T' \\ & & & & & -A_T & D_T \end{pmatrix} \tag{1.1}$$

where $D_t$'s are symmetric, $A_t$'s are square matrices, and empty blocks denote the matrices of zeros. By Molinari (2008), the determinant of $\mathrm{TriD}(A,D)_T$ is given by $\det(\mathrm{TriD}(A,D)_T) = \prod_{t=1}^{T} \det(\Lambda_t)$, where $\Lambda_1 = D_1$ and $\Lambda_t = D_t - A_t\Lambda_{t-1}^{-1}A_t'$ for $t = 2, ..., T$. By Meurant (1992) and Ran and Huang (2006), one can also calculate the inverse of $\mathrm{TriD}(A,D)_T$ recursively.

# 2 Shrinkage estimation of linear panel data models with multiple breaks

In this section we consider a linear panel data model with an unknown number of breaks, which we estimate via the adaptive group fused Lasso.

## 2.1 The model

Consider the following linear panel data model

$$y_{it} = \mu_i + \beta_t' x_{it} + u_{it},\ i = 1, ..., N,\ t = 1, \ldots, T \geq 2, \tag{2.1}$$

where $x_{it}$ is a $p \times 1$ vector of regressors, $u_{it}$ is the error term with zero mean, $\beta_t$ is a $p \times 1$ vector of unknown coefficients, and $\mu_i$ is the individual fixed effects that may be correlated with $x_{it}$. We assume that $N$ passes to infinity and $T$ can either be fixed or pass to infinity. But for clarity, we will focus on the case where $T \to \infty$ when we derive and report the theoretical results. It is evident from the derivation that all results hold under the fixed $T$ case. When $T \to \infty$, we will write $(N, T) \to \infty$ to signify that $N$ and $T$ pass to infinity jointly.

Like Qian and Su (2015), we assume that $\{\beta_1, ..., \beta_T\}$ exhibit certain *sparsity* such that the total number of distinct vectors in the set is given by $m + 1$, which is unknown but typically much smaller than $T$. More specifically, we assume that

$$\beta_t = \alpha_j \text{ for } t = T_{j-1}, ..., T_j - 1 \text{ and } j = 1, ..., m + 1$$

where we adopt the convention that $T_0 = 1$ and $T_{m+1} = T + 1$. The indices $T_1, ..., T_m$ indicate the unobserved $m$ break points/dates and the number $m + 1$ denotes the total number of regimes. We are interested in estimating the *unknown* number $m$ of *unknown* break dates and the regression coefficients. Let $\boldsymbol{\alpha}_m = (\alpha_1', ..., \alpha_{m+1}')'$ and $\mathcal{T}_m = \{T_1, ..., T_m\}$.

Throughout, we denote the true value of a parameter with a superscript 0. In particular, we use $m^0$, $\boldsymbol{\alpha}^0_{m^0} = \left(\alpha_1^{0\prime}, ..., \alpha_{m^0+1}^{0\prime}\right)'$ and $\mathcal{T}^0_{m^0} = \{T_1^0, ..., T_{m^0}^0\}$ to denote the true number of breaks, the vector of true regression coefficients, and the set of true break dates, respectively. We assume $T_1^0 \geq 2$ and allow $T^0_{m^0} = T$. When $T^0_{m^0} = T$, the last break occurs at the end of the sample (c.f., Andrews 2003) and the $(m^0 + 1)$th regime has only one observation for each individual time series. As for the true number of breaks, we allow $m^0 \to \infty$ as $T \to \infty$. Of course, in the case of fixed $T$, $m^0$ is regarded as a fixed finite integer. As for the break sizes, we allow the minimum break size $(\min_{1 \leq j \leq m^0} \left\|\alpha^0_{j+1} - \alpha^0_j\right\|)$ to shrink to zero as $N \to \infty$ or $(N, T) \to \infty$. In either case, one may write $y_{it} = y_{it,NT}$ for $1 \leq i \leq N$ and $1 \leq t \leq T$ (and similarly for $x_{it}$) to emphasize the multi-array nature of the process $\{y_{it}\}$. But for notational simplicity, we keep writing $y_{it}$ and $x_{it}$ instead.

To eliminate the effect of $\mu_i$ in the estimation procedure, we consider the first-differenced equation

$$\begin{aligned}\Delta y_{it} &= \beta_t' x_{it} - \beta_{t-1}' x_{i,t-1} + \Delta u_{it} \\ &= \beta_t' \Delta x_{it} + \left(\beta_t - \beta_{t-1}\right)' x_{i,t-1} + \Delta u_{it},\end{aligned}$$

where, e.g., $\Delta y_{it} = y_{it} - y_{i,t-1}$ for $i = 1, ..., N$ and $t = 2, ..., T$. We consider two cases:
**(a)** $E\left[\Delta u_{it} x_{it}\right] = 0$ and $E\left[\Delta u_{it} x_{i,t-1}\right] = 0$;
**(b)** $E\left[\Delta u_{it} x_{it}\right] \neq 0$ or $E\left[\Delta u_{it} x_{i,t-1}\right] \neq 0$.
Case (a) occurs when $x_{it}$ is strictly exogenous in the sense that $E\left(u_{it}|x_i\right) = 0$ a.s. where $x_i = (x_{i1}, ..., x_{iT})'$. But strict exogeneity is not necessary for case (a) and a sufficient condition for (a) to

hold is $E\left(\Delta u_{it}|x_{it},x_{it-1}\right)=0$. Case (b) occurs when $x_{it}$ contains either lagged dependent variables (e.g., $y_{i,t-1}$) or endogenous regressors that are correlated with $u_{it}$. We assume the existence of a $q\times 1$ vector of instruments $z_{it}$ in case (b) where $q\geq p$.

Note that neither $m$ nor the break dates are known and $m$ is typically much smaller than $T$. This fact motivates us to consider the estimation of $\beta_t$'s and $\mathcal{T}_m$ via a variant of fused Lasso *a la* Tibshirani et al. (2005). We propose two approaches – PLS estimation for case (a) and PGMM estimation for case (b).

## 2.2 Penalized least squares (PLS) estimation

In case (a), we propose to estimate $\boldsymbol{\beta}=\left(\beta_1',...,\beta_T'\right)'$ by minimizing the following PLS objective function

$$V_{1NT,\lambda_1}\left(\boldsymbol{\beta}\right)=\frac{1}{N}\sum_{i=1}^{N}\sum_{t=2}^{T}\left(\Delta y_{it}-\beta_t'x_{it}+\beta_{t-1}'x_{i,t-1}\right)^2+\lambda_1\sum_{t=2}^{T}\dot{w}_t\left\|\beta_t-\beta_{t-1}\right\| \tag{2.2}$$

where $\lambda_1=\lambda_1\left(N,T\right)\geq 0$ is a tuning parameter, and $\dot{w}_t$ is a data-driven weight defined by

$$\dot{w}_t=\left\|\dot{\beta}_t-\dot{\beta}_{t-1}\right\|^{-\kappa_1},\ t=2,...,T, \tag{2.3}$$

$\{\dot{\beta}_t\}$ are preliminary estimates of $\{\beta_t\}$, and $\kappa_1$ is an user-specified positive constant that usually takes value 2 in the literature. Noting that the objective function in (2.2) is convex in $\boldsymbol{\beta}$, it is easy to obtain the solution $\tilde{\boldsymbol{\beta}}=(\tilde{\beta}_1',...,\tilde{\beta}_T')'$ where we suppress the dependence of $\tilde{\beta}_t=\tilde{\beta}_t\left(\lambda_1\right)$ on $\lambda_1$ as long as no confusion arises. We will propose a data-driven method to choose $\lambda_1$ in Section 3.4.

For a given solution $\{\tilde{\beta}_t\}$, the set of estimated break dates are given by $\tilde{\mathcal{T}}_{\tilde{m}}=\{\tilde{T}_1,...,\tilde{T}_{\tilde{m}}\}$ where $1<\tilde{T}_1<...<\tilde{T}_{\tilde{m}}\leq T$ such that $\left\|\tilde{\beta}_t-\tilde{\beta}_{t-1}\right\|\neq 0$ at $t=\tilde{T}_j$ for some $j\in\{1,...,\tilde{m}\}$ and $\tilde{\mathcal{T}}_{\tilde{m}}$ divides the time interval $[1,T]$ into $\tilde{m}+1$ regimes such that the parameter estimates remain constant within each regime. Let $\tilde{T}_0=1$ and $\tilde{T}_{\tilde{m}+1}=T+1$. Define $\tilde{\alpha}_j=\tilde{\alpha}_j(\tilde{\mathcal{T}}_{\tilde{m}})=\tilde{\beta}_{\tilde{T}_{j-1}}$ as the estimate of $\alpha_j$ for $j=1,...,\tilde{m}+1$. Frequently we suppress the dependence of $\tilde{\alpha}_j$ on $\tilde{\mathcal{T}}_{\tilde{m}}$ (and $\lambda_1$) unless necessary. Let $\tilde{\boldsymbol{\alpha}}_{\tilde{m}}=\tilde{\boldsymbol{\alpha}}_{\tilde{m}}(\tilde{\mathcal{T}}_{\tilde{m}})=(\tilde{\alpha}_1(\tilde{\mathcal{T}}_{\tilde{m}})',...,\tilde{\alpha}_{\tilde{m}+1}(\tilde{\mathcal{T}}_{\tilde{m}})')'$.

Apparently, the objective function in (2.2) is closely related to the literature on adaptive Lasso (Zou 2006), group Lasso (Yuan and Lin 2006), fused Lasso (Tibshirani et al. 2005 and Rinaldo 2009), and group fused Lasso (Qian and Su 2015). Zou (2006) first shows that the Lasso could be inconsistent for model selection unless the predictor matrix satisfies a rather strong condition, and then proposes the adaptive Lasso that assigns different weights to penalize different coefficients in the $\ell_1$-penalty. Observing that the Lasso is designed for selecting individual regressors, Yuan and Lin (2006) extend the Lasso to group Lasso that selects "grouped variables". A combination of the adaptive Lasso and group Lasso yields the adaptive group Lasso that can achieve selection consistency for "grouped variables"; see, e.g., Wang and Leng (2008). In sum, such regular adaptive Lasso or adaptive group Lasso are designed to distinguish the nonzero coefficients from the zero coefficients asymptotically. They are not applicable here because our aim is not to select variables in $x_{it}$ but to determine the unknown number of breaks in $\{\beta_t\}$.

Tibshirani et al. (2005) propose the fused Lasso that is designed for problems with features that can be ordered in a meaningful way and penalizes the $\ell_1$-norm of both the coefficients and their successive differences. For a standard linear model: $y_i = \sum_{j=1}^{p} x_{ij}\gamma_j + \varepsilon_i$, $i = 1, ..., n$, the fused Lasso estimator of $\gamma = \left(\gamma_1, ..., \gamma_j\right)'$ is defined by

$$\hat{\gamma} = \arg\min_{\gamma} \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} x_{ij}\gamma_j \right)^2 + \lambda_n^{(1)} \sum_{j=1}^{p} \left|\gamma_j\right| + \lambda_n^{(2)} \sum_{j=2}^{p} \left|\gamma_{j-}\gamma_{j-1}\right|, \tag{2.4}$$

where $\lambda_n^{(1)}$ and $\lambda_n^{(2)}$ are two nonnegative tuning parameters, $\gamma_j$'s are scalar regression coefficients, and $x_{ij}$'s are regressors. Apparently, fused Lasso encourages sparsity of both the coefficients and their successive differences. Here, we can adopt the idea of fused Lasso because of the coefficient vectors $\{\beta_t\}$ in our model (2.1) have temporal order. The main difference of our PLS objective function in (2.2) from the standard Lasso objective function in (2.4) lies in three aspects: (1) we consider the vector difference $\beta_t - \beta_{t-1}$ by using the Frobenius norm $\|\cdot\|$ instead of the usual $\ell_1$-norm, (2) we assign different weights $\{\dot{w}_t\}$ to penalize different coefficient differences, and (3) we do not impose the $\ell_1$-penalty on the individual elements of $\beta_t$, $t = 1, ..., T$. Like Qian and Su (2015), the use of the Frobenius norm $\|\cdot\|$ for the vector difference $\beta_t - \beta_{t-1}$ generalizes the fused Lasso to the group fused Lasso. Unlike Qian and Su (2015) who do not assign different weights to the vector differences in their time series regression, our panel regression allows us to apply the adaptive weights $\{\dot{w}_t\}$, yielding the adaptive Lasso procedure. For this reason, we can call our estimation procedure as *adaptive group fused Lasso* (AGFL).

To obtain $\{\dot{w}_t\}$, we propose to obtain the preliminary estimate $\dot{\boldsymbol{\beta}} = (\dot{\beta}_1', ..., \dot{\beta}_T')'$ by minimizing the first term in the definition of $V_{1NT,\lambda_1}(\boldsymbol{\beta})$ in (2.2). We can readily demonstrate that

$$\dot{\boldsymbol{\beta}} = \dot{Q}_{NT}^{-1} \dot{R}_{NT}^{y}, \tag{2.5}$$

where $\dot{Q}_{NT}$ and $\dot{R}_{NT}^{y}$ are defined in (A.1) and (A.2) in Appendix A.1, respectively.

### 2.2.1 Post-Lasso least squares estimation

For any $\boldsymbol{\alpha}_m = \left(\alpha_1', ..., \alpha_{m+1}'\right)'$ and $\mathcal{T}_m = \{T_1, ..., T_m\}$ with $1 < T_1 < \cdots < T_m \leq T$, we define[4]

$$Q_{1NT}(\boldsymbol{\alpha}_m; \mathcal{T}_m) = \frac{1}{N} \sum_{j=1}^{m+1} \sum_{t=T_{j-1}+1}^{T_j-1} \sum_{i=1}^{N} \left(\Delta y_{it} - \alpha_j' \Delta x_{it}\right)^2 + \frac{1}{N} \sum_{j=1}^{m} \sum_{i=1}^{N} \left(\Delta y_{iT_j} - \alpha_{j+1}' x_{iT_j} + \alpha_j' x_{i,T_j-1}\right)^2, \tag{2.6}$$

where $\sum_{t=T_{j-1}+1}^{T_j-1} \sum_{i=1}^{N} \left(\Delta y_{it} - \alpha_j' \Delta x_{it}\right)^2$ corresponds to "the sum of squared errors" for observations in the $j$th artificial regime with time series observations indexed by integers in the interval $[T_{j-1}, T_j - 1]$, and $\sum_{i=1}^{N} \left(\Delta y_{it} - \alpha_{j+1}' x_{i,T_j} + \alpha_j' x_{i,T_j-1}\right)^2$ corresponds to the "the sum of squared errors" for observations when one moves from the $j$th regime to the $(j+1)$th regime. The second term in (2.6) is important and helps to improve the asymptotic efficiency when $T$ or the minimum length of the $m+1$ regimes is fixed. It can be omitted if $\min_{0 \leq j \leq m} |T_{j+1} - T_j| \to \infty$ as $T \to \infty$ and only the asymptotic efficiency is

[4] By default, the summation $\sum_{t=a}^{b}$ in this paper is zero if $b < a$.

concerned, but we still keep it to improve the finite sample performance of the post-Lasso estimate in this case. One can choose $\boldsymbol{\alpha}_m$ to minimize the objective function in (2.6). We denote the solution as $\tilde{\boldsymbol{\alpha}}_m^p(\mathcal{T}_m) = \left(\tilde{\alpha}_1^p(\mathcal{T}_m)', ..., \tilde{\alpha}_{m+1}^p(\mathcal{T}_m)'\right)'$. By setting $\mathcal{T}_m$ as $\tilde{\mathcal{T}}_{\tilde{m}}$, the set of estimated break dates via the AGFL procedure, we obtain the post-Lasso least squares estimator

$$\tilde{\boldsymbol{\alpha}}_{\tilde{m}}^p = \tilde{\boldsymbol{\alpha}}_{\tilde{m}}^p\left(\tilde{\mathcal{T}}_{\tilde{m}}\right) = \Phi_{NT}\left(\tilde{\mathcal{T}}_{\tilde{m}}\right)^{-1}\Psi_{NT}^y\left(\tilde{\mathcal{T}}_{\tilde{m}}\right), \tag{2.7}$$

where $\Phi_{NT}(\cdot)$ and $\Psi_{NT}^y(\cdot)$ are defined in (A.3) and (A.4) in Appendix A.1, respectively. We shall study the limiting distribution of $\tilde{\boldsymbol{\alpha}}_{\tilde{m}}^p$ in Section 3.3.

## 2.3 Penalized GMM (PGMM) estimation

In case (b), we propose to estimate $\boldsymbol{\beta}$ by minimizing the following PGMM objective function

$$V_{2NT,\lambda_2}(\boldsymbol{\beta}) = \sum_{t=2}^{T}\left\{\frac{1}{N}\sum_{i=1}^{N}\rho_{it}\left(\beta_t, \beta_{t-1}\right)\right\}' W_t\left\{\frac{1}{N}\sum_{i=1}^{N}\rho_{it}\left(\beta_t, \beta_{t-1}\right)\right\} + \lambda_2\sum_{t=2}^{T}\ddot{w}_t\left\|\beta_t - \beta_{t-1}\right\|, \tag{2.8}$$

where $\rho_{it}\left(\beta_t, \beta_{t-1}\right) = z_{it}(\Delta y_{it} - \beta_t' x_{it} + \beta_{t-1}' x_{i,t-1})$, $\lambda_2 = \lambda_2(N,T) \geq 0$ is a tuning parameter, $W_t = W_{tNT}$ is a $q \times q$ symmetric p.d. weight matrix for $t = 2, ..., T$, and $\ddot{w}_t$ is a data-driven weight defined by

$$\ddot{w}_t = \left\|\ddot{\beta}_t - \ddot{\beta}_{t-1}\right\|^{-\kappa_2}, \ t = 2, ..., T, \tag{2.9}$$

$\{\ddot{\beta}_t\}$ are preliminary estimates of $\{\beta_t\}$, and $\kappa_2$ is an user-specified positive constant that usually takes value 2 in the literature. Clearly, the first term in the definition of $V_{2NT,\lambda_2}(\boldsymbol{\beta})$ in (2.8) is different from the usual GMM objective function in the panel setting with time-invariant parameters where only one weight matrix ($W$, say) is needed and the double summation $\sum_{t=2}^{T}\sum_{i=1}^{N}$ occurs twice, one before the single weight matrix and the other after the single weight matrix. It is also different from the GMM-type objective function in Andrews (1993) who considers the test of a single structural change in a time series regression. Noting that the objective function in (2.8) is convex in $\boldsymbol{\beta}$, it is easy to obtain the solution $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1', ..., \hat{\beta}_T')'$, where we frequently suppress the dependence of $\hat{\beta}_t = \hat{\beta}_t(\lambda_2)$ on $\lambda_2$. We will propose a data-driven method to choose $\lambda_2$ in Section 4.4.

For a given solution $\{\hat{\beta}_t\}$, we can find the set of estimated break dates $\hat{\mathcal{T}}_{\hat{m}} = \{\hat{T}_1, ..., \hat{T}_{\hat{m}}\}$ as in Section 2.2. Like before, $\hat{\mathcal{T}}_{\hat{m}}$ divides $[1, T]$ into $\hat{m}+1$ regimes such that the parameter estimates remain constant within each regime and $\left\|\hat{\beta}_t - \hat{\beta}_{t-1}\right\| \neq 0$ whenever $t = \hat{T}_j$ for some $j = 1, ..., \hat{m}$. Let $\hat{T}_0 = 1$ and $\hat{T}_{\hat{m}+1} = T+1$. Define $\hat{\alpha}_j = \hat{\alpha}_j(\hat{\mathcal{T}}_{\hat{m}}) = \hat{\beta}_{\hat{T}_{j-1}}$ as the estimate of $\alpha_j$ for $j = 1, ..., \hat{m}+1$. Let $\hat{\boldsymbol{\alpha}}_{\hat{m}} = \hat{\boldsymbol{\alpha}}_{\hat{m}}(\hat{\mathcal{T}}_{\hat{m}}) = (\hat{\alpha}_1(\hat{\mathcal{T}}_{\hat{m}})', ..., \hat{\alpha}_{\hat{m}+1}(\hat{\mathcal{T}}_{\hat{m}})')'$.

To obtain the adaptive weights $\{\ddot{w}_t\}$, we propose to obtain the preliminary estimate $\ddot{\boldsymbol{\beta}} = (\ddot{\beta}_1', ..., \ddot{\beta}_T')'$ by minimizing the first term in the definition of $V_{2NT,\lambda_2}(\boldsymbol{\beta})$ in (2.8). It is easy to show that

$$\ddot{\boldsymbol{\beta}} = \ddot{Q}_{NT}^{-1}\ddot{R}_{NT}^y, \tag{2.10}$$

where $\ddot{Q}_{NT}$ and $\ddot{R}_{NT}^y$ are defined in (A.6) and (A.7) in Appendix A.2, respectively.

**Remark.** To proceed, it is worth mentioning that one might consider an alternative PGMM objective function

$$\bar{V}_{2NT,\lambda_2}(\boldsymbol{\beta}) = \left\{ \frac{1}{N} \sum_{t=2}^{T} \sum_{i=1}^{N} \rho_{it}(\beta_t, \beta_{t-1}) \right\}' W_{NT} \left\{ \frac{1}{N} \sum_{t=2}^{T} \sum_{i=1}^{N} \rho_{it}(\beta_t, \beta_{t-1}) \right\} + \lambda_2 \sum_{t=2}^{T} \ddot{w}_t \left\| \beta_t - \beta_{t-1} \right\|, \tag{2.11}$$

where $W_{NT}$ is a $q \times q$ symmetric matrix that is asymptotically nonsingular, and both $\sum_{t=2}^{T}$ and $\sum_{i=1}^{N}$ enter the first term in (2.11) twice and symmetrically. Nevertheless, this reformulation only makes sense for the over-identification case $(q > p)$. If the dimension of $z_{it}$ is the same as that of $x_{it}$, i.e., $q = p$, the resulting PGMM estimators of $\{\beta_t\}$ based on the minimization of (2.11) are given by

$$\hat{\beta}_{t,pgmm} = \left( \sum_{i=1}^{N} \sum_{s=2}^{T} z_{is} \Delta x_{is}' \right)^{-1} \sum_{i=1}^{N} \sum_{s=2}^{T} z_{is} \Delta y_{is} \text{ for each } t = 1, ..., T.$$

That is, the objective function in (2.11) always take the value zero, regardless of the choices of $W_{NT}$ and $\lambda_2$, and the PGMM estimators of $\{\beta_t\}$ remain as a constant no matter whether there is a beak in the data or not. So we cannot apply the above PGMM method to estimate the number of breaks at all. This motivates us to consider the PGMM objective function of the form given in (2.8).

### 2.3.1 Post-Lasso GMM estimation

For any $\boldsymbol{\alpha}_m = (\alpha_1', ..., \alpha_{m+1}')'$ and $\mathcal{T}_m = \{T_1, ..., T_m\}$ with $1 < T_1 < \cdots < T_m \leq T$, we define

$$\begin{aligned} Q_{2NT}(\boldsymbol{\alpha}_m; \mathcal{T}_m) = & \sum_{j=1}^{m+1} \left[ \frac{1}{N} \sum_{t=T_{j-1}+1}^{T_j - 1} \sum_{i=1}^{N} \rho_{it}(\alpha_j) \right]' W_j^p \left[ \frac{1}{N} \sum_{t=T_{j-1}+1}^{T_j - 1} \sum_{i=1}^{N} \rho_{it}(\alpha_j) \right] \\ & + \sum_{j=1}^{m} \left[ \frac{1}{N} \sum_{i=1}^{N} \rho_{1iT_j}(\alpha_{j+1}, \alpha_j) \right]' W_{T_j} \left[ \frac{1}{N} \sum_{i=1}^{N} \rho_{1iT_j}(\alpha_{j+1}, \alpha_j) \right], \end{aligned} \tag{2.12}$$

where $\rho_{it}(\alpha_j) = z_{it}(\Delta y_{it} - \alpha_j' \Delta x_{it})$, $\rho_{1iT_j}(\alpha_{j+1}, \alpha_j) = z_{iT_j}(\Delta y_{iT_j} - \alpha_{j+1}' x_{iT_j} + \alpha_j' x_{i,T_j - 1})$, and $W_j^p$ is a regime-specific $q \times q$ symmetric weight matrix that is p.d. in large samples. As in the case of PLS estimation, the second term in (2.12) is important when $T$ or the minimum regime length is fixed and can be omitted in the case where $\min_{0 \leq j \leq m} |T_{j+1} - T_j| \to \infty$ as $T \to \infty$. Let $\hat{\boldsymbol{\alpha}}_m^p(\mathcal{T}_m) = \left( \hat{\alpha}_1^p(\mathcal{T}_m)', ..., \hat{\alpha}_{m+1}^p(\mathcal{T}_m)' \right)'$ denote the minimizer of $Q_{2NT}$ defined in (2.12). By setting $\mathcal{T}_m$ as $\hat{\mathcal{T}}_{\hat{m}}$, the set of estimated break dates, we obtain the post-Lasso GMM estimator

$$\hat{\boldsymbol{\alpha}}_{\hat{m}}^p = \hat{\boldsymbol{\alpha}}_{\hat{m}}^p \left( \hat{\mathcal{T}}_{\hat{m}} \right) = \Upsilon_{NT}(\hat{\mathcal{T}}_{\hat{m}})^{-1} \Xi_{NT}^y \left( \hat{\mathcal{T}}_{\hat{m}} \right),$$

where $\Upsilon_{NT}(\cdot)$ and $\Xi_{NT}^y(\cdot)$ are defined in (A.8) and (A.9) in Appendix A.2, respectively. We shall study the limiting distribution of $\hat{\boldsymbol{\alpha}}_{\hat{m}}^p$ in Section 4.3.

To obtain the PGMM estimate and the associated post-Lasso estimate, one needs to choose the weight matrices $W_t$ $(t = 2, ..., T)$ and $W_j^p$ $(j = 1, ..., \hat{m} + 1)$. In the simulation and application below, we adopt a two-step strategy for determining both sets of weights. For $W_t$, we first obtain the estimate $\ddot{\beta}_t$ by choosing the $q \times q$ identity matrix $\mathbb{I}_q$ as the weight matrix. In the second step, we specify $W_t$ as the inverse of the estimated covariance matrix of $\rho_{it}(\ddot{\beta}_t, \ddot{\beta}_{t-1})$ and achieve an updated estimate of $\beta_t$. A similar procedure is adopted for determining the weights in post-Lasso estimation.

# 3 Asymptotic properties of the PLS estimators

In this section we address the asymptotic properties of the PLS estimators.

## 3.1 Basic assumptions

Let $I_j^0 = T_j^0 - T_{j-1}^0$ for $j = 1, ..., m^0 + 1$. Define

$$I_{\min} = \min_{1 \leq j \leq m^0+1} I_j^0, \; J_{\min} = \min_{1 \leq j \leq m^0} \left\| \alpha_{j+1}^0 - \alpha_j^0 \right\|, \text{ and } J_{\max} = \max_{1 \leq j \leq m^0} \left\| \alpha_{j+1}^0 - \alpha_j^0 \right\|.$$

Apparently, $I_{\min}$ denotes the minimum interval length among the $m^0 + 1$ regimes, and $J_{\min}$ and $J_{\max}$ denote the minimum and maximum jump sizes, respectively. In the case of fixed $T$, $I_{\min}$ does not pass to infinity as $N \to \infty$. When $T \to \infty$, $I_{\min}$ can either pass to infinity or stay fixed unless otherwise stated. We will maintain the assumption that $J_{\max}$ is always a fixed constant but $J_{\min}$ can be either fixed or shrinking to zero as either $N \to \infty$ or $(N, T) \to \infty$. Define the $p\left(m^0+1\right) \times p\left(m^0+1\right)$ matrix $\Phi_{NT}$ and $p\left(m^0+1\right) \times 1$ vector $\Phi_{NT}$ and $\Psi_{NT}^a$, respectively:

$$\Phi_{NT} = \Phi_{NT}\left(\mathcal{T}_{m^0}^0\right) \text{ and } \Psi_{NT}^a = \Psi_{NT}^a\left(\mathcal{T}_{m^0}^0\right) \text{ for } a = y \text{ or } u, \tag{3.1}$$

where $\Phi_{NT}(\cdot)$ and $\Psi_{NT}^a(\cdot)$ are defined in (A.3) and (A.4) in Appendix A.1, respectively.

To study the asymptotic properties of the PLS estimators, we make the following assumptions.

**Assumption A.1**. (i) Let $u_i = (u_{i1}, ..., u_{iT})'$. $\{x_i, u_i\}$ are independently distributed over $i$.

(ii) $E(x_{it}\Delta u_{it}) = 0$ and $E(x_{i,t-1}\Delta u_{it}) = 0$ for $i = 1, ..., N$ and $t = 2, ..., T$. $\max_{1\leq i\leq N}\max_{1\leq t\leq T} E\|\varsigma_{it}\|^{2\tau_0} < C < \infty$ for $\varsigma = x$ and $u$ and some $\tau_0 \geq 2$.

(iii) Let $\phi_{xx,t} = \frac{1}{N}\sum_{i=1}^N x_{it}x_{it}'$. There exist two constants $\underline{c}_{xx}$ and $\bar{c}_{xx}$ such that $0 < \underline{c}_{xx} \leq \min_{1\leq t\leq T} \mu_{\min}\left(E\left(\phi_{xx,t}\right)\right) \leq \max_{1\leq t\leq T}\mu_{\max}\left(E\left(\phi_{xx,t}\right)\right) \leq \bar{c}_{xx} < \infty$.

(iv) Let $\eta_{iT}$ denote the error term in the least squares projection of $x_{iT}$ on $x_{i,T-1}$. There exists a constant $\underline{c}_\eta > 0$ such that $\mu_{\min}(N^{-1}\sum_{i=1}^N E(\eta_{iT}\eta_{iT}')) \geq \underline{c}_\eta$.

**Assumption A.2**. (i) $J_{\max} = O(1)$ and $N^{1/2}J_{\min} \to c_J \in (0, \infty]$ as $(N, T) \to \infty$.

(ii) $\text{plim}_{(N,T)\to\infty} m^0 N^{1/2}\lambda_1 J_{\min}^{-\kappa_1} = c \in [0, \infty)$.

(iii) $\text{plim}_{(N,T)\to\infty} N^{(\kappa_1+1)/2}\lambda_1 = \infty$.

(iv) For some $\epsilon_0 > 1$, $N^{1-\tau_0}T(\ln T)^{\epsilon_0\tau_0} \to 0$ as $(N, T) \to \infty$.

**Assumption A.3**. Let $\mathbb{D}_{m^0+1} = \text{diag}(\sqrt{I_1^0}, ..., \sqrt{I_{m^0+1}^0}) \otimes \mathbb{I}_p$. Let $S$ denote an arbitrary $l \times p(m^0+1)$ selection matrix such that $\|S\|$ is finite, where $l \in [1, p(m^0+1)]$ is a fixed integer.

(i) There exists $\Phi_0 > 0$ such that $\left\|\mathbb{D}_{m^0+1}^{-1}\Phi_{NT}\mathbb{D}_{m^0+1}^{-1} - \Phi_0\right\|_{\text{sp}} = o_P(1)$.

(ii) $\sqrt{N}S\Phi_0^{-1}\mathbb{D}_{m^0+1}^{-1}\Psi_{NT}^u \xrightarrow{D} N\left(0, S\Phi_0^{-1}\Omega_0\Phi_0^{-1}S'\right)$.

Assumption A.1(i) requires that $\{x_i, u_i\}$ be independently distributed. It may be relaxed to allow for weak forms of cross section dependence at very lengthy arguments. A.1(ii) specifies moment conditions on $\{x_{it},\ u_{it}\}$. If $E\left(u_{it}|x_{it+1}, x_{it}\right) = 0$ a.s. for each $i$ and $t$, then the first part of A.1(ii) is satisfied. In conjunction with A.1(i), A.1(ii) implies that each block element of $\sqrt{N}\dot{R}_{NT}^u$ is $O_P(1)$ and $T^{-1}N\left\|\dot{R}_{NT}^u\right\|^2 = O_P(1)$ by Chebyshev inequality. A1(iii)-(iv) impose conditions on $E\left(\phi_{xx,t}\right)$ and $N^{-1}\sum_{i=1}^N E\left(\eta_{iT}\eta_{iT}'\right)$. They are used to ensure that the $Tp\times Tp$ matrix $\dot{Q}_{NT}$ is well behaved (see Lemma 3.1 below). Assumption A.2 mainly specifies conditions on $m^0$, $J_{\min}$, $\lambda_1$, and $N$. We use the probability limit instead of the usual limit in A.2(ii)-(iii) because we allow $\lambda_1$ to be data-driven and thus random. We allow the minimum break size $J_{\min}$ to shrink to zero as $N\to\infty$ but it cannot shrink to zero faster than $N^{-1/2}$. In addition, we allow the number of breaks $m^0$ to diverge to infinity at a slow rate. Assumption A.3 specifies conditions to ensure the asymptotic normality for any linear combinations of the Lasso or post Lasso estimators. If $m^0$ remains fixed as $(N,T)\to\infty$, we can simply replace the selection matrix $S$ by an identity matrix. From the definition of the SBTM $\Phi_{NT}(\cdot)$ in (A.3), we can easily see that the off-diagonal block matrices $\Phi_l^{\dagger}\left(\mathcal{T}_{m^0}^0\right)$, $l = 2, ..., m^0+1$, are not involved with any summation over the time index. This implies that after normalization, the probability limit of $\mathbb{D}_{m^0+1}^{-1}\Phi_{NT}\mathbb{D}_{m^0+1}^{-1}$ is given by a block diagonal matrix provided $I_{\min}\to\infty$ as $T\to\infty$.[5] That is, $\Phi_0$ is now block diagonal and one can readily check its non-singularity.

In the special case where $J_{\min}$ is bounded away from zero and $m^0$ remains fixed as $(N,T)\to\infty$, A.2 is simplified to

**Assumption A.2$^*$.** $\text{plim}_{(N,T)\to\infty}N^{1/2}\lambda_1 = c\in[0,\infty)$ and $\text{plim}_{(N,T)\to\infty}N^{(\kappa_1+1)/2}\lambda_1 = \infty$.

The following lemma studies the eigenvalue behavior of $\dot{Q}_{NT}$.

**Lemma 3.1** *Suppose Assumptions A.1 and A.2(iv) hold. Let $\dot{Q}_0 = E(\dot{Q}_{NT})$. Then*

*(i) There exist two constants $\underline{c}_{\dot{Q}_0}$ and $\bar{c}_{\dot{Q}_0}$ such that $0<\underline{c}_{\dot{Q}_0}\leq\mu_{\min}(\dot{Q}_0)\leq\mu_{\max}(\dot{Q}_0)\leq\bar{c}_{\dot{Q}_0}<\infty$,*

*(ii) $\left\|\dot{Q}_{NT}-\dot{Q}_0\right\|_{sp} = o_P(1)$,*

*(iii) $\frac{1}{2}\underline{c}_{\dot{Q}_0}\leq\mu_{\min}(\dot{Q}_{NT})\leq\mu_{\max}(\dot{Q}_{NT})\leq 2\bar{c}_{\dot{Q}_0}$ w.p.a.1.*

Lemma 3.1 indicates that despite the divergent dimensions of the $Tp\times Tp$ matrix $\dot{Q}_{NT}$ as $T\to\infty$, its eigenvalues are well behaved asymptotically. With the help of this lemma, we show in Lemma B.1 that $\sqrt{N}\left(\dot{\beta}_t-\beta_t^0\right) = O_P(1)$ for each $t = 1, ..., T$. Lemma 3.1 is also used in the proof of Theorem 3.2 below.

## 3.2 Consistency

The following theorem establishes the consistency of $\{\tilde{\beta}_t\}$.

[5] Intuitively, this means that the second term in (2.6) does not contribute to the limiting distribution of the post-Lasso estimator when $I_{\min}\to\infty$ as $T\to\infty$.

**Theorem 3.2** *Suppose that Assumptions A.1 and A.2(iv) hold. Then (i)* $T^{-1}\left\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|^2=O_P\left(N^{-1}\right)$, *and (ii)* $\tilde{\beta}_t-\beta_t^0=O_P\left(N^{-1/2}\right)$ *for each* $t=1,...,T$.

Theorems 3.2(i) and (ii) establish the mean square and pointwise convergence rates of $\{\tilde{\beta}_t\}$, respectively. The two results are equivalent in the case of fixed $T$. We allow $(N,T)\to\infty$ and then the proof of Theorem 3.2(ii) demands some extra effort. In particular, we need a close examination of the factorization and inversion properties of a SBTM.

Let $\mathcal{T}_{m^0}^{0c}=\{2,...,T\}\backslash\mathcal{T}_{m^0}^0$. Let $\theta_1^0=\beta_1^0$ and $\theta_t^0=\beta_t^0-\beta_{t-1}^0$ for $t=2,...,T$. Let $\tilde{\theta}_1=\tilde{\beta}_1$ and $\tilde{\theta}_t=\tilde{\beta}_t-\tilde{\beta}_{t-1}$ for $t=2,...,T$. The following theorem establishes the selection consistency.

**Theorem 3.3** *Suppose that Assumptions A.1-A.2 hold. Then* $P\left(\left\|\tilde{\theta}_t\right\|=0 \text{ for all } t\in\mathcal{T}_{m^0}^{0c}\right)\to 1$ *as* $N\to\infty$.

Theorem 3.2 says that w.p.a.1 all the zero vectors in $\{\theta_t^0,\ 2\leq t\leq T\}$ must be estimated as exactly zero by the PLS method so that the number of estimated breaks $\tilde{m}$ cannot be larger than $m^0$ when $N$ is sufficiently large. On the other hand, by Theorem 3.2(ii), we know that the estimates of the nonzero vectors in $\{\theta_t^0,\ 2\leq t\leq T\}$ must be consistent by noting that $\tilde{\beta}_t-\tilde{\beta}_{t-1}$ consistently estimates $\theta_t^0$ for $t\geq 2$. Put together, Theorems 3.2 and 3.3 imply that the AGFL has the ability to identify the true regression model with the correct number of breaks consistently when the minimum break size $J_{\min}$ does not shrink to zero too fast.

**Corollary 3.4** *Suppose that Assumptions A.1-A.2 hold with* $c_J=\infty$ *in Assumption A.2(i). Then (i)* $\lim_{N\to\infty}P\left(\tilde{m}=m^0\right)=1$, *and (ii)* $\lim_{N\to\infty}P(\tilde{T}_1=T_1^0,...,\tilde{T}_{m^0}=T_{m^0}^0\mid\tilde{m}=m^0)=1$.

The above corollary implies that, as long as $J_{\min}$ remains fixed or shrinks to zero at a rate slower than $N^{-1/2}$ as $N\to\infty$, we can estimate the number of structural changes and all the break dates consistently. In contrast, Qian and Su (2015, Theorem 3.3) only establish the claim that the group fused Lasso procedure can not under-estimate the number of breaks in a time series regression and that all the break fractions (but not the break dates) can be consistently estimated as in Bai and Perron (1998). More precisely, letting $\mathcal{D}(A,B)\equiv\sup_{b\in B}\inf_{a\in A}|a-b|$ for any two sets $A$ and $B$, Qian and Su (2015, Theorem 3.2) establish the claim that $\lim_{T\to\infty}P\left(\mathcal{D}\left(\tilde{\mathcal{T}}_{\tilde{m}},\mathcal{T}_{m^0}^0\right)\leq T\delta_T\right)=1$ for some sequence $\{\delta_T\}$ such that $\delta_T\to 0$ and $T\delta_T\to\infty$ as $T\to\infty$. In our panel setting, the availability of $N$ cross sectional units for each time period permits us to obtain the set of consistent preliminary estimates $\{\dot{\beta}_t\}$ used to construct the adaptive weights $\{\dot{w}_t\}$. The adaptive nature of our procedure helps to identify the exact set of break dates and yield stronger results than those in Qian and Su (2015).

## 3.3 Limiting distributions of the Lasso and post-Lasso estimators

In this subsection we study the asymptotic distributions of the Lasso and post-Lasso estimators.

Let $\mathcal{A}_{NT}=\{\tilde{T}_j=T_j^0 \text{ for } j=1,...,m^0\}$ and $\mathcal{A}_{NT}^c$ its complement. By Corollary 3.4, we have

$$\begin{aligned} & P\left\{\sqrt{N}S\mathbb{D}_{m^0+1}(\tilde{\boldsymbol{\alpha}}_{\tilde{m}}^p(\tilde{\mathcal{T}}_{\tilde{m}})-\boldsymbol{\alpha}^0)\in\mathcal{C}\mid\tilde{m}=m^0\right\} \\ = & P\left\{\sqrt{N}S\mathbb{D}_{m^0+1}(\tilde{\boldsymbol{\alpha}}_{\tilde{m}}^p(\tilde{\mathcal{T}}_{\tilde{m}})-\boldsymbol{\alpha}^0)\in\mathcal{C},\ \mathcal{A}_{NT}\mid\tilde{m}=m^0\right\} \\ & +P\left\{\sqrt{N}S\mathbb{D}_{m^0+1}(\tilde{\boldsymbol{\alpha}}_{\tilde{m}}^p(\tilde{\mathcal{T}}_{\tilde{m}})-\boldsymbol{\alpha}^0)\in\mathcal{C},\ \mathcal{A}_{NT}^c\mid\tilde{m}=m^0\right\} \\ = & P\left\{\sqrt{N}S\mathbb{D}_{m^0+1}(\tilde{\boldsymbol{\alpha}}_{m^0}^p(\mathcal{T}_{m^0}^0)-\boldsymbol{\alpha}^0)\in\mathcal{C}\right\}+o_P(1), \end{aligned}$$

where $\mathcal{C}\subset\mathbb{R}^l$, and $\tilde{\boldsymbol{\alpha}}_{m^0}^p(\mathcal{T}_{m^0}^0)$ is the infeasible estimator of $\boldsymbol{\alpha}^0$ which is obtained if one knows the exact set $\mathcal{T}_{m^0}^0$ of true break dates:

$$\tilde{\boldsymbol{\alpha}}_{m^0}^p(\mathcal{T}_{m^0}^0)=\Phi_{NT}^{-1}\Psi_{NT}^y, \tag{3.2}$$

where $\Phi_{NT}$ and $\Psi_{NT}^y$ are defined in (3.1).

The following theorem reports the limiting distributions of the Lasso estimator $\tilde{\boldsymbol{\alpha}}_{\tilde{m}}(\tilde{\mathcal{T}}_{\tilde{m}})$ and the post-Lasso estimator $\tilde{\boldsymbol{\alpha}}_{\tilde{m}}^p(\tilde{\mathcal{T}}_{\tilde{m}})$ conditional on the large probability event $\{\tilde{m}=m^0\}$.

**Theorem 3.5** *Suppose that Assumptions A.1-A.3 hold with $c_J=\infty$ in Assumption A.2(i). Then conditional on $\tilde{m}=m^0$, we have (i) $\sqrt{N}S\mathbb{D}_{m^0+1}(\tilde{\boldsymbol{\alpha}}_{\tilde{m}}(\tilde{\mathcal{T}}_{\tilde{m}})-\boldsymbol{\alpha}^0)\xrightarrow{D}N\left(0,S\Phi_0^{-1}\Omega_0\Phi_0^{-1}S'\right)$, and (ii) $\sqrt{N}S\mathbb{D}_{m^0+1}(\tilde{\boldsymbol{\alpha}}_{\tilde{m}}^p(\tilde{\mathcal{T}}_{\tilde{m}})-\boldsymbol{\alpha}^0)\xrightarrow{D}N\left(0,S\Phi_0^{-1}\Omega_0\Phi_0^{-1}S'\right)$.*

Noting that the dimensions of $\tilde{\boldsymbol{\alpha}}_{\tilde{m}}(\tilde{\mathcal{T}}_{\tilde{m}})$ and $\tilde{\boldsymbol{\alpha}}_{\tilde{m}}^p(\tilde{\mathcal{T}}_{\tilde{m}})$ diverge to infinity when $m^0\to\infty$, we cannot derive their asymptotic normality directly. For this reason, we follow the literature on inferences with a diverging number of parameters (e.g., Fan and Peng (2004), Lam and Fan (2008), Lu and Su (2015a)) and prove the asymptotic normality for any arbitrary linear combinations of elements of $\tilde{\boldsymbol{\alpha}}_{\tilde{m}}(\tilde{\mathcal{T}}_{\tilde{m}})$ or $\tilde{\boldsymbol{\alpha}}_{\tilde{m}}^p(\tilde{\mathcal{T}}_{\tilde{m}})$. Since we allow $I_j^0$ to be either fixed or diverge to infinity as $T\to\infty$, $\tilde{\alpha}_j(\tilde{\mathcal{T}}_{\tilde{m}})$ and $\tilde{\alpha}_j^p(\tilde{\mathcal{T}}_{\tilde{m}})$'s may have different convergence rates to their true values. In the special case where $I_j^0$ is proportional to $T$, both achieve the usual $\sqrt{NT}$-rate of consistency.

Theorem 3.5 indicate that both the Lasso estimator $\tilde{\boldsymbol{\alpha}}_{\tilde{m}}(\tilde{\mathcal{T}}_{\tilde{m}})$ and the post-Lasso version $\tilde{\boldsymbol{\alpha}}_{\tilde{m}}^p(\tilde{\mathcal{T}}_{\tilde{m}})$ are asymptotically equivalent to the infeasible estimator $\tilde{\boldsymbol{\alpha}}_{m^0}^p(\mathcal{T}_{m^0}^0)$ conditional on the large probability event $\tilde{m}=m^0$. The latter can be obtained only if one knows all break dates. In this sense, our Lasso and post-Lasso estimators have the oracle efficiency. Despite the asymptotic equivalence between the Lasso and post-Lasso estimators, it is well known that the post-Lasso estimator typically outperforms the Lasso estimator and is thus recommended for practical use.

### 3.4 Choosing the tuning parameter $\lambda_1$

Let $\tilde{\boldsymbol{\alpha}}_{\tilde{m}_{\lambda_1}}\equiv\tilde{\boldsymbol{\alpha}}_{\tilde{m}_{\lambda_1}}(\tilde{\mathcal{T}}_{\tilde{m}_{\lambda_1}})=(\tilde{\alpha}_1(\tilde{\mathcal{T}}_{\tilde{m}_{\lambda_1}})',...,\hat{\alpha}_{\tilde{m}_{\lambda_1}+1}(\tilde{\mathcal{T}}_{\tilde{m}_{\lambda_1}})')'$ denote the set of post-Lasso estimates of the regression coefficients based on the break dates in $\tilde{\mathcal{T}}_{\tilde{m}_{\lambda_1}}=\tilde{\mathcal{T}}_{\tilde{m}_{\lambda_1}}(\lambda_1)$, where we make the dependence of various estimates on $\lambda_1$ explicit. Let $\tilde{\sigma}^2_{\tilde{\mathcal{T}}_{\tilde{m}_{\lambda_1}}}\equiv\frac{1}{T-1}Q_{1NT}(\tilde{\boldsymbol{\alpha}}_{\tilde{m}_{\lambda_1}};\tilde{\mathcal{T}}_{\tilde{m}_{\lambda_1}})$. Following Wang et al. (2007), Zhang et al. (2010) and Su and Qian (2014), we propose to select the tuning parameter $\lambda_1$ by minimizing the following information criterion (IC):

$$IC_1(\lambda_1) = \tilde{\sigma}^2_{\tilde{\mathcal{T}}_{\tilde{m}_{\lambda_1}}} + \rho_{1NT}\, p\left(\tilde{m}_{\lambda_1} + 1\right), \tag{3.3}$$

where $\rho_{1NT}$ is a tuning parameter which plays a similar role to that of $\frac{2}{NT}$ and $\frac{\ln(NT)}{NT}$ in Akaike and Bayesian information criteria, respectively.

Denote $\Omega = [0, \mu_{\max}]$, a bounded interval in $\mathbb{R}^+$. We divide $\Omega$ into three random subsets $\Omega_0$, $\Omega_-$ and $\Omega_+$ as follows

$$\Omega_0 = \left\{\lambda_1 \in \Omega : \tilde{m}_{\lambda_1} = m^0\right\},\ \Omega_- = \left\{\lambda_1 \in \Omega : \tilde{m}_{\lambda_1} < m^0\right\},\ \text{and } \Omega_+ = \left\{\lambda_1 \in \Omega : \tilde{m}_{\lambda_1} > m^0\right\}.$$

Clearly, $\Omega_0$, $\Omega_-$ and $\Omega_+$ denote the three subsets of $\Omega$ in which the correct-, under- and over-number of breaks are selected by our AGFL procedure, respectively. We suppress their dependence on the sample sizes $N$ and $T$ for notational simplicity. They are random because $\tilde{m}_{\lambda_1}$ has to be determined based on the random sample. Let $\lambda^0_{1NT}$ denote an element in $\Omega_0$ that satisfies the conditions on $\lambda_1$ in Assumptions A.2(ii)-(iii).

Let $\bar{\sigma}^2_{NT} \equiv \frac{1}{N(T-1)} \sum_{i=1}^{N} \sum_{t=2}^{T} \left(\Delta u_{it}\right)^2$ and $\sigma^2_0 \equiv \text{plim}\bar{\sigma}^2_{NT}$. To state the next result, we add the following assumptions.

**Assumption A.4**. (i) $\text{plim}_{N\to\infty} \min_{1\leq j\leq m^0} \min_{\alpha\in\mathbb{R}^p} \frac{1}{NJ^2_{\min}} \sum_{i=1}^{N} [(\alpha^0_{j+1} - \alpha)' x_{iT^0_j} - (\alpha^0_j - \alpha)' x_{i,T^0_j-1}]^2 \geq \underline{c}_\alpha > 0$.

(ii) $\frac{1}{\sqrt{N(T-1)}} \sum_{t=2}^{T} \sum_{i=1}^{N} \Delta x_{it} \Delta u_{it} = O_P(1)$.

(iii) As $(N,T) \to \infty$, $\frac{T}{(I_{\min} J_{\min})^2 N} \to 0$.

**Assumption A.5**. As $(N,T) \to \infty$, $\left(m^0 + \frac{T}{I_{\min} J^2_{\min}}\right) \rho_{1NT} \to 0$ and $N\rho_{1NT} \to \infty$.

A.4(i) imposes conditions on the parameters and the observations that are either at the break dates or immediately preceding the break dates. The scalar $J^2_{\min}$ reflects the fact that we allow the minimum break size $J_{\min}$ to shrink to zero. In the latter case, pulling observations in two adjacent regimes with the break size of order $O(J_{\min})$ together to estimate the regression coefficients within these two regimes is still consistent with $J^{-1}_{\min}$-rate of consistency. Under A.2(i)-(ii), A.4(ii) can be verified under various weak dependence conditions, say, strong mixing or martingale difference sequence-type of conditions. A.4(iii) imposes restriction on $I_{\min}$, $J_{\min}$ and the sample sizes. It is trivially satisfied if $I_{\min} \propto T$ and $J_{\min}$ remains fixed as $N \to \infty$ or $(N,T) \to \infty$, and reduces to the condition that $c_J = \infty$ in Assumption A.2(i) in the case where $T$ is fixed. A.5 reflects the usual conditions for the consistency of model selection, that is, the penalty coefficient $\rho_{1NT}$ cannot shrink to zero either too fast or too slowly. If $I_{\min} \propto T$ and $J^{-1}_{\min} = O(1)$, the first part of A.5 requires that $\rho_{1NT} \to 0$, which is standard for a typical IC function. The second condition in A.5 is different from the typical IC requirement that $NT\rho_{1NT} \to \infty$ in the model selection literature because it is possible for a regime in a over-parametrized model to have only one time series observation, and $N^{-1}$ indicates the probability order of the distance between the first term in our IC function for an over-parametrized model and that for the true model.

**Theorem 3.6** *Suppose that Assumptions A.1, A.2(i) and A.3-A.5 hold with $c_J = \infty$ in Assumption A.2(i). Then $P\left(\inf_{\lambda_1 \in \Omega_- \cup \Omega_+} IC_1(\lambda_1) > IC_1\left(\lambda_{1NT}^0\right)\right) \to 1$ as $N \to \infty$.*

Theorem 3.6 implies that the $\lambda_1$'s that yield the over-estimated or under-estimated number of breaks fail to minimize the information criterion w.p.a.1. Consequently, the minimizer of $IC_1(\lambda_1)$ can only be the one that produces the correct number of estimated breaks in large samples. Note that we prove the above theorem without requiring $\lambda_1$ to satisfy Assumptions A.2(ii)-(iii). It indicates that if the number of corrected breaks is of our major concern, we can simply choose $\lambda_1$ to minimize $IC_1(\lambda_1)$.

# 4 Asymptotic properties of the PGMM estimators

In this section we address the statistical properties of the PGMM estimators.

## 4.1 Assumptions

Define the $p\left(m^0+1\right) \times p\left(m^0+1\right)$ matrix $\Upsilon_{NT}$ and $p\left(m^0+1\right) \times 1$ vector $\Xi_{NT}^a$, respectively:

$$\Upsilon_{NT} = \Upsilon_{NT}\left(\mathcal{T}_{m^0}^0\right) \text{ and } \Xi_{NT}^a = \Xi_{NT}^a\left(\mathcal{T}_{m^0}^0\right) \text{ for } a = y \text{ or } u, \qquad (4.1)$$

where $\Upsilon_{NT}(\cdot)$ and $\Xi_{NT}^a(\cdot)$ are defined in (A.8) and (A.9) in Appendix A.2, respectively. To study the asymptotic properties of the PGMM estimators, we make the following assumptions.

**Assumption B.1**. (i) Let $z_i = (z_{i2}, ..., z_{iT})'$. $\{x_i, z_i, u_i\}$ are independently distributed over $i$.

(ii) $E(z_{it}\Delta u_{it}) = 0$ for $i = 1, ..., N$ and $t = 2, ..., T$. $\max_{1\le i\le N}\max_{1\le t\le T} E\|\varsigma_{it}\|^{2\tau_0} < C < \infty$ for $\varsigma_{it} = x_{it}$, $z_{it}$ and $u_{it}$ and some $\tau_0 > 2$.

(iii) Let $\phi_{zx,t,s} = \frac{1}{N}\sum_{i=1}^N z_{it}x_{is}'$, $A_{t,s} = E\left(\phi_{zx,t,s}'\right)E\left(\phi_{zx,t,s}\right)$, and $A_t = A_{t,t}$. There exist two constants $\underline{c}_w$ and $\bar{c}_w$ such that $0 < \underline{c}_w \le \min_{1\le t\le T}\mu_{\min}(W_t) \le \max_{1\le t\le T}\mu_{\max}(W_t) \le \bar{c}_w < \infty$. There exist two constants $\underline{c}_{zx}$ and $\bar{c}_{zx}$ such that $0 < \underline{c}_{zx} \le \min_{2\le t\le T}\mu_{\min}(A_{t,t-1}) \le \max_{2\le t\le T}\mu_{\max}(A_{t,t-1}) \le \bar{c}_{zx} < \infty$, and $0 < \underline{c}_{zx} \le \min_{1\le t\le T}\mu_{\min}(A_t) \le \max_{1\le t\le T}\mu_{\max}(A_t) \le \bar{c}_{zx} < \infty$.

(iv) Let $\hat{\eta}_{iT}$ denote the the residual from the auxiliary GMM estimation of $x_{iT} = \alpha_t x_{i,T-1} + \eta_{iT}$ with $z_{iT}$ as the IV for $x_{i,T-1}$ and $W_T$ as the weight function. Let $\phi_{z\hat{\eta},t,t} = \frac{1}{N}\sum_{i=1}^N z_{it}\hat{\eta}_{it}'$. There exists a constant $\underline{c}_{z\eta} > 0$ such that $\text{plim}_{(N,T)\to\infty}\mu_{\min}(\phi_{z\hat{\eta},T,T}'\phi_{z\hat{\eta},T,T}) \ge \underline{c}_{z\eta}$.

**Assumption B.2**. (i) $J_{\max} = O(1)$ and $N^{1/2}J_{\min} \to c_J \in (0, \infty]$ as $(N,T) \to \infty$.

(ii) $\text{plim}_{(N,T)\to\infty} m^0 N^{1/2}\lambda_2 J_{\min}^{-\kappa_2} = c \in [0, \infty)$.

(iii) $\text{plim}_{(N,T)\to\infty} N^{(\kappa_2+1)/2}\lambda_2 = \infty$.

(iv) For some $\epsilon_0 > 1$, $N^{1-\tau_0}T(\ln T)^{\epsilon_0\tau_0} \to 0$ as $(N,T) \to \infty$.

**Assumption B.3**. (i) $\left\|\mathbb{D}_{m^0+1}^{-3}\Upsilon_{NT}\mathbb{D}_{m^0+1}^{-1} - \Upsilon_0\right\|_{\text{sp}} = o_P(1)$.

(ii) $\sqrt{N}S\Upsilon_0^{-1}\mathbb{D}_{m^0+1}^{-3}\Xi_{NT}^u \xrightarrow{D} N\left(0, S\Upsilon_0^{-1}\Sigma_0\Upsilon_0^{-1}S'\right)$ where $S$ is as defined in Assumption A.3

Assumptions B.1-B.3 parallel Assumptions A.1-A.3. B.1(ii) specifies moment conditions on $\{x_{it}$, $z_{it}$, $u_{it}\}$. In conjunction with B.1(i), B.1(ii) implies that each block element of $\sqrt{N}\ddot{R}_{NT}^u$ is $O_P(1)$ and

$T^{-1}N\left\|\dot{R}^{u}_{NT}\right\|^{2}=O_{P}\left(1\right)$ by Chebyshev inequality. B.1(iii) requires that $W_{t}$, $E\left(\phi'_{zx,t,t-1}\right)E\left(\phi_{zx,t,t-1}\right)$, and $E\left(\phi'_{zx,t}\right)E\left(\phi_{zx,t}\right)$ be nonsingular uniformly in $t$. B.1(v) requires that $\phi'_{z\hat{\eta},T,T}\phi_{z\hat{\eta},T,T}$ be asymptotically nonsingular. In conjunction with Assumption B.2(iv), B.1 implies that both $\phi'_{zx,t,t-1}W_{t}\phi_{zx,t,t-1}$ and $\phi'_{zx,t}W_{t}\phi_{zx,t}$ have eigenvalues that are bounded away from zero and infinity w.p.a.1. Assumption B.2 mainly specifies conditions on $m^{0}$, $J_{\min}$, $\lambda_{2}$, and $N$. Assumption B.3 specifies conditions to ensure the asymptotic normality of the post Lasso estimator. Note that the normalizations in B.3 is different from those in A.3 because the dominant terms in the definitions of $\Upsilon_{NT}$ and $\Xi^{u}_{NT}$ are now involved with two summations over the time index instead of one. From the definition of the SBTM $\Upsilon_{NT}\left(\cdot\right)$ in (A.8), we can easily see that the off-diagonal block matrices $\Upsilon^{\dagger}_{l}\left(\mathcal{T}^{0}_{m^{0}}\right)$, $l=2,...,m^{0}+1$, are not involved with any summation over the time index. This implies that after normalization, the probability limit of $\mathbb{D}^{-3}_{m^{0}+1}\Upsilon_{NT}\mathbb{D}^{-1}_{m^{0}+1}$ is given by a block diagonal matrix provided $I_{\min}\rightarrow\infty$ as $T\rightarrow\infty$.[6] That is, $\Upsilon_{0}$ is now block diagonal and one can readily check its non-singularity.

In the special case where $J_{\min}$ is bounded away from zero and $m^{0}$ remains fixed as $T\rightarrow\infty$, B.2 reduces to

**Assumption B.2***. $\text{plim}_{(N,T)\rightarrow\infty}N^{1/2}\lambda_{2}=c\in[0,\infty)$ and $\text{plim}_{(N,T)\rightarrow\infty}N^{(\kappa_{2}+1)/2}\lambda_{2}=\infty$.

The following lemma studies the eigenvalue behavior of $\ddot{Q}_{NT}$.

**Lemma 4.1** *Suppose Assumptions B.1 and B.2(iv) hold. Then w.p.a.1 the eigenvalues of $\ddot{Q}_{NT}$ are bounded away from zero and infinity, i.e., there exist two constants $\underline{c}_{\ddot{Q}}$ and $\bar{c}_{\ddot{Q}}$ such that $0<\underline{c}_{\ddot{Q}}\leq\mu_{\min}(\ddot{Q}_{NT})\leq\mu_{\max}(\ddot{Q}_{NT})\leq\bar{c}_{\ddot{Q}}<\infty$.*

Lemma 4.1 indicates that despite the divergent dimensions of the $Tp\times Tp$ matrix $\ddot{Q}_{NT}$ as $T\rightarrow\infty$, its eigenvalues are well behaved. This lemma is used to prove $\sqrt{N}\left(\ddot{\beta}_{t}-\beta^{0}_{t}\right)=O_{P}\left(1\right)$ for each $t=1,...,T$ and Theorem 4.2 below.

## 4.2 Consistency

The following theorem establishes the consistency of $\{\hat{\beta}_{t}\}$.

**Theorem 4.2** *Suppose that Assumptions B.1 and B.2(iv) hold. Then (i) $T^{-1}\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^{0}\right\|^{2}=O_{P}\left(N^{-1}\right)$, and (ii) $\hat{\beta}_{t}-\beta^{0}_{t}=O_{P}\left(N^{-1/2}\right)$ for each $t=1,...,T$.*

Theorems 4.2(i) and (ii) establish the mean square and pointwise convergence rates of $\{\hat{\beta}_{t}\}$, respectively. The two results are equivalent in the case of fixed $T$ and are not in the case of large $T$. If $(N,T)\rightarrow\infty$, the proof of Theorem 4.2(ii) requires the use of the factorization and inversion properties of a SBTM as in the proof of Theorem 3.2(ii).

Let $\hat{\theta}_{1}=\hat{\beta}_{1}$ and $\hat{\theta}_{t}=\hat{\beta}_{t}-\hat{\beta}_{t-1}$ for $t=2,...,T$. The following theorem establishes the selection consistency.

[6] Intuitively, this means that the second term in (2.12) does not contribute to the limiting distribution of the post-Lasso GMM estimator when $I_{\min}\rightarrow\infty$ as $T\rightarrow\infty$.

**Theorem 4.3** *Suppose that Assumptions B.1-B.2 hold. Then* $P\left(\left\|\hat{\theta}_t\right\| = 0 \text{ for all } t \in \mathcal{T}_{m^0}^{0c}\right) \to 1$ *as* $N \to \infty$.

Theorem 4.3 says that w.p.a.1 all the zero vectors in $\{\theta_t^0, \, 2 \le t \le T\}$ must be estimated as exactly zero by the PGMM method. On the other hand, by Theorem 4.2(ii), we know that the estimates of the nonzero vectors in $\{\theta_t^0, \, 2 \le t \le T\}$ must be consistent by noting that $\hat{\beta}_t - \hat{\beta}_{t-1}$ consistently estimates $\theta_t^0$ for $t \ge 2$. Put together, Theorems 4.2 and 4.3 imply that the AGFL has the ability to identify the true regression model with the correct number of breaks consistently when the minimum break size $J_{\min}$ does not shrink to zero too fast.

**Corollary 4.4** *Suppose that Assumptions B.1-B.2 hold with* $c_J = \infty$ *in Assumption B.2(i). Then (i)* $\lim_{N\to\infty} P\left(\hat{m} = m^0\right) = 1$ *and (ii)* $\lim_{N\to\infty} P(\hat{T}_1 = T_1^0, ..., \hat{T}_{m^0} = T_{m^0}^0 \mid \hat{m} = m^0) = 1$.

The above corollary implies that the PGMM method helps us to estimate the number of structural changes and all the break dates consistently.

## 4.3 Limiting distribution of the post-Lasso estimator

In this subsection we study the limiting distribution of the post-Lasso estimator $\hat{\boldsymbol{\alpha}}_{\hat{m}}^p(\hat{\mathcal{T}}_{\hat{m}})$. The study of the asymptotic distribution of the Lasso estimator $\hat{\boldsymbol{\alpha}}_{\hat{m}}(\hat{\mathcal{T}}_{\hat{m}})$ needs the introduction of a different set of notations and conditions. Because of the special feature of the first term in (2.8), the PGMM-based Lasso estimator is generally not as asymptotically efficient as the post-Lasso estimator.[7] To save space, we relegate the study of its limiting distribution to the supplementary Appendix E.

Using arguments as used in Section 3.3, we can argue that the post-Lasso GMM estimator $\hat{\boldsymbol{\alpha}}_{\hat{m}}^p(\hat{\mathcal{T}}_{\hat{m}})$ is asymptotically equivalent to the infeasible estimator $\hat{\boldsymbol{\alpha}}_{m^0}^p(\mathcal{T}_{m^0}^0)$ which is obtained if one knows the exact set $\mathcal{T}_{m^0}^0$ of true break dates:

$$\hat{\boldsymbol{\alpha}}_{m^0}(\mathcal{T}_{m^0}^0) = \Upsilon_{NT}^{-1}\Xi_{NT}^y,$$

where $\Upsilon_{NT}$ and $\Xi_{NT}^y$ are defined in (4.1). The following theorem reports the limiting distribution of $\hat{\boldsymbol{\alpha}}_{\hat{m}}^p(\hat{\mathcal{T}}_{\hat{m}})$ conditional on the large probability event $\left\{\hat{m} = m^0\right\}$.

**Theorem 4.5** *Suppose that Assumptions B.1-B.3 hold. Then conditional on* $\hat{m} = m^0$, *we have* $\sqrt{N}S\mathbb{D}_{m^0+1}$ $(\hat{\boldsymbol{\alpha}}_{\hat{m}}^p(\hat{\mathcal{T}}_{\hat{m}}) - \boldsymbol{\alpha}^0) \xrightarrow{D} N\left(0, S\Upsilon_0^{-1}\Sigma_0\Upsilon_0^{-1}S'\right)$.

Since we allow $I_j^0$ to be either fixed or diverge to infinity in the case of large $T$, $\hat{\alpha}_j^p(\hat{\mathcal{T}}_{\hat{m}})$'s may have different convergence rates to their true values. In the special case where $I_j^0$ is proportional to $T$, $\hat{\alpha}_j^p(\hat{\mathcal{T}}_{\hat{m}})$ achieves the usual $\sqrt{NT}$-rate of consistency.

[7] Notice that the derivative of (2.8) with respect to (wrt) $\beta_t$ does not involve with any summation over $t$ at all. To derive the limiting distribution of the Lasso estimator $\hat{\boldsymbol{\alpha}}_{\hat{m}}(\hat{\mathcal{T}}_{\hat{m}})$, we need to sum both sides of the first order conditions (FOCs) wrt $\beta_t$ over $t$ for each of the $\hat{m}+1$ estimated regimes and apply the fact that $\hat{\beta}_t = \hat{\alpha}_j$ if $t$ belongs to the $j$th estimated regime. But this device cannot generate the type of FOCs that are used to obtain the post-Lasso GMM estimator in view of the fact that $\sum_{t=\hat{T}_{j-1}+1}^{\hat{T}_j-1}$, like $\sum_{i=1}^N$, appears in the first term of the post-Lasso GMM objective function in (2.12) twice.

## 4.4 Choosing the tuning parameter $\lambda_2$

Let $\hat{\boldsymbol{\alpha}}_{\hat{m}_{\lambda_2}} = \hat{\boldsymbol{\alpha}}_{\hat{m}_{\lambda_2}}(\hat{\mathcal{T}}_{\hat{m}_{\lambda_2}}) = (\hat{\alpha}_1(\hat{\mathcal{T}}_{\hat{m}_{\lambda_2}})', ..., \hat{\alpha}_{\hat{m}_{\lambda_2}+1}(\hat{\mathcal{T}}_{\hat{m}_{\lambda_2}})')'$ denote the set of post-Lasso estimates of the regression coefficients based on the break dates in $\hat{\mathcal{T}}_{\hat{m}_{\lambda_2}} = \hat{\mathcal{T}}_{\hat{m}_{\lambda_2}}(\lambda_2)$, where we make the dependence of various estimates on $\lambda_2$ explicit. Let $\hat{\sigma}^2_{\hat{\mathcal{T}}_{\hat{m}_{\lambda_2}}} \equiv \frac{1}{T-1} Q_{2NT}(\hat{\boldsymbol{\alpha}}_{\hat{m}_{\lambda_2}}, \hat{\mathcal{T}}_{\hat{m}_{\lambda_2}})$. We propose to select the tuning parameter $\lambda_2$ by minimizing the following information criterion:

$$IC_2(\lambda_2) = \hat{\sigma}^2_{\hat{\mathcal{T}}_{\hat{m}_{\lambda_2}}} + \rho_{2NT}\, p\, (\hat{m}_{\lambda_2} + 1), \tag{4.2}$$

where $\rho_{2NT}$ is a tuning parameter. Denote $\Omega_2 = [0, \lambda_{2\max}]$, a bounded interval in $\mathbb{R}^+$. We divide $\Omega_2$ into three subsets $\Omega_{20}$, $\Omega_{2-}$ and $\Omega_{2+}$ as follows

$$\Omega_{20} = \left\{\lambda_2 \in \Omega_2 : \hat{m}_{\lambda_2} = m^0\right\},\ \Omega_{2-} = \left\{\lambda_2 \in \Omega_2 : \hat{m}_{\lambda_2} < m^0\right\},\ \text{and } \Omega_{2+} = \left\{\lambda_2 \in \Omega_2 : \hat{m}_{\lambda_2} > m^0\right\}.$$

Let $\lambda^0_{2NT}$ denote an element in $\Omega_{20}$ that also satisfies the conditions on $\lambda_2$ in Assumptions B.2(ii)-(iii).

To state the next result, we add the following assumptions.

**Assumption B.4**. (i) $\text{plim}_{N\to\infty} \min_{1\le j\le m^0} \min_{\alpha\in\mathbb{R}^p} \frac{1}{J^2_{\min}} \eta_j(\alpha)' W_{T^0_j} \eta_j(\alpha) \ge \underline{c}_\alpha > 0$, where $\eta_j(\alpha) = \frac{1}{N}\sum_{i=1}^N [(\alpha^0_{j+1} - \alpha)' x_{iT^0_j} - (\alpha^0_j - \alpha)' x_{i,T^0_j-1}] z_{iT^0_j}$.

(ii) $\frac{1}{\sqrt{N(I^0_j-1)}} \sum_{t=T^0_{j-1}+1}^{T^0_j-1} \sum_{i=1}^N z_{it}\Delta u_{it} = O_P(1)$ for each $j = 1, ..., m^0+1$.

(iii) As $(N,T)\to\infty$, $I_{\min}\to\infty$ and $\frac{T}{(I_{\min}J_{\min})^2 N} \to 0$.

**Assumption B.5**. As $(N,T)\to\infty$, $\left(1 + \frac{T}{I_{\min}J^2_{\min}}\right)\rho_{2NT} \to 0$ and $N\rho_{2NT}\to\infty$.

Assumptions B.4-B.5 parallel A.4-A.5. Note that we now require $I_{\min}\to\infty$ as $T\to\infty$. The following theorem implies that the minimizer of $IC_2(\lambda_2)$ can only be the one that produces the correct number of estimated breaks in large samples.

**Theorem 4.6** *Suppose that Assumptions B.1, B.2(i) and B.3-B.5 hold with $c_J = \infty$ in Assumption B.2(i). Then $P\left(\inf_{\lambda_2\in\Omega_{2-}\cup\Omega_{2+}} IC_2(\lambda_2) > IC_2\left(\lambda^0_{2NT}\right)\right) \to 1$ as $N\to\infty$.*

## 4.5 The case of fixed $T$

So far we have derived the results for both the PLS and PGMM estimation under the condition $(N,T)\to\infty$. From the proofs of the above results, we can easily tell that all results continue to hold in the fixed $T$ framework. Noticeable differences mainly lie in two aspects. First, when $T$ is fixed, both $I_{\min}$ and $m^0$ are fixed integers too and all the conditions that are involved with either one can be simplified. Second, some of the proofs (e.g., those of Lemmas 3.1, B.1-B.3, 4.1, and C.1 and Theorems 3.2 and 4.2) can be greatly simplified in this case. In particular, now we can allow consecutive breaks for both PLS and PGMM estimation.

# 5 Monte Carlo simulations

In this section we conduct a set of Monte Carlo experiments to evaluate the finite sample performance of our AGFL method. The first set of experiments are concerned with the PLS or PGMM estimation of static panel data models. We first evaluate the probability of falsely detecting breaks when there are none. Then we experiment on the data generating processes (DGPs) with one or two breaks. In this case, we evaluate both the probability of correctly detecting the number of breaks and the accuracy of estimating the break dates. The second set of experiments deal with the PGMM estimation of dynamic panel data models. We focus on DGPs with a lagged dependent variable and an exogenous variable. Finally we consider the case where number of breaks increases with the time dimension.

For fast computation, we use the block-coordinate descent algorithm (see, e.g., Angelosante and Giannakis (2012)) implemented in MATLAB Excutable (MEX) to solve the minimization problem in (2.2) for the PLS case and (2.8) for the PGMM case. We select the tuning parameters $\lambda_1$ and $\lambda_2$ that minimize the information criterion in (3.3) and (4.2) for the cases of PLS and PGMM estimation, respectively. Specifically, we choose a tuning parameter $\lambda_{\max}$ that would yield zero break in every DGP and a $\lambda_{\min}$ that would yield many breaks. In practice, we can easily find such $\lambda_{\max}$ and $\lambda_{\min}$ by trial and error. We then search for the optimal tuning parameter on the 50 evenly-distributed logarithmic grids in the interval $[\lambda_{\min},\ \lambda_{\max}]$. We choose $\rho_{1NT} = \rho_{2NT} = c \ln(NT)/\sqrt{NT}$ in (3.3) or (4.2) with $c = 0.05$. Simulations (not reported here) show that the performance of our method is not sensitive to the choice of $c$, especially when $N$ or $T$ is large.

Following the literature on the adaptive Lasso, we set $\kappa_1 = \kappa_2 = 2$ in the construction of the adaptive weights $\{\dot{w}_t\}$ and $\{\ddot{w}_t\}$ that are used for the PLS and PGMM estimation, respectively. In addition, we choose all weight matrices $\{W_t,\, t = 2, ..., T\}$ and $\{W_j^p,\, j = 1, ..., \hat{m}+1\}$ as detailed in the last paragraph of Section 2.3. The number of repetitions in all subsequent Monte Carlo experiments is 1000.

## 5.1 The case of static panel

We consider the following DGPs:

$$y_{it} = \beta_t x_{it} + \mu_i + \sigma_u u_{it},\ i = 1, \dots, N,\ t = 1, \dots, T, \tag{5.1}$$

where $\mu_i = T^{-1}\sum_{t=1}^{T} x_{it}$ and

- DGP 1: $x_{it} \sim i.i.d.\ N(0,1)$, $u_{it} \sim i.i.d.\ N(0,1)$.
- DGP 2: Same as DGP 1 except $u_{it} \sim \text{AR}(1)$ for each $i : u_{it} = 0.5u_{i,t-1} + \epsilon_{it}$, $\epsilon_{it} \sim i.i.d.\ N(0, 0.75)$.
- DGP 3: Same as DGP 1 except $u_{it} \sim \text{GARCH}(1,1)$ for each $i : u_{it} = \sqrt{h_{it}}\epsilon_{it}$, $h_{it} = 0.05 + 0.05u_{i,t-1}^2 + 0.9h_{i,t-1}$, $\epsilon_{it} \sim i.i.d.\ N(0,1)$.
- DGP 4: $x_{it} = \xi_{it} + 0.3u_{it}$, $u_{it}$ and $\xi_{it}$ are $i.i.d.\ N(0,1)$ and mutually independent, $z_{it} = \xi_{it} + 0.3\epsilon_{it}$, $\epsilon_{it} \sim i.i.d.\ N(0,1)$ and independent of $u_{it}$.

- DGP 5: Same as DGP 4 except $\xi_{it} \sim \text{AR}(1)$ for each $i : \xi_{it} = 0.5\xi_{i,t-1} + \epsilon_{it}$, $\epsilon_{it} \sim i.i.d.\ N(0, 0.75)$.
- DGP 6: Same as DGP 4 except $u_{it} \sim \text{GARCH}(1,1)$ for each $i$ : $u_{it} = \sqrt{h_{it}}\epsilon_{it}$, $h_{it} = 0.05 + 0.05u_{i,t-1}^2 + 0.9h_{i,t-1}$, $\epsilon_{it} \sim i.i.d.\ N(0,1)$.

We consider $T = 6$, 12, 50 or 100, and $N = 50$, 100, and 200. For each DGP, we set $\beta_t = 1$ for all $t$ when no break exists, $\beta_t = \mathbf{1}\{1 \leq t \leq T/2\}$ when there is one break, and $\beta_t = \mathbf{1}\{1 \leq t \leq T/2\} + \mathbf{1}\{T/2 < t \leq \lfloor 2T/3 \rfloor\}$ when there are two breaks, where $\mathbf{1}\{\cdot\}$ denotes the usual indicator function and $\lfloor \cdot \rfloor$ takes the integer part. If $T = 6$, the last case allows consecutive breaks at $t = 4$ and 5.

Note that the individual effects are generated from within-average and thus regarded as "fixed effects". In the first three DGPs, no endogeneity issue exists and we use PLS to estimate the models. DGP 1 serves as the benchmark case where both the regressor and the idiosyncratic error processes are strong white noise. DGP 2 allows serial correlation in the idiosyncratic error process and DGP 3 allows conditional heteroskedasticity. DGPs 4-6 contain an endogenous variable $x_{it}$ and a variable $z_{it}$ that generates a valid IV. We apply PGMM to estimate the models, using $(z_{it}, z_{i,t-1})'$ as the instrument. DGP 4 serves as the benchmark case where both the regressor and the error terms are i.i.d. across $i$ and $t$. $x_{it}$ and $u_{it}$ are obviously correlated and $z_{it}$ is correlated with $x_{it}$ due to the presence of $\xi_{it}$ in both. DGP 5 allows serial correlation in $x_{it}$ and $z_{it}$, and DGP 6 allows conditional heteroskedasticity in $u_{it}$.

To evaluate the performance of the PLS or PGMM estimation under different noise levels, we select the scale parameter $\sigma_u$ to be 0.5 and 1. In DGP 1 without break, these values for $\sigma_u$ correspond to signal-to-noise ratios of 4 and 1 (or in terms of the goodness of fit $R^2$ of the model, 0.8 and 0.5), respectively.

Tables 1 and 2 report simulation results from the above DGPs. The first panel of Table 1 reports the percentages of falsely detecting breaks when there are none ($m^0 = 0$). The second and the third panels report the percentages of correctly estimating the number of breaks when the true numbers of breaks are 1 and 2, respectively. In the following we summarize some important findings from Table 1. First, the simulations confirm that when there are no breaks, probabilities of falsely detecting breaks decline to zero as either $N$ or $T$ increases. This is true for both the PLS estimation in DGPs 1-3 in the case of no endogenous regressor and the PGMM estimation in DGPs 4-6 in the case of an endogenous regressor. When $N = 50$ and $T = 6$ or 12, PLS and PGMM tend to over-estimate the number of breaks, especially when noise level is high. Second, when there is one or two breaks, the probabilities of correctly detecting one break converge to 100% as $N$ or $T$ increases. In the one-break case, when both $N$ and $T$ are small and the noise level is high ($\sigma_u = 1$), PLS gives poor performance in DGPs 1 and 3. However, with $N = 50$ and $T = 50$, PLS already correctly detects one break in 87% of all repetitions for DGP-1. The case for DGP 3 is similar. For DGP 2, where the error is serially correlated, PLS performs much better at small $N$ and $T$. Results from the two-break case are similar. Third, holding $T$ fixed, an increase in $N$ always leads to higher probability of correct detection. But holding $N$ fixed, an increase in $T$ does not always bring a better performance for the PGMM estimation. For example, when $N = 50$ and $T$ increases from 50 to 100, the probability of correct detection may decline slightly for the PGMM estimation. The reason

Table 1: The determination of the number of breaks for DGPs 1-6 (static panels)

| DGP | $\sigma_u$ | $N = 50$ | | | | $N = 100$ | | | | $N = 200$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $T: 6$ | 12 | 50 | 100 | 6 | 12 | 50 | 100 | 6 | 12 | 50 | 100 |
| $m^0 = 0$, % of falsely detecting breaks when there are none | | | | | | | | | | | | | |
| 1 | 0.5 | 4.4 | 0.8 | 0 | 0 | 1 | 0.1 | 0 | 0 | 0.1 | 0 | 0 | 0 |
| | 1 | 61.3 | 57.7 | 7.1 | 1.2 | 37.7 | 28 | 1.5 | 0.1 | 15.3 | 5.6 | 0.1 | 0 |
| 2 | 0.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 21.7 | 8.4 | 0.1 | 0 | 6.4 | 1.6 | 0 | 0 | 0.7 | 0.1 | 0 | 0 |
| 3 | 0.5 | 4.3 | 0.4 | 0 | 0 | 0.9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 60.3 | 55.6 | 8.4 | 1.1 | 37.8 | 24.9 | 1.4 | 0.2 | 16.4 | 7.4 | 0.1 | 0 |
| 4 | 0.5 | 27.2 | 11.8 | 0.3 | 0 | 7.9 | 2 | 0.1 | 0 | 1.8 | 0.2 | 0 | 0 |
| | 1 | 29.5 | 11.4 | 0.5 | 0 | 10.9 | 2 | 0 | 0 | 1.9 | 0.4 | 0 | 0 |
| 5 | 0.5 | 32 | 18.7 | 0.2 | 0 | 10.9 | 4.2 | 0 | 0 | 1.6 | 0.6 | 0 | 0 |
| | 1 | 31.3 | 19.1 | 0.7 | 0 | 11.1 | 3.4 | 0 | 0 | 2.1 | 0.2 | 0 | 0 |
| 6 | 0.5 | 27.6 | 13.7 | 0.5 | 0 | 11.1 | 3.3 | 0 | 0 | 0.8 | 0.2 | 0 | 0 |
| | 1 | 27.7 | 13 | 0.2 | 0 | 10.6 | 3.3 | 0 | 0 | 2.2 | 0.2 | 0 | 0 |
| $m^0 = 1$, % of correctly detecting one break | | | | | | | | | | | | | |
| 1 | 0.5 | 96.1 | 98.7 | 100 | 100 | 99.4 | 99.9 | 100 | 100 | 99.9 | 100 | 100 | 100 |
| | 1 | 43.8 | 40.3 | 87 | 94.3 | 61.2 | 71.3 | 97.3 | 99.9 | 83.9 | 92.4 | 100 | 100 |
| 2 | 0.5 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | 1 | 80.6 | 88.5 | 99.9 | 99.9 | 92 | 97.6 | 100 | 100 | 98.9 | 100 | 100 | 100 |
| 3 | 0.5 | 95.6 | 99.1 | 100 | 100 | 99.4 | 99.9 | 100 | 100 | 99.9 | 100 | 100 | 100 |
| | 1 | 46.1 | 41.8 | 86.2 | 94.5 | 63.7 | 71.4 | 97.9 | 99.6 | 85 | 90.9 | 99.8 | 100 |
| 4 | 0.5 | 75.1 | 84.1 | 97.4 | 96.7 | 92.6 | 96.5 | 100 | 100 | 98.9 | 99.5 | 100 | 100 |
| | 1 | 72.1 | 80.2 | 84.7 | 82.6 | 89.5 | 95.5 | 91.9 | 87.8 | 98.1 | 99.4 | 99.6 | 99.4 |
| 5 | 0.5 | 70.3 | 82.9 | 99 | 99.5 | 89.6 | 95 | 100 | 100 | 99 | 99.6 | 100 | 100 |
| | 1 | 65.7 | 77.8 | 93.7 | 89.5 | 88.4 | 95.5 | 98.8 | 97.5 | 97.8 | 99.8 | 99.9 | 99.9 |
| 6 | 0.5 | 74.8 | 83.1 | 99.1 | 98.7 | 92.1 | 95.2 | 100 | 100 | 98.6 | 99.6 | 100 | 100 |
| | 1 | 74.4 | 82.1 | 86.9 | 84.9 | 90.3 | 96 | 97.3 | 93.9 | 98.1 | 99.3 | 100 | 99.6 |
| $m^0 = 2$, % of correctly detecting two breaks | | | | | | | | | | | | | |
| 1 | 0.5 | 96.9 | 99 | 100 | 100 | 99.6 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | 1 | 50.5 | 40.5 | 81.3 | 91 | 71.1 | 69.8 | 96.7 | 99.7 | 86 | 90.5 | 99.8 | 100 |
| 2 | 0.5 | 99.9 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | 1 | 84.1 | 89.8 | 99.7 | 100 | 93.3 | 98.8 | 100 | 100 | 99.1 | 99.9 | 100 | 100 |
| 3 | 0.5 | 95.5 | 98.2 | 100 | 100 | 99.1 | 99.9 | 100 | 100 | 100 | 100 | 100 | 100 |
| | 1 | 49.9 | 40.9 | 82.7 | 92.3 | 68.7 | 72.6 | 97.2 | 99.8 | 86.3 | 91.6 | 99.8 | 100 |
| 4 | 0.5 | 81.6 | 84.5 | 97.2 | 96.2 | 92.8 | 95.7 | 99.9 | 99.9 | 99.4 | 99.5 | 100 | 100 |
| | 1 | 66.2 | 75.6 | 77.1 | 71.8 | 86.4 | 91.7 | 90.8 | 83.9 | 97.4 | 99.2 | 99 | 98.9 |
| 5 | 0.5 | 74.9 | 81.1 | 98.4 | 99.7 | 92.9 | 95 | 99.9 | 100 | 99.3 | 99.8 | 100 | 100 |
| | 1 | 54.5 | 69 | 87.5 | 85.1 | 79.7 | 89.8 | 98.6 | 97.8 | 96.7 | 99.7 | 99.9 | 100 |
| 6 | 0.5 | 77.7 | 80.6 | 98.7 | 98.2 | 92.2 | 96 | 100 | 99.9 | 98.2 | 99.8 | 100 | 100 |
| | 1 | 67.4 | 79.4 | 83.3 | 75.4 | 88.8 | 94 | 95 | 92.6 | 98.1 | 99.3 | 99.9 | 99.6 |

Table 2: The accuracy of estimating the break dates for DGPs 1-6 (static panels)

| DGP | $\sigma_u$ | $N = 50$ | | | | $N = 100$ | | | | $N = 200$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $T: 6$ | 12 | 50 | 100 | 6 | 12 | 50 | 100 | 6 | 12 | 50 | 100 |
| $m^0 = 1$ | | | | | | | | | | | | | |
| 1 | 0.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 0 | 0.02 | 0.00 | 0.01 | 0 | 0 | 0 | 0.00 | 0 | 0 | 0 | 0 |
| 2 | 0.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 0 | 0.02 | 0.01 | 0.01 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0.5 | 0.09 | 0.08 | 0.05 | 0.02 | 0 | 0 | 0.00 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 2.10 | 1.28 | 0.89 | 1.03 | 0.43 | 0.33 | 0.20 | 0.19 | 0 | 0.03 | 0.03 | 0.02 |
| 5 | 0.5 | 0.09 | 0.01 | 0.02 | 0.01 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 0.89 | 1.09 | 0.76 | 0.87 | 0.25 | 0.20 | 0.11 | 0.09 | 0 | 0 | 0 | 0 |
| 6 | 0.5 | 0.02 | 0.05 | 0.03 | 0.02 | 0.02 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 1.64 | 0.98 | 0.64 | 0.69 | 0.24 | 0.16 | 0.09 | 0.10 | 0 | 0.01 | 0.01 | 0.01 |
| $m^0 = 2$ | | | | | | | | | | | | | |
| 1 | 0.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 0.17 | 0.04 | 0.03 | 0.01 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0.5 | 0 | 0 | 0.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 0 | 0 | 0.02 | 0.02 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0.5 | 0.20 | 0.17 | 0.07 | 0.05 | 0.02 | 0.01 | 0 | 0.00 | 0 | 0 | 0 | 0 |
| | 1 | 2.52 | 2.07 | 1.62 | 1.49 | 0.50 | 0.35 | 0.37 | 0.31 | 0.03 | 0.07 | 0.03 | 0.03 |
| 5 | 0.5 | 0.09 | 0.08 | 0.01 | 0.02 | 0 | 0.02 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 1.65 | 1.36 | 1.20 | 1.17 | 0.19 | 0.22 | 0.13 | 0.08 | 0 | 0.01 | 0 | 0 |
| 6 | 0.5 | 0.21 | 0.12 | 0.04 | 0.01 | 0 | 0.01 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 2.32 | 2.03 | 0.89 | 1.10 | 0.73 | 0.26 | 0.20 | 0.10 | 0.02 | 0 | 0.01 | 0.01 |

Note: The table reports the ratio of the average Hausdoff distance between the estimated and true sets of break dates to $T$, i.e., 100$\cdot$ HD$(\tilde{\mathcal{T}}^0_{\tilde{m}}, \mathcal{T}^0_{m^0})/T$ in DGPs 1-3 and 100$\cdot$ HD $(\hat{\mathcal{T}}^0_{\hat{m}}, \mathcal{T}^0_{m^0})/T$ in DGPs 4-6.

is that we do not restrict the number of breaks when $T$ increases and there is a slight chance for false detection of breaks when $T$ increases. When $N$ and $T$ increase together, as in the case from $N = T = 50$ to $N = T = 100$, the performance substantially improves. Fourth, the simulations confirm that our procedure works for the case where there are two consecutive breaks. This can be observed in columns corresponding to $T = 6$ in the third panel ($m^0 = 2$).

To measure the accuracy of break-date estimation, we define the Hausdorff error of an estimated break date by its Hausdoff distance (HD) to the true sets of break date, $\mathrm{HD}(\tilde{\mathcal{T}}^0_{\tilde{m}}, \mathcal{T}^0_{m^0})$ in the case of PLS estimation and $\mathrm{HD}(\hat{\mathcal{T}}^0_{\hat{m}}, \mathcal{T}^0_{m^0})$ in the case of PGMM estimation, conditional on correction estimation of the number of breaks.[8] Note that in the case of one break, the Hausdorff error reduces to the absolute error of the estimated break. Table 2 reports the mean Hausdoff error (MHE) in percentages of $T$ (i.e., $100\cdot \mathrm{HD}(\tilde{\mathcal{T}}^0_{\tilde{m}}, \mathcal{T}^0_{m^0})/T$ in the case of PLS estimation and $100\cdot$ $\mathrm{HD}(\hat{\mathcal{T}}^0_{\hat{m}}, \mathcal{T}^0_{m^0})/T$ in the case of PGMM estimation, averaged across the 1000 replications). Conditional on the correct estimation of the number of breaks, both PLS and PGMM estimate the break dates very accurately. Even with $N = 50$ and $T = 6$, the MHE's are close to zero for PLS at both noise levels. For DGPs with endogeneity, the PGMM estimation of break-dates is less accurate, especially at high noise level, but the performance quickly improves as $N$ or $T$ increases.

## 5.2 The case of dynamic panel

We consider the following DGP's with an AR(1) dynamics:

$$y_{it} = \beta_{1t} y_{i,t-1} + \beta_{2t} x_{2it} + \mu_i + \sigma_u u_{it},$$

where $\mu_i \sim i.i.d.$ Uniform$[-0.1, 0.1]$ and

- DGP 1d: $x_{2it} \sim i.i.d.\ N(0,1)$, $u_{it} \sim i.i.d.\ N(0,1)$.
- DGP 2d: Same as DGP 1d except $x_{2it} \sim$ AR(1) for each $i : x_{2it} = 0.5x_{2i,t-1} + v_{it}$, $v_{it} \sim i.i.d.\ N(0, 0.75)$.
- DGP 3d: Same as DGP 1d except $u_{it} \sim$ GARCH(1,1) for each $i : u_{it} = \sqrt{h_{it}}\epsilon_{it}$, $h_{it} = 0.05 + 0.05u^2_{i,t-1} + 0.9h_{i,t-1}$, $\epsilon_{it} \sim i.i.d.\ N(0,1)$.

As in the static case, we take $T = 6$, 12, 50 or 100, and $N = 50$, 100, or 200. For each DGP, we set either $\beta_{1t} = \beta_{2t} = 0.5$ or more persistently, $\beta_{1t} = \beta_{2t} = 0.8$ for all $t$ when no break exists, $\beta_{1t} = \beta_{2t} = 0.3 \cdot \mathbf{1}\{1 \le t \le T/2\} + 0.7 \cdot \mathbf{1}\{T/2 < t \le T\}$ when there is one break, and $\beta_{1t} = \beta_{2t} = 0.3 \cdot \mathbf{1}\{1 \le t \le T/2\} + 0.7 \cdot \mathbf{1}\{T/2 + 1 \le t < \lfloor 2T/3 \rfloor\} + 0.3 \cdot \mathbf{1}\{\lfloor 2T/3 \rfloor + 1 \le t \le T\}$ when there are two breaks. Note that when $T = 6$, there are consecutive breaks at $t = 4$ and 5.

DGP 1d is the benchmark case with i.i.d. $x_{it}$ and $u_{it}$ across both $i$ and $t$. DGP 2d allows serial correlation in $x_{it}$ and DGP 3d allows conditional heteroskedasticity in $u_{it}$. We choose the scale parameter

[8] Let $\mathcal{D}(A,B) \equiv \sup_{b\in B} \inf_{a\in A} |a-b|$ for any two sets $A$ and $B$. The Hausdorff distance between $A$ and $B$ is defined as $\mathrm{HD}(A,B) \equiv \max\{\mathcal{D}(A,B), \mathcal{D}(B,A)\}$.

$\sigma_u$ to be 0.2, 0.4, and 0.6. The relatively lower noise levels are justified by the usually high goodness-of-fit of many dynamic panels in applications. To obtain the PGMM estimate, we use $z_{it} = (y_{i,t-2}, x_{2it}, x_{2i,t-1})'$ as the instrument.

Tables 3 and 4 report the performance of estimating the number of breaks and break dates, respectively, for these three DGPs. The first two panels of Table 3 report the percentages of falsely detecting breaks when there are none ($m^0 = 0$). The AR coefficient is 0.5 in the first panel and 0.8 in the second panel. The third and the fourth panels report the percentages of correctly estimating the number of breaks when the true numbers of breaks are 1 and 2, respectively. As in Table 2, Table 4 gives the mean Hausdorff error (MHE) for the break date estimation. We summarize the results in Tables 3 and 4 as follows. First, the simulations confirm that when there are no breaks, the probabilities of falsely detecting breaks decline to zero when $N$ or $T$ increases. When the AR coefficient increases from 0.5 to 0.8 and the dynamic panel becomes more persistent, the probabilities of false detection decrease in general, thanks to the fact that the signal-to-noise ratio is higher at higher persistence level. Second, when there is one break, the probabilities of correctly detecting one break converge to one at all noise levels. This is true even if we fix $T = 6$, in which case there would be two consecutive breaks at $t = 4$ and 5. Third, as in the static panel case, fixing $T$ and increasing $N$ always results in better performance, but not the other way around. When $N = 50$, for example, the percentages of correctly estimating the number of breaks are highest at $T = 12$ in some cases. Fourth, as in the static panel case, conditional on the correct estimation of the number of breaks, our procedure estimates the break dates very accurately. Even with $N = 50$ and $T = 6$, the MHE's are close to zero at all noise levels. When $N = 200$, the break dates are exactly estimated in most cases, conditional on correct the estimation of the number of breaks.

## 5.3 The case of increasing number of breaks

Finally we consider the case where the true number of breaks increases with the time dimension. We let $m^0 = \lfloor T^{1/3} \rfloor$ and consider the static panel equation in (5.1) with $\beta_t = \mathbf{1}\{2k\delta + 1 \leq t < (2k+1)\delta + 1\}$, $k = 0, 1, \ldots$, where $\delta = \lfloor T / \left(m^0 + 1\right) \rfloor$. Furthermore, $(x_{it}, u_{it})$ are generated from the following two DGPs,

- DGP 1i: $x_{it} \sim \text{AR}(1)$ for each $i : x_{it} = 0.5x_{i,t-1} + v_{it}$, $v_{it} \sim i.i.d.\ N(0, 0.75)$. $u_{it} \sim \text{GARCH}(1,1)$ for each $i : u_{it} = \sqrt{h_{it}}\epsilon_{it}$, $h_{it} = 0.05 + 0.05u_{i,t-1}^2 + 0.9h_{i,t-1}$, $\epsilon_{it} \sim i.i.d.\ N(0,1)$.
- DGP 2i: $x_{it} = \xi_{it} + 0.3u_{it}$, $\xi_{it} \sim \text{AR}(1)$ for each $i : \xi_{it} = 0.5\xi_{i,t-1} + v_{it}$, $v_{it} \sim i.i.d.\ N(0, 0.75)$. $z_{it} = \xi_{it} + 0.3\epsilon_{it}$, $\epsilon_{it} \sim i.i.d.\ N(0,1)$ independent of $u_{it}$. $u_{it}$ is the same as in DGP 1i.

Note that $x_{it}$ and $u_{it}$ are correlated in DGP 2i and hence $z_{it}$ is generated to form a valid IV for $x_{it}$. We use PLS in the case of DGP 1i and PGMM in the case of DGP 2i. We consider $T = 50, 100$, 200, and $N =$ 100, 200. Simulation results from 1000 repetitions are summarized in Table 5. For DGP 1i, PLS accurately estimates the number of breaks and the break dates in all cases we consider. For DGP 2i, PGMM seems to require a bigger $N$ for satisfactory performance, especially under higher noise

Table 3: The determination of the number of breaks for DGPs 1d-3d (dynamic panels)

| DGP | $\sigma_u$ | $N = 50$ | | | | $N = 100$ | | | | $N = 200$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $T : 6$ | 12 | 50 | 100 | 6 | 12 | 50 | 100 | 6 | 12 | 50 | 100 |
| | | $m^0 = 0$, $\beta_t = 0.5$, % of falsely detecting breaks when there are none. | | | | | | | | | | | |
| | 0.2 | 8.5 | 1.1 | 0 | 0 | 1.8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1d | 0.4 | 8.8 | 1.2 | 0 | 0 | 0.9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 0.6 | 11 | 1.6 | 0 | 0 | 1 | 0.1 | 0 | 0 | 0.1 | 0 | 0 | 0 |
| | 0.2 | 9.7 | 0.3 | 0 | 0 | 1.3 | 0 | 0 | 0 | 0 | 0.1 | 0 | 0 |
| 2d | 0.4 | 8.8 | 0.6 | 0 | 0 | 1.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 0.6 | 10.2 | 0.6 | 0 | 0 | 1.4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 0.2 | 9.1 | 0.8 | 0 | 0 | 1 | 0.1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3d | 0.4 | 8.1 | 0.9 | 0 | 0 | 1.1 | 0 | 0 | 0 | 0 | 0.1 | 0 | 0 |
| | 0.6 | 10 | 1.1 | 0 | 0 | 0.9 | 0.1 | 0 | 0 | 0.1 | 0 | 0 | 0 |
| | | $m^0 = 0$, $\beta_t = 0.8$, % of falsely detecting breaks when there are none. | | | | | | | | | | | |
| | 0.2 | 7.2 | 0.4 | 0 | 0 | 0.6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1d | 0.4 | 7.8 | 1.2 | 0 | 0 | 1.1 | 0 | 0 | 0 | 0.1 | 0 | 0 | 0 |
| | 0.6 | 8.1 | 1.3 | 0 | 0 | 0.8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 0.2 | 7.9 | 1.3 | 0 | 0 | 0.8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2d | 0.4 | 8.8 | 0.4 | 0 | 0 | 1 | 0.1 | 0 | 0 | 0.1 | 0 | 0 | 0 |
| | 0.6 | 8.1 | 0.7 | 0 | 0 | 1.2 | 0 | 0 | 0 | 0.1 | 0 | 0 | 0 |
| | 0.2 | 8.3 | 0.6 | 0 | 0 | 1.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3d | 0.4 | 7.5 | 0.7 | 0 | 0 | 1.2 | 0.1 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 0.6 | 9.5 | 0.9 | 0 | 0 | 1.4 | 0 | 0 | 0 | 0.1 | 0 | 0 | 0 |
| | | $m^0 = 1$, % of correctly detecting one break | | | | | | | | | | | |
| | 0.2 | 90.5 | 98.5 | 99.3 | 98.6 | 98 | 99.7 | 100 | 100 | 99.8 | 100 | 100 | 100 |
| 1d | 0.4 | 90.6 | 96 | 89.9 | 85.6 | 98 | 99.6 | 99 | 98.5 | 100 | 100 | 100 | 100 |
| | 0.6 | 85.7 | 93.1 | 89.5 | 83.1 | 98.6 | 99.3 | 98.3 | 96.2 | 99.9 | 100 | 100 | 100 |
| | 0.2 | 91.5 | 98.8 | 99.6 | 99.7 | 98.8 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| 2d | 0.4 | 89.5 | 97.6 | 93.3 | 86.7 | 98.7 | 99.9 | 99.7 | 98.7 | 100 | 100 | 100 | 100 |
| | 0.6 | 86 | 92.4 | 91.1 | 84.3 | 98.8 | 99.3 | 97.9 | 96.8 | 100 | 100 | 100 | 100 |
| | 0.2 | 91.2 | 98.7 | 99.5 | 98.2 | 98.7 | 99.8 | 100 | 100 | 99.9 | 100 | 100 | 100 |
| 3d | 0.4 | 90.5 | 94.2 | 91.4 | 84.3 | 99.5 | 99.4 | 98.9 | 98.2 | 100 | 100 | 100 | 100 |
| | 0.6 | 87.1 | 92.6 | 87.3 | 82.5 | 97.8 | 98.8 | 96.4 | 95.3 | 100 | 100 | 99.8 | 99.8 |
| | | $m^0 = 2$, % of correctly detecting two breaks | | | | | | | | | | | |
| | 0.2 | 94.6 | 98.4 | 98.6 | 98.3 | 99.4 | 99.7 | 100 | 100 | 100 | 100 | 100 | 100 |
| 1d | 0.4 | 88 | 87.7 | 81.6 | 72.5 | 99.1 | 99.7 | 99.3 | 97.4 | 100 | 100 | 99.9 | 100 |
| | 0.6 | 61.3 | 46.4 | 35.5 | 44.2 | 90.2 | 86.6 | 94.3 | 92.1 | 99.9 | 99.9 | 99.9 | 99.9 |
| | 0.2 | 94.1 | 98.5 | 99.8 | 98.8 | 99.4 | 99.9 | 100 | 100 | 100 | 100 | 100 | 100 |
| 2d | 0.4 | 92.2 | 91.7 | 89.1 | 79.5 | 99.2 | 100 | 99.7 | 98.7 | 99.9 | 100 | 100 | 100 |
| | 0.6 | 75.7 | 57.4 | 22.8 | 35.5 | 96.6 | 93.2 | 93.9 | 94.8 | 100 | 100 | 100 | 100 |
| | 0.2 | 93 | 97.6 | 99.2 | 97.6 | 99.2 | 99.9 | 100 | 100 | 100 | 100 | 100 | 100 |
| 3d | 0.4 | 89 | 88.8 | 78.3 | 70.7 | 99.4 | 99.4 | 99.1 | 97.5 | 99.9 | 100 | 100 | 100 |
| | 0.6 | 63.3 | 43.2 | 33.8 | 37.8 | 89 | 85 | 92.3 | 89.5 | 99.6 | 99.9 | 99.7 | 99.5 |

Table 4: The accuracy of estimating the break dates for DGPs 1d-3d (dynamic panels)

| DGP | $\sigma_u$ | $N=50$ | | | | $N=100$ | | | | $N=200$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $T:6$ | 12 | 50 | 100 | 6 | 12 | 50 | 100 | 6 | 12 | 50 | 100 |
| | | $m^0=1$ | | | | | | | | | | | |
| | 0.2 | 0 | 0 | 0.01 | 0.01 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1d | 0.4 | 0.06 | 0.06 | 0.25 | 0.30 | 0 | 0 | 0.00 | 0.01 | 0 | 0 | 0 | 0 |
| | 0.6 | 0.89 | 0.70 | 0.88 | 1.46 | 0.02 | 0.08 | 0.06 | 0.11 | 0 | 0 | 0 | 0.00 |
| | 0.2 | 0 | 0 | 0 | 0.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2d | 0.4 | 0.02 | 0.03 | 0.23 | 0.33 | 0 | 0 | 0.02 | 0.01 | 0 | 0 | 0 | 0 |
| | 0.6 | 0.19 | 0.37 | 0.98 | 1.51 | 0.03 | 0.06 | 0.06 | 0.07 | 0 | 0 | 0.00 | 0 |
| | 0.2 | 0 | 0 | 0.01 | 0.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3d | 0.4 | 0.09 | 0.08 | 0.29 | 0.25 | 0 | 0 | 0.02 | 0.02 | 0 | 0 | 0 | 0 |
| | 0.6 | 0.69 | 0.74 | 1.19 | 1.49 | 0.05 | 0.13 | 0.07 | 0.11 | 0 | 0 | 0 | 0 |
| | | $m^0=2$ | | | | | | | | | | | |
| | 0.2 | 0 | 0 | 0.01 | 0.00 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1d | 0.4 | 0.15 | 0.06 | 0.07 | 0.27 | 0 | 0 | 0.01 | 0.01 | 0 | 0 | 0 | 0 |
| | 0.6 | 0.60 | 0.57 | 0.17 | 0.23 | 0.04 | 0.03 | 0.04 | 0.05 | 0 | 0 | 0.00 | 0 |
| | 0.2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2d | 0.4 | 0.02 | 0.01 | 0.03 | 0.11 | 0 | 0 | 0.00 | 0.01 | 0 | 0 | 0 | 0 |
| | 0.6 | 0.40 | 0.13 | 0.01 | 0.04 | 0.07 | 0 | 0.00 | 0.03 | 0 | 0 | 0 | 0 |
| | 0.2 | 0 | 0 | 0.00 | 0.01 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3d | 0.4 | 0.13 | 0.05 | 0.10 | 0.21 | 0 | 0 | 0.00 | 0.01 | 0 | 0 | 0 | 0 |
| | 0.6 | 0.66 | 0.37 | 0.15 | 0.29 | 0.11 | 0.06 | 0.04 | 0.10 | 0.02 | 0 | 0.00 | 0 |

Note: The table reports the ratio of the average Hausdoff distance between the estimated and true sets of break dates to $T$, i.e., 100$\cdot$ HD$(\hat{\mathcal{T}}^0_{\hat{m}}, \mathcal{T}^0_{m^0})/T$.

level. Note that fixing $N$, increasing $T$ results in slightly lower percentage of correct estimation of the number of breaks in DGP 2i. Conditional on the correct estimation of the number of breaks, PGMM also performs well in the estimation of break dates. Overall, we may conclude that both PLS and PGMM can satisfactorily deal with the case of increasing number of breaks.

# 6 An empirical application

In this section we offer an illustration of the use of our method. We seek to evaluate the effect of FDI inflow on economic growth by using a dynamic panel data model with an unknown number of breaks.

Table 5: Monte Carlo simulations for the case of increasing number of breaks

| DGP | $\sigma_u$ | $N=100$ | | | $N=200$ | | | $N=100$ | | | $N=200$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $T:50$ | 100 | 200 | 50 | 100 | 200 | 50 | 100 | 200 | 50 | 100 | 200 |
| 1i | 0.5 | 100 | 100 | 100 | 100 | 100 | 100 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 1 | 96.9 | 99.9 | 99.9 | 100 | 100 | 100 | 0 | 0.00 | 0.00 | 0 | 0 | 0 |
| 2i | 0.5 | 99.8 | 99.9 | 100 | 100 | 100 | 100 | 0 | 0.00 | 0.00 | 0 | 0 | 0 |
| | 1 | 94.2 | 94.3 | 90.5 | 100 | 99.8 | 99.8 | 0.15 | 0.17 | 0.10 | 0.00 | 0.01 | 0.01 |

Note: The true number of breaks is $m^0 = \lfloor T^{1/3} \rfloor$. The left panel reports the percentages of correctly estimating the number of breaks. The right panel reports the ratio of the average Hausdoff distance (HD) between the estimated and the true sets of breaks to $T$ (100$\cdot$HD$(\tilde{\mathcal{T}}^0_{\tilde{m}}, \mathcal{T}^0_{m^0})/T$)

The possible existence of breaks may be justified theoretically. In the endogenous growth model of Romer (1986), for example, economic growth may behave differently in different policy environments. Furthermore, in the growth model of Jones (2002), the regime shifts may be common across countries in "a world of ideas", assuming that ideas propagate fast enough. Empirically, there is ample evidence of the existence of breaks in growth path (e.g., Ben-David and Papell 1995). However, most of existing studies rely on time series structural break tests for individual economies, the United States in particular.

In this empirical exercise, we use a panel data of 88 countries or regions from 1972 to 2012. We obtain the natural logarithm of per capita GDP ($y_{i\tau}$), net FDI inflow ($F_{i\tau}$), and GDP ($GDP_{i\tau}$), $\tau = 1, \ldots, 41$, from the UNCTAD (United Nations Conference on Trade and Development) database.[9] Following the literature on growth empirics (e.g., Islam 1995), we work on five-year averages of the data. For $t = 0, \ldots, 7$, define $R_{it} = (y_{i,5(t+1)+1} - y_{i,5t+1})/5$, which is the $t$th five-year average growth in GDP per capita, and $Y_{it}^0 = \sum_{s=1}^{5} y_{i,5t+s}/5$, which is the $t$th five-year average of log per capita GDP with one-year-lag behind $R_{it}$. Furthermore, we construct the ratio of the net FDI inflow to GDP ($F_{i\tau}/GDP_{i\tau}$), obtain similar five-year averages, and denote them by $FDI_{it}$. The averaging gives us eight five-year time periods for each economy. Due to the fact that there is one lagged dependent variable in the model, the effective number of data points for each economy is seven. We apply the PGMM method to estimate the following dynamic panel data model with an unknown number of breaks,

$$R_{it} = \mu_i + \beta_{1t} R_{i,t-1} + \beta_{2t} FDI_{it} + \beta_{3t} Y_{it}^0 + u_{it}, \quad t = 1, \ldots, 7,$$

where $\mu_i$ is the country-specific effect, $\beta_{1t}$ is the AR(1) coefficient, $\beta_{2t}$ is the parameter of interest that measures the effect of FDI on growth, and $\beta_{3t} Y_{it}^0$ controls the "initial" income level. A negative $\beta_{3t}$ would imply "convergence" in economic growth. As in the simulations, we set $\kappa_2 = 2$ in the construction of the adaptive weights, choose the weight matrices $(W_t, W_j^p)$ as detailed in the last paragraph of Section 2.3, and adopt $z_{it} = \left(R_{i,t-2}, FDI_{it}, FDI_{i,t-1}, Y_{it}^0, Y_{i,t-1}^0\right)'$ as the instrument.

We choose $\lambda_{\max} = 10$, which results in zero break, and $\lambda_{\min} = 0.002$, which results in six breaks. We then search on the interval $[\lambda_{\min},\ \lambda_{\max}]$ with fifty evenly-distributed logarithmic grids. As in the simulations, we set $\rho_{2NT} = 0.05 \ln(NT)/\sqrt{NT}$. The information criterion $IC_2$ selects a model that contains three breaks at $t = 5$, 6 and 7, corresponding to the 1998-2002, 2003-2007, and 2008-2012 periods. Figure 1 shows how $IC_2(\lambda_2)$ (left axis) and the estimated number of breaks (right axis) change with the tuning parameter $\lambda_2$. We can see that the $IC_2$ declines until the estimated number of breaks reaches three and rises as $\lambda_2$ gets bigger. It is notable that there are five $\lambda_2$'s that result in three breaks, ranging from 0.195 to 0.343, and the IC curve is flat over this segment (and similarly over several other segments).[10] This suggests that the penalized GMM estimation is not very sensitive to the tuning parameter.

[9] The UNCTAD database covers 237 countries and regions. We delete those economies with missing values over 1972-2012.

[10] When $\lambda_2$ changes from 0.195 to 0.343, the number of breaks and the set of estimated break dates remain unchanged so that neither the first term (corresponding to the *post* Lasso regression) nor the second term (the penalty term) in (4.2) changes.

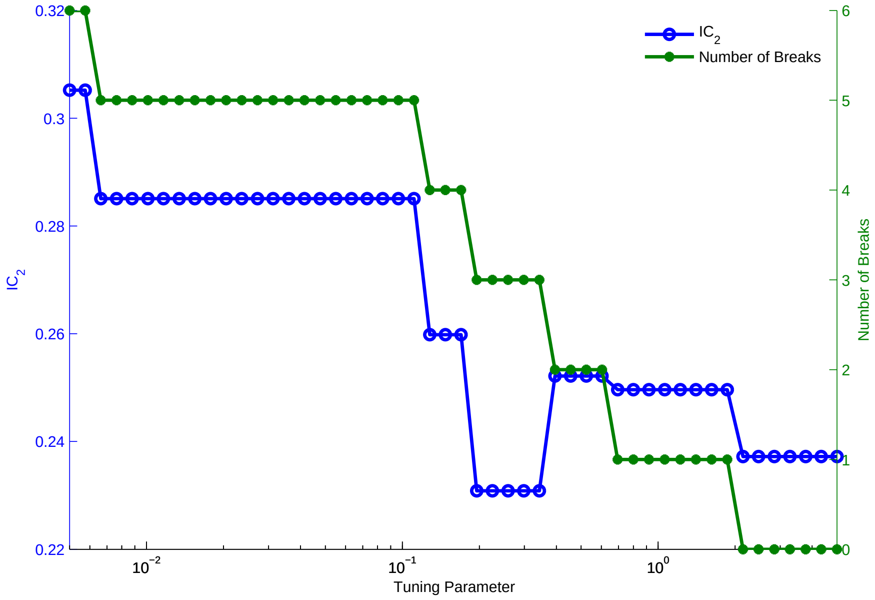

Figure 1: Selecting the optimal tuning parameter by minimizing the information criterion ($IC_2$). Horizontal axis: tuning parameter, Left vertical axis: $IC_2$, right vertical axis: number of breaks.

Table 6: The effect of FDI on the economic growth (88 countries and regions, 1978-2012)

| $\hat{m}$ | $t$ | 1 (78-82) | 2 (83-87) | 3 (88-92 | 4 (93-97) | 5 (98-02) | 6 (03-07) | 7 (08-12) |
|---|---|---|---|---|---|---|---|---|
| 0 | $R_{i,t-1}$ | -.118(.052)$^b$ | | | | | | |
| | $FDI_{it}$ | .103(.053)$^a$ | | | | | | |
| | $Y_{it}^0$ | -1.518(.294)$^c$ | | | | | | |
| 1 | $R_{i,t-1}$ | -.090(.054)$^a$ | | | | | | .169(.327) |
| | $FDI_{it}$ | .115(.053)$^b$ | | | | | | .084(.080) |
| | $Y_{it}^0$ | -.633(.299)$^b$ | | | | | | -.876(.320)$^c$ |
| 2 | $R_{i,t-1}$ | -.027(.053) | | | | -.148(.102) | | .141(.335) |
| | $FDI_{it}$ | .548(.086)$^c$ | | | | .154(.045)$^c$ | | .110(.076) |
| | $Y_{it}^0$ | -.162(.337) | | | | -.186(.320) | | -.449(.347) |
| 3 | $R_{i,t-1}$ | -.014(.052) | | | | -.180(.126) | .284(.179) | .206(.288) |
| | $FDI_{it}$ | .556(.091)$^c$ | | | | .148(.051)$^c$ | .261(.075)$^c$ | .150(.068)$^b$ |
| | $Y_{it}^0$ | -1.085(.373)$^c$ | | | | -1.067(.357)$^c$ | -1.063(.357)$^c$ | -1.203(.360)$^c$ |
| 4 | $R_{i,t-1}$ | -.034(.117) | .046(.061) | | | -.154(.132) | .303(.182)$^a$ | .248(.291) |
| | $FDI_{it}$ | .899 (.239)$^c$ | .544 (.093)$^c$ | | | .153(.052)$^c$ | .264(.077)$^c$ | .149(.070)$^b$ |
| | $Y_{it}^0$ | -1.467(.365)$^c$ | -1.367(.367)$^c$ | | | 1.341(.358)$^c$ | -1.327(.357)$^c$ | -1.462(.359)$^c$ |
| 5 | $R_{i,t-1}$ | -.061(.121) | .070(.104) | | .100(.076) | -.123(.135) | .332(.197)$^a$ | .343(.305) |
| | $FDI_{it}$ | .354(.240) | -.070(.227) | | .516(.099)$^c$ | 0.155(.054)$^c$ | .265(.077)$^c$ | .143(.074)$^a$ |
| | $Y_{it}^0$ | -2.120(.573)$^c$ | -2.018(.575)$^c$ | | -2.004(.543)$^c$ | -1.967(.536)$^c$ | -1.936(.525)$^c$ | -2.064(.524)$^c$ |
| 6 | $R_{i,t-1}$ | -.093(.120) | .076(.095) | .018(.082) | .149(.097) | -.063(.149) | .388(.218)$^a$ | .521(.337) |
| | $FDI_{it}$ | .417(.261) | .001(.231) | .606(.187)$^c$ | .526(.101)$^c$ | .158(.055)$^c$ | .267(.079)$^c$ | .133(.083) |
| | $Y_{it}^0$ | -3.491(.640)$^c$ | -3.377(.640)$^c$ | -3.304(.639)$^c$ | -3.182(.600)$^c$ | -3.122(.605)$^c$ | -3.062(.597)$^c$ | -3.177(.596)$^c$ |

Note: Standard errors are in parentheses. The superscripts $a$, $b$, and $c$ indicate statistical significance at 10%, 5%, and 1% levels, respectively.

It is well known that information criteria may not be able to select the right model in finite samples. It is thus prudent to examine the cases with the number of breaks other than three. Table 6 shows regime segmentation, parameter estimates, and standard errors (in parentheses) from the post-Lasso estimation for the cases where $\hat{m} = 0, 1, \ldots, 6$. Note that in the last case ($\hat{m} = 6$), there is a structural break at every time point.

As shown in Table 6, the set of break dates is an increasing sequence as the tuning parameter decreases. It starts from an empty set when $\hat{m} = 0$. When $\hat{m} = 1$, we have one break at $t = 7$, which corresponds to the five-year period of 2008-2012. As the tuning parameter decreases, another break (in addition to the one at $t = 7$) is detected at $t = 5$, which corresponds to the period 1998-2002. As the tuning parameter decreases further, we arrive at the case of $\hat{m} = 3$ that achieves the minimum $IC_2$ and the set of breaks is now $\{5, 6, 7\}$. When $\hat{m} = 4$, the set of breaks becomes $\{2, 5, 6, 7\}$, and when $\hat{m} = 5$, it is enlarged to $\{2, 3, 5, 6, 7\}$. Finally, $\hat{m} = 6$ corresponds to the case where breaks occur at every period.

Table 6 demonstrate that the determination of structural change in the model is crucial for the quantitative evaluation of the effect of FDI on the economic growth. In the model chosen by $IC_2$ ($\hat{m} = 3$) the coefficients of FDI are significantly positive at 5% level in all regimes, and the FDI effect on growth has declined substantially since the turn of the new millennium. If we assume that no break exists and estimate a textbook dynamic panel data model, the time-varying character of the FDI effect would be lost.

Furthermore, as shown in the top panel of Table 6, the magnitude of the estimated FDI effect is smaller than any regime in the model with three breaks. Indeed, it fails to pass the significance test at 5% level. In the model with three breaks, the coefficients of initial per capita GDP are significantly negative in all regimes, confirming the convergence story. The magnitudes of the convergence effects are, however, lower than that in the model with time-invariant coefficients. This empirical exercise suggests that the time-invariant parameter in the textbook dynamic panel data model is an unnecessarily restrictive assumption and may lead to erroneous conclusions. Our shrinkage-based method, by allowing multiple breaks in panel data model, provides applied economists with a natural approach to relaxing this assumption.

# 7 Conclusion

We propose two shrinkage procedures for the determination of the number of structural changes in linear panel data models via adaptive group fused Lasso: PLS estimation for first-differenced models without endogeneity and PGMM estimation for first-differenced models with endogeneity. We show that with probability tending to one our methods can correctly determine the true number of breaks and estimate the break dates consistently. Simulation results suggest that our methods perform well in finite samples.

There are several interesting topics for further research. First, we do not allow cross section dependence in our models. Given the large literature on cross section dependence, it is interesting to extend our methodology to panel data models with cross section dependence. Second, if we model the cross section dependence through a factor structure, the factor loadings may also exhibit structural changes over time (see, e.g., Breitung and Eickmeier 2011, Cheng et al. 2015, and Su and Wang 2015) and this further complicates the analysis. Third, we consider the common shocks for homogenous panel data models. It is also interesting to consider heterogeneous panel data models and to allow the break dates to be different across individuals. We leave these topics for future research.

# APPENDIX

# A Definitions of several matrices

In this appendix, we define several matrices used in the main text.

## A.1 Penalized least squares estimation

Let $\phi_{ab,ts} = \frac{1}{N}\sum_{i=1}^{N} a_{it}b_{is}'$ and $\phi_{ab,t} = \phi_{ab,tt}$ for $t,\ s = 1, ..., T$, and $a,\ b = x,\ \Delta x,\ \Delta y,\ \Delta^2 y,\ \Delta u$ or $\Delta^2 u$. For example, $\phi_{x\Delta^2 y,t,t+1} = \frac{1}{N}\sum_{i=1}^{N} x_{it}\Delta^2 y_{i,t+1}$ for $t = 2, ..., T-1$. Let $\mathrm{TriD}(\cdot,\cdot)_T$ be as defined in (1.1). Define

$$\dot{Q}_{NT} = \mathrm{TriD}(Q^{\dagger}, Q)_T, \tag{A.1}$$

$$\dot{R}_{NT}^{a} = (-\phi_{x\Delta a,2}', -\phi_{x\Delta^2 a,2,3}', -\phi_{x\Delta^2 a,3,4}' ..., -\phi_{x\Delta^2 a,T-1,T}', \phi_{x\Delta a,T}')',\ a = y \text{ or } u, \tag{A.2}$$

where $Q_t = \phi_{xx,t}$ for $t = 1$ and $T$, $Q_t = 2\phi_{xx,t}$ for $2 \leq t \leq T-1$, and $Q_t^{\dagger} = \phi_{xx,t,t-1}$ for $t = 2, ..., T$. (2.5) indicates that $\dot{\boldsymbol{\beta}} = \dot{Q}_{NT}^{-1}\dot{R}_{NT}^{y}$.

Recall that $\mathcal{T}_m = \{T_1, ..., T_m\}$ where $T_0 = 1$ and $T_{m+1} = T$. Let $\Phi_{ab,l}(\mathcal{T}_m) = \frac{1}{N}\sum_{t=T_{l-1}+1}^{T_l-1}\sum_{i=1}^{N} a_{it}b_{it}'$ for $l = 1, ..., m+1$ and $a, b = \Delta x,\ x,$ or $\Delta y$. Define the $p(m+1) \times p(m+1)$ matrix $\Phi_{NT}(\mathcal{T}_m)$ and $p(m+1) \times 1$ vector $\Psi_{NT}^{a}(\mathcal{T}_m)$, respectively:

$$\Phi_{NT}(\mathcal{T}_m) = \mathrm{TriD}\left(\Phi^{\dagger}(\mathcal{T}_m), \Phi(\mathcal{T}_m)\right)_{m+1}, \tag{A.3}$$

$$\begin{aligned}\Psi_{NT}^{a}(\mathcal{T}_m) = \big(&\Phi_{\Delta x\Delta a,1}' - \phi_{x\Delta y,T_1-1,T_1}', \Phi_{\Delta x\Delta a,2}' - \phi_{x\Delta a,T_2-1,T_2}' + \phi_{x\Delta a,T_1}', ..., \\ &\Phi_{\Delta x\Delta a,m}' - \phi_{x\Delta a,T_m-1,T_m}' + \phi_{x\Delta a,T_{m-1}}',\ \Phi_{\Delta x\Delta a,m+1}' + \phi_{x\Delta a,T_m}'\Big)',\ a = y \text{ or } u,\end{aligned} \tag{A.4}$$

where $\mathrm{TriD}(\cdot,\cdot)_{m+1}$ is defined analogously to $\mathrm{TriD}(\cdot,\cdot)_T$ in (1.1), $\Phi_1(\mathcal{T}_m) = \Phi_{\Delta x\Delta x,1} + \phi_{xx,T_1-1}$, $\Phi_l(\mathcal{T}_m) = \Phi_{\Delta x\Delta x,l} + \phi_{xx,T_l-1} + \phi_{xx,T_{l-1}}$ for $l = 2, ..., m$, $\Phi_{m+1}(\mathcal{T}_m) = \Phi_{\Delta x\Delta x,m+1} + \phi_{xx,T_m}$, and $\Phi_{l+1}^{\dagger}(\mathcal{T}_m) = \phi_{xx,T_l,T_l-1}$ for $l = 1, ..., m$. Then the post Lasso least squares estimator of $\boldsymbol{\alpha}^0$ and its infeasible version are respectively given by

$$\tilde{\boldsymbol{\alpha}}_{\tilde{m}}^{p}(\tilde{\mathcal{T}}_{\tilde{m}}) = \Phi_{NT}\left(\tilde{\mathcal{T}}_{\tilde{m}}\right)^{-1}\Psi_{NT}^{y}\left(\tilde{\mathcal{T}}_{\tilde{m}}\right) \text{ and } \tilde{\boldsymbol{\alpha}}_{m^0}^{p}(\mathcal{T}_{m^0}^{0}) = \Phi_{NT}\left(\mathcal{T}_{m^0}^{0}\right)^{-1}\Psi_{NT}^{y}\left(\mathcal{T}_{m^0}^{0}\right). \tag{A.5}$$

## A.2 Penalized GMM estimation

Let $\ddot{Q}_{zx,t,s} = \phi_{zx,t,s}' W_t \phi_{zx,t,s}$ and $\ddot{Q}_{zx,t} = \ddot{Q}_{zx,t,t}$ for $t, s = 1, 2, ..., T$. Let $\dot{Q}_{zx,t,t-1} = \phi_{zx,t}' W_t \phi_{zx,t,t-1}$ for $t = 2, ..., T$. Define

$$\ddot{Q}_{NT} = \mathrm{TriD}\left(\dot{Q}, \ddot{Q}\right)_T, \tag{A.6}$$

$$\begin{aligned}\ddot{R}_{NT}^{a} = \Big(&-\left(\phi_{zx,2}' W_2 \phi_{z\Delta a,2}\right)',\ \left(\phi_{zx,2}' W_2 \phi_{z\Delta a,2} - \phi_{zx,3,2}' W_3 \phi_{z\Delta a,3}\right)', ..., \\ &\left(\phi_{zx,T-1}' W_{T-1} \phi_{z\Delta a,T-1} - \phi_{zx,T,T-1}' W_T \phi_{z\Delta a,T}\right)',\ \left(\phi_{zx,T}' W_T \phi_{z\Delta a,T}\right)'\Big)',\ a = y \text{ or } u,\end{aligned} \tag{A.7}$$

where $\ddot{Q}_1 = \ddot{Q}_{zx,1,2}$, $\ddot{Q}_t = \ddot{Q}_{zx,t} + \ddot{Q}_{zx,t+1,t}$ for $t = 2, ..., T-1$, $\ddot{Q}_T = \ddot{Q}_{zx,T}$, and $\dot{Q}_t = \dot{Q}_{zx,t,t-1}$ for $t = 2, ..., T$. Note that $\ddot{\boldsymbol{\beta}} = \ddot{Q}_{NT}^{-1} \ddot{R}_{NT}^{y}$ by (2.10).

Define the $p(m+1) \times p(m+1)$ matrix $\Upsilon_{NT}(\mathcal{T}_m)$ and $p(m+1) \times 1$ vector $\Xi_{NT}^{a}(\mathcal{T}_m)$, respectively:

$$\begin{aligned} \Upsilon_{NT}(\mathcal{T}_m) &= \text{TriD}\left(\Upsilon^{\dagger}(\mathcal{T}_m), \Upsilon(\mathcal{T}_m)\right)_{m+1}, & \text{(A.8)} \\ \Xi_{NT}^{a}(\mathcal{T}_m) &= \left(\Xi_{a,1}(\mathcal{T}_m)',\ \Xi_{a,2}(\mathcal{T}_m)', ...,\ \Xi_{a,m+1}(\mathcal{T}_m)'\right)',\ a = y \text{ or } u, & \text{(A.9)} \end{aligned}$$

where TriD$(\cdot,\cdot)_{m+1}$ is defined analogously to TriD$(\cdot,\cdot)_T$ in (1.1),

$$\begin{aligned} \Upsilon_1(\mathcal{T}_m) &= \Phi_{z\Delta x,1}(\mathcal{T}_m)' W_1^p \Phi_{z\Delta x,1}(\mathcal{T}_m) + \phi'_{zx,T_1,T_1-1} W_{T_1} \phi_{zx,T_1,T_1-1}, \\ \Upsilon_l(\mathcal{T}_m) &= \Phi_{z\Delta x,l}(\mathcal{T}_m)' W_l^p \Phi_{z\Delta x,l}(\mathcal{T}_m) + \phi'_{zx,T_l,T_l-1} W_{T_l} \phi_{zx,T_l,T_l-1} + \phi'_{zx,T_{l-1}} W_{T_{l-1}} \phi_{zx,T_{l-1}} \text{ for } l\text{=}2, ..., m, \\ \Upsilon_{m+1}(\mathcal{T}_m) &= \Phi_{z\Delta x,m+1}(\mathcal{T}_m)' W_{m+1}^p \Phi_{z\Delta x,m+1}(\mathcal{T}_m) + \phi'_{zx,T_m} W_{T_m} \phi_{zx,T_m}; \\ \Upsilon_l^{\dagger}(\mathcal{T}_m) &= \phi'_{xx,T_{l-1}} W_{T_{l-1}} \phi_{xx,T_{l-1},T_{l-1}-1} \text{ for } l\text{=}2, ..., m+1; \\ \Xi_{a,1}(\mathcal{T}_m) &= \Phi_{z\Delta x,1}(\mathcal{T}_m)' W_1^p \Phi_{z\Delta a,1}(\mathcal{T}_m) - \phi'_{zx,T_1,T_1-1} W_{T_1} \phi_{z\Delta a,T_1}, \\ \Xi_{a,l}(\mathcal{T}_m) &= \Phi_{z\Delta x,l}(\mathcal{T}_m)' W_l^p \Phi_{z\Delta a,l}(\mathcal{T}_m) - \phi'_{zx,T_l,T_l-1} W_{T_l} \phi_{z\Delta a,T_l} + \phi'_{zx,T_{l-1}} W_{T_{l-1}} \phi_{z\Delta a,T_{l-1}} \text{ for } l\text{=}2, ..., m, \\ \Xi_{a,m+1}(\mathcal{T}_m) &= \Phi_{z\Delta x,m+1}(\mathcal{T}_m)' W_{m+1}^p \Phi_{z\Delta a,m+1}(\mathcal{T}_m) + \phi'_{zx,T_m} W_{T_m} \phi_{z\Delta a,T_m}. \end{aligned}$$

Then the post Lasso GMM estimator of $\boldsymbol{\alpha}^0$ and its infeasible version are respectively given by

$$\hat{\boldsymbol{\alpha}}_{\hat{m}}^{p}(\hat{\mathcal{T}}_{\hat{m}}) = \Upsilon_{NT}\left(\hat{\mathcal{T}}_{\hat{m}}\right)^{-1} \Xi_{NT}^{y}\left(\hat{\mathcal{T}}_{\hat{m}}\right) \text{ and } \hat{\boldsymbol{\alpha}}_{m^0}^{p}(\mathcal{T}_{m^0}^{0}) = \Upsilon_{NT}\left(\mathcal{T}_{m^0}^{0}\right)^{-1} \Xi_{NT}^{y}\left(\mathcal{T}_{m^0}^{0}\right). \tag{A.10}$$

# B Proof of the results in Section 3

**Proof of Lemma 3.1** (i) Recall that $\dot{Q}_{NT}$ =TriD$(Q^{\dagger}, Q)_T$ by (A.1). So $\dot{Q}_0$ is also a SBTM. Define

$$\begin{aligned} \Lambda_1 &= E\left(\phi_{xx,1}\right), \\ \Lambda_t &= 2E\left(\phi_{xx,t}\right) - E\left(\phi_{xx,t,t-1}\right)\Lambda_{t-1}^{-1}E\left(\phi_{xx,t-1,t}\right) \text{ for } t = 2, ..., T-1, \\ \Lambda_T &= E\left(\phi_{xx,T}\right) - E\left(\phi_{xx,T,T-1}\right)\Lambda_{T-1}^{-1}E\left(\phi_{xx,T-1,T}\right). \end{aligned}$$

We first argue that the above notations are well defined under Assumptions A.1(iii)-(iv) and that

$$0 < \min(\underline{c}_{xx}, \underline{c}_{\eta}) \leq \min_{1 \leq t \leq T} \mu_{\min}(\Lambda_t) \leq \max_{1 \leq t \leq T} \mu_{\max}(\Lambda_t) \leq 2\bar{c}_{xx} < \infty. \tag{B.1}$$

By Assumption A.1(iii), $\Lambda_1$ is p.d. To study the behavior of $\Lambda_t$ for $t = 2, ..., T$, we consider the auxiliary least squares projection of $x_{it}$ on $x_{i,t-1}$ :

$$x_{it} = \alpha_t^* x_{i,t-1} + \eta_{it},\ i = 1, ..., N,$$

where the pseudo true parameter $\alpha_t^*$ is chosen such that $N^{-1}\sum_{i=1}^{N} E\left(\eta_{it} x'_{i,t-1}\right) = 0$. It is easy to verify that $\alpha_t^* = E\left(\phi_{xx,t,t-1}\right)\left[E\left(\phi_{xx,t-1}\right)\right]^{-1}$ and that

$$\begin{aligned} E\left(\phi_{xx,t}\right) &= N^{-1}\sum_{i=1}^{N} E\left(x_{it}x'_{it}\right) = N^{-1}\sum_{i=1}^{N} E\left[\left(\alpha_t^* x_{i,t-1} + \eta_{it}\right)\left(\alpha_t^* x_{i,t-1} + \eta_{it}\right)'\right] \\ &= E\left(\phi_{xx,t,t-1}\right)\left[E\left(\phi_{xx,t-1}\right)\right]^{-1} E\left(\phi_{xx,t-1,t}\right) + E\left(\phi_{\eta\eta,t}\right), \end{aligned} \tag{B.2}$$

where $\phi_{\eta\eta,t} = N^{-1}\sum_{i=1}^{N}\eta_{it}\eta_{it}'$. It follows that

$$E\left(\phi_{xx,t}\right) \geq E\left(\phi_{xx,t,t-1}\right)\left[E\left(\phi_{xx,t-1}\right)\right]^{-1}E\left(\phi_{xx,t-1,t}\right) \text{ for } t = 2, ..., T,$$

which further implies that

$$\Lambda_2 = 2E\left(\phi_{xx,2}\right) - E\left(\phi_{xx,2,1}\right)\left[E\left(\phi_{xx,1}\right)\right]^{-1}E\left(\phi_{xx,1,2}\right) \geq E\left(\phi_{xx,2}\right), \tag{B.3}$$

and by induction that

$$\Lambda_t \geq 2E\left(\phi_{xx,t}\right) - E\left(\phi_{xx,t,t-1}\right)\left[E\left(\phi_{xx,t-1}\right)\right]^{-1}E\left(\phi_{xx,t-1,t}\right) \geq E\left(\phi_{xx,t}\right) \text{ for } t = 2, ..., T-1. \tag{B.4}$$

In addition, by (B.4) and (B.2)

$$\Lambda_T \geq E\left(\phi_{xx,T}\right) - E\left(\phi_{xx,T,T-1}\right)\left[E\left(\phi_{xx,T-1}\right)\right]^{-1}E\left(\phi_{xx,T-1,T}\right) = E\left(\phi_{\eta\eta,T}\right). \tag{B.5}$$

Consequently, $\min_{1\leq t\leq T}\mu_{\min}\left(\Lambda_t\right) \geq \min\left(\min_{1\leq t\leq T-1}\mu_{\min}\left\{E\left(\phi_{xx,t}\right)\right\}, \mu_{\min}\left\{E\left(\phi_{\eta\eta,T}\right)\right\}\right) \geq \min(\underline{c}_{xx}, \underline{c}_{\eta})$. In view of the fact that $E\left(\phi_{xx,t,t-1}\right)\Lambda_{t-1}^{-1}E\left(\phi_{xx,t-1,t}\right)$ is p.s.d. for $t = 2, ..., T$, we have $\Lambda_t \leq 2E\left(\phi_{xx,t}\right)$ for $t = 2, ..., T-1$ and $\Lambda_T \leq E\left(\phi_{xx,T}\right)$. It follows that $\max_{1\leq t\leq T}\mu_{\max}\left(\Lambda_t\right) \leq 2\max_{1\leq t\leq T}\mu_{\max}\left(E\left(\phi_{xx,t}\right)\right) \leq 2\bar{c}_{xx} < \infty$. That is, (B.1) follows.

Let $\Lambda$ denote a block diagonal matrix whose diagonal blocks are denoted by $\Lambda_t$ for $t = 1, ..., T$. Let $L$ denote the block lower part of $\dot{Q}_0$. By (B.1), the inverse $\Lambda^{-1}$ of $\Lambda$ exists and we can consider the block LU factorization of the SBTM $\dot{Q}_0 : \dot{Q}_0 = \left(\Lambda + L\right)\Lambda^{-1}\left(\Lambda + L'\right)$; see, e.g., Meurant (1992). By Lemma 21.2.1 in Harville (1997), the eigenvalues of the lower block triangular matrix $\Lambda + L$ and the upper block triangular matrix $\Lambda + L'$ are given by the collection of the eigenvalues of their diagonal blocks. This, in conjunction with (B.1), implies that

$$\mu_{\max}\left(\Lambda + L\right) = \mu_{\max}\left(\Lambda + L'\right) = \mu_{\max}\left(\Lambda\right) = \max_{1\leq t\leq T}\mu_{\max}\left(\Lambda_t\right) \leq 2\bar{c}_{xx} < \infty, \tag{B.6}$$

and

$$\mu_{\min}\left(\Lambda + L\right) = \mu_{\min}\left(\Lambda + L'\right) = \mu_{\min}\left(\Lambda\right) = \min_{1\leq t\leq T}\mu_{\min}\left(\Lambda_t\right) \geq \min(\underline{c}_{xx}, \underline{c}_{\eta}) > 0. \tag{B.7}$$

Then by the fact that $\mu_{\max}\left(AB\right) \leq \mu_{\max}\left(A\right)\mu_{\max}\left(B\right)$ and that $\mu_{\min}\left(AB\right) \geq \mu_{\min}\left(A\right)\mu_{\min}\left(B\right)$ for two conformable p.s.d. matrices $A$ and $B$ (see, e.g., Fact 8.14.20 in Bernstein 2005, p.329),

$$\mu_{\max}\left(\dot{Q}_0\right) = \mu_{\max}\left\{\left(\Lambda + L\right)\Lambda^{-1}\left(\Lambda + L'\right)\right\} \leq \left[\mu_{\max}\left(\Lambda + L\right)\right]^2\mu_{\max}\left(\Lambda^{-1}\right) \leq \left(2\bar{c}_{xx}\right)^2\left[\min(\underline{c}_{xx}, \underline{c}_{\eta})\right]^{-1},$$

and

$$\mu_{\min}\left(\dot{Q}_0\right) = \mu_{\min}\left\{\left(\Lambda + L\right)\Lambda^{-1}\left(\Lambda + L'\right)\right\} \geq \left[\mu_{\min}\left(\Lambda + L\right)\right]^2\mu_{\min}\left(\Lambda^{-1}\right) \geq \left(\min(\underline{c}_{xx}, \underline{c}_{\eta})\right)^2\left[2\bar{c}_{xx}\right]^{-1}.$$

So part (i) of the lemma holds with $\underline{c}_{\dot{Q}_0} = [\min(\underline{c}_{xx}, \underline{c}_{\eta})]^2[2\bar{c}_{xx}]^{-1}$ and $\bar{c}_{\dot{Q}_0} = \left(2\bar{c}_{xx}\right)^2[\min(\underline{c}_{xx}, \underline{c}_{\eta})]^{-1}$.

(ii) For notational simplicity, we assume that $p = 1$ in this proof. For any $m \times n$ matrix $A = \left(a_{ij}\right)$, define $\left\|A\right\|_1 = \max_{1\leq j\leq n}\sum_{i=1}^{m}\left|a_{ij}\right|$ and $\left\|A\right\|_\infty = \max_{1\leq i\leq m}\sum_{j=1}^{n}\left|a_{ij}\right|$. Note that $\left\|A\right\|_{\text{sp}}^2 \leq \left\|A\right\|_1\left\|A\right\|_\infty$.

Let $\bar{\phi}_{xx,t,s} = \phi_{xx,t,s} - E\left(\phi_{xx,t,s}\right)$ and $\bar{\phi}_{xx,t} = \bar{\phi}_{xx,t,t}$ for $t, s = 1, ..., T$. Then

$$\dot{Q}_{NT} - \dot{Q}_0 = \begin{pmatrix} \bar{\phi}_{xx,1} & -\bar{\phi}'_{xx,2,1} & & & & & \\ -\bar{\phi}_{xx,2,1} & 2\bar{\phi}_{xx,2} & -\bar{\phi}'_{xx,3,2} & & & & \\ & -\bar{\phi}_{xx,3,2} & 2\bar{\phi}_{xx,3} & -\bar{\phi}'_{xx,4,3} & & & \\ & \ddots & \ddots & \ddots & & & \\ & & & & -\bar{\phi}_{xx,T-1,T-2} & 2\bar{\phi}_{xx,T-1} & -\bar{\phi}'_{xx,T,T-1} \\ & & & & & -\bar{\phi}_{xx,T,T-1} & \bar{\phi}_{xx,T} \end{pmatrix}.$$

Let $[B]_{st}$ denote the $(s,t)$th element of a matrix $B$. By the symmetry of $\dot{Q}_{NT} - \dot{Q}_0$, we have

$$\begin{aligned} \left\|\dot{Q}_{NT} - \dot{Q}_0\right\|_\infty &= \left\|\dot{Q}_{NT} - \dot{Q}_0\right\|_1 = \max_{1\leq t\leq T}\sum_{s=1}^T \left|\left[\dot{Q}_{NT} - \dot{Q}_0\right]_{st}\right| \\ &= \max\left\{\left|\bar{\phi}_{xx,1}\right| + \left|\bar{\phi}_{xx,2,1}\right|,\ \left|\bar{\phi}_{xx,2,1}\right| + 2\left|\bar{\phi}_{xx,2}\right| + \left|\bar{\phi}_{xx,3,2}\right|,\ \left|\bar{\phi}_{xx,3,2}\right| + 2\left|\bar{\phi}_{xx,3}\right| + \left|\bar{\phi}_{xx,4,3}\right|,\right. \\ &\qquad \left. ...,\ \left|\bar{\phi}_{xx,T-1,T-2}\right| + 2\left|\bar{\phi}_{xx,T-1}\right| + \left|\bar{\phi}_{xx,T,T-1}\right|,\ \left|\bar{\phi}_{xx,T,T-1}\right| + \left|\bar{\phi}_{xx,T}\right|\right\} \\ &\leq 2\left\{\max_{1\leq t\leq T}\left|\bar{\phi}_{xx,t}\right| + \max_{2\leq t\leq T}\left|\bar{\phi}_{xx,t,t-1}\right|\right\}. \end{aligned}$$

The upper bound is $o_P(1)$ provided that $\max_{1\leq t\leq T}\left|\bar{\phi}_{xx,t}\right| = o_P(1)$ and $\max_{2\leq t\leq T}\left|\bar{\phi}_{xx,t,t-1}\right| = o_P(1)$. We only show the former one as the proof of the second claim is similar. Let $c_{NT} = N(\ln T)^{-\epsilon_0}$ for some $\epsilon_0 > 1$. Define $\varsigma_{it}^{(1)} = x_{it}^2 \mathbf{1}_{it} - E\left(x_{it}^2 \mathbf{1}_{it}\right)$ and $\varsigma_{it}^{(2)} = x_{it}^2 \bar{\mathbf{1}}_{it} - E\left(x_{it}^2 \bar{\mathbf{1}}_{it}\right)$ where $\mathbf{1}_{it} = \mathbf{1}\left\{x_{it}^2 \leq c_{NT}\right\}$ and $\bar{\mathbf{1}}_{it} = 1 - \mathbf{1}_{it}$. Then $\bar{\phi}_{xx,t} = \frac{1}{N}\sum_{i=1}^N \varsigma_{it}^{(1)} + \frac{1}{N}\sum_{i=1}^N \varsigma_{it}^{(2)}$. For any $\epsilon > 0$, by Bernstein inequality for independent random variables (e.g., Serfling 1980, p.95)

$$\begin{aligned} P\left(\max_{1\leq t\leq T}\left|\frac{1}{N}\sum_{i=1}^N \varsigma_{it}^{(1)}\right| \geq \epsilon\right) &= \sum_{t=1}^T P\left(\left|\sum_{i=1}^N \varsigma_{it}^{(1)}\right| \geq N\epsilon\right) \\ &\leq 2T \max_{1\leq t\leq T} \exp\left(-\frac{N^2\epsilon^2}{2\sum_{i=1}^N \mathrm{Var}\left(\varsigma_{it}^{(1)}\right) + \frac{2}{3}c_{NT}N\epsilon}\right) = o(1) \end{aligned}$$

where we use the fact that $\sum_{i=1}^N \mathrm{Var}\left(\varsigma_{it}^{(1)}\right) \leq \max_{1\leq t\leq T}\sum_{i=1}^N E\left(X_{it}^4\right) = O(N)$ by Assumption A.1(ii). By Markov inequality, Lebesgue dominated convergence theorem, and Assumptions A.1(ii) and A2(iv)

$$\begin{aligned} P\left(\max_{1\leq t\leq T}\left|\frac{1}{N}\sum_{i=1}^N \varsigma_{it}^{(2)}\right| \geq \epsilon\right) &\leq P\left(\max_{1\leq t\leq T}\max_{1\leq i\leq N} x_{it}^2 > c_{NT}\right) \\ &\leq \frac{1}{c_{NT}^{\tau_0}}\sum_{i=1}^N\sum_{t=1}^T E\left[x_{it}^{2\tau_0}\mathbf{1}\left\{x_{it}^2 > c_{NT}\right\}\right] = o(1), \end{aligned}$$

where we use the fact that $NTc_{NT}^{\tau_0} = N^{1-\tau_0}T(\ln T)^{\epsilon_0\tau_0} = o(1)$ under Assumption A.2(iv). Consequently $P\left(\max_{1\leq t\leq T}\left|\bar{\phi}_{xx,t}\right| \geq 2\epsilon\right) = o(1)$ for any $\epsilon > 0$ and $\max_{1\leq t\leq T}\left|\bar{\phi}_{xx,t}\right| = o_P(1)$. Analogously, we can show that $\max_{2\leq t\leq T}\left|\bar{\phi}_{xx,t,t-1}\right| = o_P(1)$. It follows that $\left\|\dot{Q}_{NT} - \dot{Q}_0\right\|_1 = \left\|\dot{Q}_{NT} - \dot{Q}_0\right\|_\infty = o_P(1)$ and $\left\|\dot{Q}_{NT} - \dot{Q}_0\right\|_{\mathrm{sp}} = o_P(1)$.

(iii) By (i)-(ii), we have w.p.a.1

$$\mu_{\min}\left(\dot{Q}_{NT}\right) = \min_{\|\varkappa\|=1}\left\{\varkappa'\dot{Q}_0\varkappa + \varkappa'\left(\dot{Q}_{NT} - \dot{Q}_0\right)\varkappa\right\} \geq \mu_{\min}\left(\dot{Q}_0\right) - \left\|\dot{Q}_{NT} - \dot{Q}_0\right\|_{\mathrm{sp}} \geq \underline{c}_{\dot{Q}_0}/2$$

and

$$\mu_{\max}\left(\dot{Q}_{NT}\right) = \max_{\|\varkappa\|=1}\left\{\varkappa'\dot{Q}_0\varkappa + \varkappa'\left(\dot{Q}_{NT} - \dot{Q}_0\right)\varkappa\right\} \leq \mu_{\max}\left(\dot{Q}_0\right) + \left\|\dot{Q}_{NT} - \dot{Q}_0\right\|_{\mathrm{sp}} \geq 2\bar{c}_{\dot{Q}_0}. \blacksquare$$

Below, we use $\sum_{t\in\mathcal{T}^0_{m^0}}$ and $\sum_{t\in\mathcal{T}^{0c}_{m^0}}$ to denote $\sum_{t=2,t\in\mathcal{T}^0_{m^0}}^T$ and $\sum_{t=2,t\in\mathcal{T}^{0c}_{m^0}}^T$, respectively. The following lemmas are used in the proofs of our main results and their proofs are given in the online supplemental appendix.

**Lemma B.1** *Suppose Assumptions A.1 and A.2(iv) hold. Then* $\dot{\beta}_t - \beta^0_t = O_P\left(N^{-1/2}\right)$ *for each* $t = 1, 2, ..., T$.

**Lemma B.2** *Suppose that the conditions in Theorem 3.6 hold. Let* $\mathbb{T}_m = \{\mathcal{T}_m = \{T_1, ..., T_m\} : 2 \leq T_1 < ... < T_m \leq T,\ T_0 = 1$ *and* $T_{m+1} = T+1\}$. *Then* $\min_{0\leq m<m^0}\inf_{\mathcal{T}_m\in\mathbb{T}_m}\frac{(T-1)}{I_{\min}J^2_{\min}}(\tilde{\sigma}^2_{\mathcal{T}_m} - \tilde{\sigma}^2_{\mathcal{T}^0_{m^0}}) \geq c + o_P(1)$ *for some* $c > 0$.

**Lemma B.3** *Suppose that the conditions in Theorem 3.6 hold. Let* $\bar{\mathbb{T}}_m = \{\mathcal{T}_m = \{T_1, ..., T_m\} : \mathcal{T}_{m^0} \subset \mathcal{T}_m,$ $2 \leq T_1 < ... < T_m \leq T\}$ *where* $m^0 < m \leq m_{\max}$. *Then* $\max_{m^0<m\leq m_{\max}}\sup_{\mathcal{T}_m\in\bar{\mathbb{T}}_m} N^{-1}\left|\tilde{\sigma}^2_{\mathcal{T}_m} - \tilde{\sigma}^2_{\mathcal{T}_{m^0}}\right| = O_P(1)$.

**Proof of Theorem 3.2.** (i) Let $\beta_t = \beta^0_t + N^{-1/2}b_t$ for $t = 1, ..., T$ and $\mathbf{b} \equiv (b'_1, ..., b'_T)'$. Note that $\boldsymbol{\beta} = \boldsymbol{\beta}^0 + N^{-1/2}\mathbf{b}$. Let $\tilde{b}_t = N^{1/2}(\tilde{\beta}_t - \beta^0_t)$ and $\tilde{\mathbf{b}} = N^{1/2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)$. Noting that $\Delta y_{it} - x'_{it}\beta_t + x'_{i,t-1}\beta_{t-1} = \Delta u_{it} - N^{-1/2}(x'_{it}b_t - x'_{i,t-1}b_{t-1})$, we have

$$\begin{aligned}
& N\left[V_{1NT,\lambda_1}(\boldsymbol{\beta}) - V_{1NT,\lambda_1}(\boldsymbol{\beta}^0)\right] \\
&= \frac{1}{N}\sum_{i=1}^N\sum_{t=2}^T\left(x'_{it}b_t - x'_{i,t-1}b_{t-1}\right)^2 - \frac{2}{N^{1/2}}\sum_{i=1}^N\sum_{t=2}^T\Delta u_{it}\left(x'_{it}b_t - x'_{i,t-1}b_{t-1}\right) \\
&+ N\lambda_1\sum_{t=2}^T\dot{w}_t\left[\left\|\beta^0_t - \beta^0_{t-1} + N^{-1/2}(b_t - b_{t-1})\right\| - \left\|\beta^0_t - \beta^0_{t-1}\right\|\right] \\
&= \mathbf{b}'\dot{Q}_{NT}\mathbf{b} - 2\mathbf{b}'\sqrt{N}\dot{R}^u_{NT} + N\lambda_1\sum_{t\in\mathcal{T}^0_{m^0}}\dot{w}_t\left[\left\|\beta^0_t - \beta^0_{t-1} + N^{-1/2}(b_t - b_{t-1})\right\| - \left\|\beta^0_t - \beta^0_{t-1}\right\|\right] \\
&+ N\lambda_1\sum_{t\in\mathcal{T}^{0c}_{m^0}}\dot{w}_t\left\|N^{-1/2}(b_t - b_{t-1})\right\| \\
&\equiv A_1(\mathbf{b}) - 2A_2(\mathbf{b}) + A_3(\mathbf{b}) + A_4(\mathbf{b}),\ \text{say},
\end{aligned}$$

where $\dot{Q}_{NT}$ and $\dot{R}^u_{NT}$ are defined in (A.1) and (A.2), respectively. By Lemma B.1 and Assumption A.2(i), $\max_{t\in\mathcal{T}^0_{m^0}}\dot{w}_t = \max_{t\in\mathcal{T}^0_{m^0}}\left\|\dot{\beta}_t - \dot{\beta}_{t-1}\right\|^{-\kappa_1} = \max_{t\in\mathcal{T}^0_{m^0}}\left\|\beta^0_t - \beta^0_{t-1} + O_P\left(N^{-1/2}\right)\right\|^{-\kappa_1} = O_P\left(J^{-\kappa_1}_{\min}\right)$. By the Jensen, triangle and Cauchy-Schwarz inequalities, and Assumption A.2(ii),

$$\begin{aligned}
\left|T^{-1}A_3\left(\mathbf{b}\right)\right| &\leq m^0 T^{-1} N^{1/2}\lambda_1 \max_{s\in\mathcal{T}^0_{m^0}} \dot{w}_s \left\{\frac{1}{m^0}\sum_{t\in\mathcal{T}^0_{m^0}} \left\|b_t - b_{t-1}\right\|\right\} \\
&\leq m^0 T^{-1} N^{1/2}\lambda_1 \max_{s\in\mathcal{T}^0_{m^0}} \dot{w}_s \left\{\frac{1}{m^0}\sum_{t\in\mathcal{T}^0_{m^0}} \left\|b_t - b_{t-1}\right\|^2\right\}^{1/2} \\
&\leq 2\left(m^0\right)^{1/2} T^{-1/2} N^{1/2}\lambda_1 \max_{s\in\mathcal{T}^0_{m^0}} \dot{w}_s T^{-1/2}\left\|\mathbf{b}\right\| \\
&= O_P\left(\left(m^0 N\right)^{1/2}\lambda_1 T^{-1/2} J_{\min}^{-\kappa_1}\right) T^{-1/2}\left\|\mathbf{b}\right\| = O_P\left(1\right) T^{-1/2}\left\|\mathbf{b}\right\|. \qquad \text{(B.8)}
\end{aligned}$$

In conjunction with the analyses of $A_1\left(\mathbf{b}\right)$ and $A_2\left(\mathbf{b}\right)$ in the proof of Lemma B.1, this implies that w.p.a.1

$$T^{-1}\left[A_1\left(\mathbf{b}\right) - 2A_2\left(\mathbf{b}\right) + A_3\left(\mathbf{b}\right)\right] \geq \mu_{\min}\left(\dot{Q}_{NT}\right) T^{-1}\left\|\mathbf{b}\right\|^2 - O_P\left(1\right) T^{-1/2}\left\|\mathbf{b}\right\| > 0$$

if $T^{-1/2}\left\|\mathbf{b}\right\| = L$ is sufficiently large. That is, $A_1\left(\mathbf{b}\right)$ dominates $-2A_2\left(\mathbf{b}\right)+A_3\left(\mathbf{b}\right)$ for large $L$. In addition, $A_4\left(\mathbf{b}\right)\geq 0$. Consequently, $N\left[V_{1NT,\lambda_1}\left(\boldsymbol{\beta}\right) - V_{1NT,\lambda_1}\left(\boldsymbol{\beta}^0\right)\right] > 0$ w.p.a.1 for large $L$ and $V_{1NT,\lambda_1}\left(\boldsymbol{\beta}\right)$ cannot be minimized in this case. This further implies that $T^{-1/2}\left\|\tilde{\mathbf{b}}\right\|$ has to be stochastically bounded and Theorem 3.2 (i) holds.

(ii) Let $\dot{L}$, $\dot{\Lambda}$, $\mathbf{b}^{\dagger}$, $\dot{R}^{\dagger}_{NT}$, and $\left\{\omega_{ts}\right\}_{t,s=1}^{T}$ be as defined in the proof of Lemma B.1. Let $\omega^{\dagger}_{ts} = \omega_{ts} - \omega_{t-1,s}$. Then $b_t - b_{t-1} = \sum_{s=t}^{T}\omega_{ts}b^{\dagger}_s - \sum_{s=t-1}^{T}\omega_{t-1,s}b^{\dagger}_s = \sum_{s=t-1}^{T}\omega^{\dagger}_{ts}b^{\dagger}_s$ as $\omega_{ts} = 0$ for $s = t-1$. So we can rewrite $N\left[V_{1NT,\lambda_1}\left(\boldsymbol{\beta}\right) - V_{1NT,\lambda_1}\left(\boldsymbol{\beta}^0\right)\right]$ in terms of $\mathbf{b}^{\dagger}$ :

$$\begin{aligned}
&N\left[V_{1NT,\lambda_1}\left(\boldsymbol{\beta}\right) - V_{1NT,\lambda_1}\left(\boldsymbol{\beta}^0\right)\right] \\
&= \sum_{t=1}^{T}\left[b^{\dagger\prime}_t \dot{\Lambda}^{-1}_t b^{\dagger}_t - 2b^{\dagger\prime}_t \dot{R}^{\dagger}_{tNT}\right] + N\lambda_1 \sum_{t\in\mathcal{T}^0_{m^0}} \dot{w}_t \left[\left\|\beta^0_t - \beta^0_{t-1} + N^{-1/2}\sum_{s=t-1}^{T}\omega^{\dagger}_{ts}b^{\dagger}_s\right\| - \left\|\beta^0_t - \beta^0_{t-1}\right\|\right] \\
&+ N^{1/2}\lambda_1 \sum_{t\in\mathcal{T}^{0c}_{m^0}} \dot{w}_t \left\|\sum_{s=t-1}^{T}\omega^{\dagger}_{ts}b^{\dagger}_s\right\| \equiv N V^{\dagger}_{1NT,\lambda_1}\left(\mathbf{b}^{\dagger}\right), \text{ say.}
\end{aligned}$$

Let $\tilde{\mathbf{b}}^{\dagger} = (\dot{\Lambda}+\dot{L}^{\prime})\tilde{\mathbf{b}} = (\tilde{b}^{\dagger\prime}_1, ..., \tilde{b}^{\dagger\prime}_T)^{\prime}$. Noting that $m_0 N^{1/2}\lambda_1 \max_{s\in\mathcal{T}^0_{m^0}} \dot{w}_s = O_P\left(m_0 N^{1/2}\lambda_1 J_{\min}^{-\kappa_1}\right) = O_P\left(1\right)$, $\sum_{s\in\mathcal{T}^0_{m^0}}\left\|\omega^{\dagger}_{st}\right\| = O_P\left(m^0\right)$ and $\left\|\dot{R}^{\dagger}_{tNT}\right\| = O_P\left(1\right)$ for each $t$, we have by the triangle inequality

$$\begin{aligned}
0 \geq N V^{\dagger}_{1NT,\lambda_1}(\tilde{\mathbf{b}}^{\dagger}) &\geq \sum_{t=1}^{T}\left[\tilde{b}^{\dagger\prime}_t \dot{\Lambda}^{-1}_t \tilde{b}^{\dagger}_t - 2\tilde{b}^{\dagger\prime}_t \dot{R}^{\dagger}_{tNT}\right] - N^{1/2}\lambda_1 \max_{s\in\mathcal{T}^0_{m^0}} \dot{w}_s \sum_{s\in\mathcal{T}^0_{m^0}}\left\|\sum_{t=s-1}^{T}\omega^{\dagger}_{st}\tilde{b}^{\dagger}_t\right\| \\
&\geq \sum_{t=1}^{T}\left[\tilde{b}^{\dagger\prime}_t \dot{\Lambda}^{-1}_t \tilde{b}^{\dagger}_t - \left(2\left\|\dot{R}^{\dagger}_{tNT}\right\| + N^{1/2}\lambda_1 \max_{s\in\mathcal{T}^0_{m^0}} \dot{w}_s \sum_{s\in\mathcal{T}^0_{m^0}}\left\|\omega^{\dagger}_{st}\right\|\right)\left\|\tilde{b}^{\dagger}_t\right\|\right] \\
&= \sum_{t=1}^{T}\left[\tilde{b}^{\dagger\prime}_t \dot{\Lambda}^{-1}_t \tilde{b}^{\dagger}_t - O_P\left(1\right)\left\|\tilde{b}^{\dagger}_t\right\|\right].
\end{aligned}$$

It follows that $\tilde{b}_t^{\dagger} = O_P(1)$ for each $t$ by arguments as used in the proof of Lemma B.1. Otherwise, $\{\tilde{b}_t^{\dagger}\}$ cannot minimize $V_{1NT,\lambda_1}^{\dagger}(\mathbf{b}^{\dagger})$. This implies that $\tilde{b}_t = N^{1/2}(\tilde{\beta}_t - \beta_t^0) = O_P(1)$ by the same arguments as used in the proof of Lemma B.1. ■

**Proof of Theorem 3.3.** We want to demonstrate that

$$P\left(\left\|\tilde{\theta}_t\right\| = 0 \text{ for all } t \in \mathcal{T}_{m^0}^{0c}\right) \to 1 \text{ as } N \to \infty. \tag{B.9}$$

Suppose that to the contrary, $\tilde{\theta}_t = \tilde{\beta}_t - \tilde{\beta}_{t-1} \neq 0$ for some $t \in \mathcal{T}_{m^0}^{0c}$ for sufficiently large $N$. Then there exists $r \in \{1, ..., p\}$ such that $\left|\tilde{\theta}_{t,r}\right| = \max\left\{\left|\tilde{\theta}_{t,l}\right|,\ l = 1, ..., p\right\}$, where for any $p \times 1$ vector $a_t$, $a_{t,l}$ denotes its $l$th element. Without loss of generality (wlog) assume that $r = p$, implying that $\left|\tilde{\theta}_{t,p}\right| / \left\|\tilde{\theta}_t\right\| \geq 1/\sqrt{p}$. To consider the first order condition (FOC) with respect to (wrt) $\beta_t$, $t \geq 2$, based on subdifferential calculus (e.g., Bersekas 1995, Appendix B.5), we distinguish two cases: (a) $2 \leq t \leq T-1$ and (b) $t = T$ and $T \in \mathcal{T}_{m^0}^{0c}$.

In case (a), we consider two subcases: (a1) $t+1 = T_j^0 \in \mathcal{T}_{m^0}^{0}$ for some $j = 1, ..., m^0$, and (a2) $t+1 \in \mathcal{T}_{m^0}^{0c}$. In either case, we can apply the FOC wrt $\beta_{t,p}$ and the equality $\Delta y_{it} = \beta_t^{0\prime} x_{it} - \beta_{t-1}^{0\prime} x_{i,t-1} + \Delta u_{it}$ to obtain

$$\begin{aligned}
0 &= \frac{-2}{\sqrt{N}} \sum_{i=1}^{N} \left(\Delta y_{it} - \tilde{\beta}_t' x_{it} + \tilde{\beta}_{t-1}' x_{i,t-1}\right) x_{it,p} + \frac{2}{\sqrt{N}} \sum_{i=1}^{N} \left(\Delta y_{i,t+1} - \tilde{\beta}_{t+1}' x_{i,t+1} + \tilde{\beta}_t' x_{it}\right) x_{it,p} \\
&\quad + \sqrt{N}\lambda_1 \dot{w}_t \frac{\tilde{\theta}_{t,p}}{\left\|\tilde{\theta}_t\right\|} - \sqrt{N}\lambda_1 \dot{w}_{t+1} e_{t+1,p} \qquad\qquad (\text{B.10}) \\
&= -\frac{2}{\sqrt{N}} \sum_{i=1}^{N} \left[\left(\tilde{\beta}_{t+1} - \beta_{t+1}^0\right)' x_{i,t+1} - 2\left(\tilde{\beta}_t - \beta_t^0\right)' x_{it} + \left(\tilde{\beta}_{t-1} - \beta_{t-1}^0\right)' x_{i,t-1}\right] x_{it,p} \\
&\quad + \frac{2}{\sqrt{N}} \sum_{i=1}^{N} \Delta^2 u_{i,t+1} x_{it,p} + \sqrt{N}\lambda_1 \dot{w}_t \frac{\tilde{\theta}_{t,p}}{\left\|\tilde{\theta}_t\right\|} - \sqrt{N}\lambda_1 \dot{w}_{t+1} e_{t+1,p} \\
&\equiv B_{1t} + B_{2t} + B_{3t} - B_{4t}, \text{ say,}
\end{aligned}$$

where $e_{t+1} = \tilde{\theta}_{t+1} / \left\|\tilde{\theta}_{t+1}\right\|$ if $\left\|\tilde{\theta}_{t+1}\right\| \neq 0$ and $\|e_{t+1}\| \leq 1$ otherwise, $e_{t+1,p}$ is the $p$th element in $e_{t+1}$. By Assumptions A.1(i)-(ii) and Theorem 3.2, $B_{1t} = O_P(1)$ and $B_{2t} = O_P(1)$. In view of the fact that $\dot{w}_t^{-1} = O_P(N^{-\kappa_1/2})$ for $t \in \mathcal{T}_{m^0}^{0c}$, $|B_{3t}| \geq \sqrt{N}\lambda_1 \dot{w}_t / \sqrt{p}$, which is explosive in probability under Assumption A.2(iii) (i.e., $N^{(\kappa_1+1)/2}\lambda_1 \to \infty$).

To bound the probability order of $B_{4t}$, we distinguish two subcases. In subcase (a1), noting that $\dot{\beta}_{t+1} - \dot{\beta}_t \xrightarrow{P} \theta_{t+1}^0 \neq 0$ by Theorem 3.2, we have $\dot{w}_{t+1} = \left\|\theta_{t+1}^0 + O_P(N^{-1/2})\right\|^{-\kappa_1} = O_P\left(J_{\min}^{-\kappa_1}\right)$ and $B_{4t} = \sqrt{N}\lambda_1 \dot{w}_{t+1} e_{t+1,p} = O_P(\sqrt{N}\lambda_1 J_{\min}^{-\kappa_1}) = O_P(1)$. Consequently, $|B_{3t}| \gg |B_{1t} + B_{2t} - B_{4t}|$ so that (B.10) cannot be true for sufficiently large $N$ or $(N, T)$. Then we conclude that w.p.a.1, $\tilde{\theta}_t$ must be in a position where $\left\|\tilde{\theta}_t\right\|$ is not differentiable in subcase (a1). In addition, a direct implication of this result is that if $t = T_j^0 - 1 \in \mathcal{T}_{m^0}^{0c}$ for some $j = 1, ..., m^0$, then $P\left(\left\|\tilde{\theta}_{T_j^0 - 1}\right\| = 0\right) \to 1$ as $N \to \infty$ and $\sqrt{N}\lambda_1 \dot{w}_{T_j^0 - 1} e_{T_j^0 - 1} = O_P(1)$ in order for the FOC to hold for $t = T_j^0 - 1$.

In subcase (a2), difficulty arises as $\dot{w}_{t+1} = O_P(N^{\kappa_1/2})$ and $\sqrt{N}\lambda_1\dot{w}_{t+1} = O_P(N^{(1+\kappa_1)/2}\lambda_1)$. But we can apply the implication from the result in subcase (a1) recursively. When $t = T_j^0 - 2 \in \mathcal{T}_{m^0}^{0c}$ for some $j = 1, ..., m^0$, $B_{4t} = \sqrt{N}\lambda_1\dot{w}_{T_j^0-1}e_{T_j^0-1,p} = O_P(1)$ and $|B_{3t}| \gg |B_{1t} + B_{2t} - B_{4t}|$. Thus (B.10) cannot hold for $t = T_j^0 - 2 \in \mathcal{T}_{m^0}^{0c}$ either and we must have $P\left(\left\|\tilde{\theta}_{T_j^0-2}\right\| = 0\right) \to 1$ as $N \to \infty$ and $\sqrt{N}\lambda_1\dot{w}_{T_j^0-2}e_{T_j^0-2} = O_P(1)$ in order for the FOC to hold for $t = T_j^0 - 2$. Deducting in this way until we reach $t = T_{j-1}^0 + 1 \in \mathcal{T}_{m^0}^{0c}$. Consequently, $\tilde{\theta}_t$ must be in a position that $\left\|\tilde{\theta}_t\right\|$ is not differentiable for all $t \in \mathcal{T}_{m^0}^{0c}$ and $t \neq T$.

In case (b), noting that only one term in the penalty term ($\lambda_1 \sum_{t=2}^T \dot{w}_t \left\|\beta_t - \beta_{t-1}\right\|$) is involved with $\beta_T$, it is easy to show that $\tilde{\theta}_T = \tilde{\beta}_T - \tilde{\beta}_{T-1}$ must be in a position where $\left\|\tilde{\theta}_T\right\|$ is not differentiable if $T \in \mathcal{T}_{m^0}^{0c}$. Consequently (B.9) follows. ■

**Proof of Corollary 3.4.** We consider two cases: (a) $t \in \mathcal{T}_{m^0}^{0c}$, and (b) $t \in \mathcal{T}_{m^0}^{0}$. In case (a), Theorem 3.3 implies that asymptotically no time point in $\mathcal{T}_{m^0}^{0c}$ can be identified as an estimated break date so that $\tilde{m} \leq m^0$. In case (b), we want to show that all break points in $\mathcal{T}_{m^0}^{0}$ must be identified as an estimated break point. Suppose not. Then there exists $t \in \mathcal{T}_{m^0}^{0}$ such that $\left\|\tilde{\theta}_t\right\| = 0$. By the $\sqrt{N}$-consistency of $\tilde{\theta}_t$ and the fact $\tilde{\theta}_t = \tilde{\beta}_t - \tilde{\beta}_{t-1} = \beta_t^0 - \beta_{t-1}^0 + O_P(N^{-1/2}) = \theta_t^0 + O_P(N^{-1/2})$ by Theorem 3.2, we have $\left\|\theta_t^0\right\| = O(N^{-1/2})$, which contradicts the assumption that $N^{1/2}J_{\min} \to \infty$ as $N \to \infty$ as $\left\|\theta_t^0\right\| \geq J_{\min}$ for any $t \in \mathcal{T}_{m^0}^{0}$. ■

**Proof of Theorem 3.5.** (i) The FOCs wrt $\beta_t$, $t = 1, ..., T$, for the PLS problem are given by

$$\begin{aligned}
\mathbf{0}_{p\times 1} = & \frac{-2}{N}\sum_{i=1}^{N}\left(\Delta y_{it} - \tilde{\beta}_t' x_{it} + \tilde{\beta}_{t-1}' x_{i,t-1}\right) x_{it}\mathbf{1}\left\{t > 1\right\} \\
& + \frac{2}{N}\sum_{i=1}^{N}\left(\Delta y_{i,t+1} - \tilde{\beta}_{t+1}' x_{i,t+1} + \tilde{\beta}_t' x_{it}\right) x_{it}\mathbf{1}\left\{t < T\right\} \\
& + \lambda_1\left[\dot{w}_t e_t \mathbf{1}\left\{t > 1\right\} - \dot{w}_{t+1}e_{t+1}\mathbf{1}\left\{t < T\right\}\right],
\end{aligned} \tag{B.11}$$

where $e_t = \tilde{\theta}_t / \left\|\tilde{\theta}_t\right\|$ if $\left\|\tilde{\theta}_t\right\| \neq 0$ and $\|e_t\| \leq 1$ otherwise. Summing both sides of the above equation over $t$ for each of the $\tilde{m}+1$ estimated regimes and using the fact that $\tilde{\beta}_t = \tilde{\alpha}_j$ if $t$ belongs to the $j$th estimated regime yield

$$\begin{aligned}
\mathbf{0}_{p\times 1} &= \frac{-2}{N}\sum_{t=2}^{\tilde{T}_1-1}\sum_{i=1}^{N}\left(\Delta y_{it} - \tilde{\alpha}_1'\Delta x_{it}\right)\Delta x_{it} + \frac{2}{N}\sum_{i=1}^{N}\left(\Delta y_{i\tilde{T}_1} - \tilde{\alpha}_2' x_{i,\tilde{T}_1} + \tilde{\alpha}_1' x_{i,\tilde{T}_1-1}\right) x_{i,\tilde{T}_1-1} + \mathcal{R}_{NT,1}, \\
\mathbf{0}_{p\times 1} &= \frac{-2}{N}\sum_{t=\tilde{T}_{j-1}+1}^{\tilde{T}_j-1}\sum_{i=1}^{N}\left(\Delta y_{it} - \tilde{\alpha}_j'\Delta x_{it}\right)\Delta x_{it} + \frac{2}{N}\sum_{i=1}^{N}\left(\Delta y_{i\tilde{T}_j} - \tilde{\alpha}_{j+1}' x_{i\tilde{T}_j} + \tilde{\alpha}_j' x_{i,\tilde{T}_j-1}\right) x_{i,\tilde{T}_j-1} \\
&\quad - \frac{2}{N}\sum_{i=1}^{N}\left(\Delta y_{i\tilde{T}_{j-1}} - \tilde{\alpha}_j' x_{i\tilde{T}_{j-1}} + \tilde{\alpha}_{j-1}' x_{i,\tilde{T}_{j-1}-1}\right) x_{i\tilde{T}_{j-1}} + \mathcal{R}_{NT,j} \quad \text{for } j = 2, ..., \tilde{m}, \\
\mathbf{0}_{p\times 1} &= \frac{-2}{N}\sum_{t=\tilde{T}_{\tilde{m}}+1}^{T}\sum_{i=1}^{N}\left(\Delta y_{it} - \tilde{\alpha}_{\tilde{m}+1}'\Delta x_{it}\right)\Delta x_{it} - \frac{2}{N}\sum_{i=1}^{N}\left(\Delta y_{i\tilde{T}_{\tilde{m}}} - \tilde{\alpha}_{\tilde{m}+1}' x_{i\tilde{T}_{\tilde{m}}} + \tilde{\alpha}_{\tilde{m}}' x_{i,\tilde{T}_{\tilde{m}}-1}\right) x_{i\tilde{T}_{\tilde{m}}} + \mathcal{R}_{NT,\tilde{m}+1},
\end{aligned}$$

where $\mathcal{R}_{NT,1} = \mathcal{R}_{NT,1}(\tilde{\mathcal{T}}_{\tilde{m}}) = -\lambda_1 \dot{w}_{\tilde{T}_1} e_{\tilde{T}_1}$, $\mathcal{R}_{NT,j} = \mathcal{R}_{NT,j}(\tilde{\mathcal{T}}_{\tilde{m}}) = \lambda_1(\dot{w}_{\tilde{T}_{j-1}} e_{\tilde{T}_{j-1}} - \dot{w}_{\tilde{T}_j} e_{\tilde{T}_j})$ for $j = 2, ..., \tilde{m}$, $\mathcal{R}_{NT,\tilde{m}+1} = \mathcal{R}_{NT,\tilde{m}+1}(\tilde{\mathcal{T}}_{\tilde{m}}) = \lambda_1 \dot{w}_{\tilde{T}_{\tilde{m}}} e_{\tilde{T}_{\tilde{m}}}$, and we have suppressed the dependence of $\tilde{\alpha}_j$ on $\tilde{\mathcal{T}}_{\tilde{m}}$. Let $\mathcal{R}_{NT}(\tilde{\mathcal{T}}_{\tilde{m}}) = \left(\mathcal{R}'_{NT,1}(\tilde{\mathcal{T}}_{\tilde{m}}), ..., \mathcal{R}'_{NT,\tilde{m}+1}(\tilde{\mathcal{T}}_{\tilde{m}})\right)'$. One can readily solve for $\tilde{\boldsymbol{\alpha}}_{\tilde{m}} = \tilde{\boldsymbol{\alpha}}_{\tilde{m}}(\tilde{\mathcal{T}}_{\tilde{m}})$ to obtain

$$\tilde{\boldsymbol{\alpha}}_{\tilde{m}}\left(\tilde{\mathcal{T}}_{\tilde{m}}\right) = \Phi_{NT}\left(\tilde{\mathcal{T}}_{\tilde{m}}\right)^{-1}\left[\Psi^y_{NT}\left(\tilde{\mathcal{T}}_{\tilde{m}}\right) - \frac{1}{2}\mathcal{R}_{NT}\left(\tilde{\mathcal{T}}_{\tilde{m}}\right)\right],$$

where $\Phi_{NT}(\cdot)$ and $\Psi^y_{NT}(\cdot)$ are defined in (A.3) and (A.4) in Appendix A.1, respectively.

By Corollary 3.4, $\tilde{\boldsymbol{\alpha}}_{\tilde{m}}(\tilde{\mathcal{T}}_{\tilde{m}}) = \tilde{\boldsymbol{\alpha}}_{m^0}\left(\mathcal{T}^0_{m^0}\right)$ w.p.a.1. Therefore we can study the asymptotic distribution of $\tilde{\boldsymbol{\alpha}}_{\tilde{m}}(\tilde{\mathcal{T}}_{\tilde{m}})$ by studying that of $\tilde{\boldsymbol{\alpha}}_{m^0}\left(\mathcal{T}^0_{m^0}\right)$. Note that $\tilde{\boldsymbol{\alpha}}_{m^0}\left(\mathcal{T}^0_{m^0}\right) = \Phi^{-1}_{NT}[\Psi^y_{NT} - \frac{1}{2}\mathcal{R}_{NT}(\mathcal{T}^0_{m^0})]$, where $\Phi_{NT} = \Phi_{NT}\left(\mathcal{T}^0_{m^0}\right)$ and $\Psi^y_{NT} = \Psi^y_{NT}(\mathcal{T}^0_{m^0})$ (see (3.1)). It is easy to verify that

$$\begin{aligned}\sqrt{N}S\mathbb{D}_{m^0+1}\left(\tilde{\boldsymbol{\alpha}}_{m^0}\left(\mathcal{T}_{m^0}\right) - \boldsymbol{\alpha}^0\right) &= \sqrt{N}S\mathbb{D}_{m^0+1}\left(\tilde{\boldsymbol{\alpha}}^p_{m^0}\left(\mathcal{T}^0_{m^0}\right) - \boldsymbol{\alpha}^0\right)\\ &\quad -\frac{1}{2}S\left(\mathbb{D}^{-1}_{m^0+1}\Phi_{NT}\mathbb{D}^{-1}_{m^0+1}\right)^{-1}\sqrt{N}\mathbb{D}^{-1}_{m^0+1}\mathcal{R}_{NT}\left(\mathcal{T}^0_{m^0}\right).\end{aligned}$$

By the proof of (ii) below, $\sqrt{N}S\mathbb{D}_{m^0+1}\left(\tilde{\boldsymbol{\alpha}}^p_{m^0}\left(\mathcal{T}^0_{m^0}\right) - \boldsymbol{\alpha}^0\right) \xrightarrow{D} N\left(0, S\Phi_0^{-1}\Omega_0\Phi_0^{-1}S'\right)$. By the fact that $\|A - B\|^2 \leq 2\{\|A\|^2 + \|B\|^2\}$, $\|e_t\| \leq 1$, and $\max_{t\in\mathcal{T}^0_{m^0}} \dot{w}_t = O_P\left(J^{-\kappa_1}_{\min}\right)$ under Assumption A.2(i), we have

$$\begin{aligned}& N\left\|\mathbb{D}^{-1}_{m^0+1}\mathcal{R}_{NT}\left(\mathcal{T}^0_{m^0}\right)\right\|^2\\ = \;& N\lambda_1^2\left\{\left(I^0_1\right)^{-1}\left\|\dot{w}_{T^0_1}e_{T^0_1}\right\|^2 + \sum_{j=2}^{m^0}\left(I^0_j\right)^{-1}\left\|\dot{w}_{T^0_{j-1}}e_{T^0_{j-1}} - \dot{w}_{T^0_j}e_{T^0_j}\right\|^2 + \left(I^0_{m^0+1}\right)^{-1}\left\|\dot{w}_{T^0_{m^0}}e_{T^0_{m^0}}\right\|^2\right\}\\ \leq \;& 4\left(m^0 + 1\right)N\lambda_1^2 I^{-1}_{\min}\max_{t\in\mathcal{T}^0_{m^0}}\|\dot{w}_t\|^2 = O_P\left(m^0 N\lambda_1^2 I^{-1}_{\min}J^{-2\kappa_1}_{\min}\right) = o_P(1) \text{ under Assumption A.2(ii).}\end{aligned}$$

With this, we can show that $\left\|S(\mathbb{D}^{-1}_{m^0+1}\Phi_{NT}\mathbb{D}^{-1}_{m^0+1})^{-1}\sqrt{N}\mathbb{D}^{-1}_{m^0+1}\mathcal{R}_{NT}\left(\mathcal{T}^0_{m^0}\right)\right\|^2 \leq \left[\mu_{\min}(\mathbb{D}^{-1}_{m^0+1}\Phi_{NT}\mathbb{D}^{-1}_{m^0+1})\right]^{-2}$ $\times\|S\|^2 N\left\|\mathbb{D}^{-1}_{m^0+1}\mathcal{R}_{NT}\left(\mathcal{T}^0_{m^0}\right)\right\|^2 = O_P(1)\,O(1)\,o_P(1) = o_P(1)$. Consequently, $\sqrt{N}S\mathbb{D}_{m^0+1}\left(\tilde{\boldsymbol{\alpha}}_{m^0}\left(\mathcal{T}_{m^0}\right) - \boldsymbol{\alpha}^0\right)$ $\xrightarrow{D} N\left(0, S\Phi_0^{-1}\Omega_0\Phi_0^{-1}S'\right)$.

(ii) Note that $\tilde{\boldsymbol{\alpha}}^p_{\tilde{m}}(\tilde{\mathcal{T}}_{\tilde{m}}) = (\tilde{\alpha}^p_1(\tilde{\mathcal{T}}_{\tilde{m}})', ..., \tilde{\alpha}^p_{\tilde{m}+1}(\tilde{\mathcal{T}}_{\tilde{m}})')' = \arg\min_{\boldsymbol{\alpha}_{\tilde{m}}} Q_{1NT}\left(\boldsymbol{\alpha}_{\tilde{m}}; \tilde{\mathcal{T}}_{\tilde{m}}\right)$. The FOCs for this minimization problem are

$$\begin{aligned}\mathbf{0}_{p\times 1} &= \frac{-2}{N}\sum_{t=2}^{\tilde{T}_1-1}\sum_{i=1}^{N}\left(\Delta y_{it} - \tilde{\alpha}^{p\prime}_1\Delta x_{it}\right)\Delta x_{it} + \frac{2}{N}\sum_{i=1}^{N}\left(\Delta y_{i\tilde{T}_1} - \tilde{\alpha}^{p\prime}_2 x_{i,\tilde{T}_1} + \tilde{\alpha}^{p\prime}_1 x_{i,\tilde{T}_1-1}\right)x_{i,\tilde{T}_1-1},\\ \mathbf{0}_{p\times 1} &= \frac{-2}{N}\sum_{t=\tilde{T}_{j-1}+1}^{\tilde{T}_j-1}\sum_{i=1}^{N}\left(\Delta y_{it} - \tilde{\alpha}^{p\prime}_j\Delta x_{it}\right)\Delta x_{it} + \frac{2}{N}\sum_{i=1}^{N}\left(\Delta y_{i\tilde{T}_j} - \tilde{\alpha}^{p\prime}_{j+1}x_{i\tilde{T}_j} + \tilde{\alpha}^{p\prime}_j x_{i,\tilde{T}_j-1}\right)x_{i,\tilde{T}_j-1}\\ &\quad -\frac{2}{N}\sum_{i=1}^{N}\left(\Delta y_{i\tilde{T}_{j-1}} - \tilde{\alpha}^{p\prime}_j x_{i\tilde{T}_{j-1}} + \tilde{\alpha}^{p\prime}_{j-1}x_{i,\tilde{T}_{j-1}-1}\right)x_{i\tilde{T}_{j-1}} \text{ for } j = 2, ..., \tilde{m}, \text{ and}\\ \mathbf{0}_{p\times 1} &= \frac{-2}{N}\sum_{t=\tilde{T}_{\tilde{m}}+1}^{T}\sum_{i=1}^{N}\left(\Delta y_{it} - \tilde{\alpha}^{p\prime}_{\tilde{m}+1}\Delta x_{it}\right)\Delta x_{it} - \frac{2}{N}\sum_{i=1}^{N}\left(\Delta y_{i\tilde{T}_{\tilde{m}}} - \tilde{\alpha}^{p\prime}_{\tilde{m}+1}x_{i\tilde{T}_{\tilde{m}}} + \tilde{\alpha}^{p\prime}_{\tilde{m}}x_{i,\tilde{T}_{\tilde{m}}-1}\right)x_{i\tilde{T}_{\tilde{m}}},\end{aligned}$$

where we suppress the dependence of $\tilde{\alpha}_j^p$'s on $\tilde{\mathcal{T}}_{\tilde{m}}$. One can readily solve for $\tilde{\boldsymbol{\alpha}}_{\tilde{m}}^p = \tilde{\boldsymbol{\alpha}}_{\tilde{m}}^p(\tilde{\mathcal{T}}_{\tilde{m}})$ to obtain $\tilde{\boldsymbol{\alpha}}_{\tilde{m}}^p(\tilde{\mathcal{T}}_{\tilde{m}}) = \Phi_{NT}\left(\tilde{\mathcal{T}}_{\tilde{m}}\right)^{-1}\Psi_{NT}^y\left(\tilde{\mathcal{T}}_{\tilde{m}}\right)$.

By Corollary 3.4, $\tilde{\boldsymbol{\alpha}}_{\tilde{m}}^p(\tilde{\mathcal{T}}_{\tilde{m}}) = \tilde{\boldsymbol{\alpha}}_{m^0}^p\left(\mathcal{T}_{m^0}^0\right)$ w.p.a.1. Therefore we can study the asymptotic distribution of $\tilde{\boldsymbol{\alpha}}_{\tilde{m}}(\tilde{\mathcal{T}}_{\tilde{m}})$ by studying that of $\tilde{\boldsymbol{\alpha}}_{m^0}\left(\mathcal{T}_{m^0}^0\right)$. Using $\tilde{\boldsymbol{\alpha}}_{m^0}^p\left(\mathcal{T}_{m^0}^0\right) = \Phi_{NT}^{-1}\Psi_{NT}^y$, it is easy to verify that

$$\begin{aligned}
\sqrt{N}S\mathbb{D}_{m^0+1}\left(\tilde{\boldsymbol{\alpha}}_{m^0}^p\left(\mathcal{T}_{m^0}^0\right)-\boldsymbol{\alpha}^0\right) &= S\left(\mathbb{D}_{m^0+1}^{-1}\Phi_{NT}\mathbb{D}_{m^0+1}^{-1}\right)^{-1}\sqrt{N}\mathbb{D}_{m^0+1}^{-1}\Psi_{NT}^u \\
&= S\Phi_0^{-1}V_{NT} + S\left(\bar{\Phi}_{NT}^{-1}-\Phi_0^{-1}\right)V_{NT},
\end{aligned}$$

where $\Psi_{NT}^u$ is defined in (3.1), $V_{NT} = \sqrt{N}\mathbb{D}_{m^0+1}^{-1}\Psi_{NT}^u$, and $\bar{\Phi}_{NT} = \mathbb{D}_{m^0+1}^{-1}\Phi_{NT}\mathbb{D}_{m^0+1}^{-1}$. By Assumption A.3(ii), $S\Phi_0^{-1}V_{NT} \xrightarrow{D} N\left(0, S\Phi_0^{-1}\Omega_0\Phi_0^{-1}S'\right)$, implying that $\left\|S\Phi_0^{-1}V_{NT}\right\| = O_P(1)$. Assumption A.3(i) implies that $0 < \frac{1}{2}\mu_{\min}(\Phi_0) \le \mu_{\min}\left(\bar{\Phi}_{NT}\right) \le \mu_{\max}\left(\bar{\Phi}_{NT}\right) \le 2\lambda_{\max}(\Phi_0) < \infty$ w.p.a.1. Then by the fact that $\mu_{\min}(A)\text{tr}(B) \le\text{tr}(AB) \le \mu_{\max}(A)\text{tr}(B)$ for any symmetric matrix $A$ and conformable p.s.d. matrix $B$ (see, e.g., Proposition 8.4.13 in Bernstein 2005, p.275) and that $\max(\left|\mu_{\min}(A)\right|, \mu_{\max}(A)) = \|A\|_{\text{sp}}$ for any symmetric matrix $A$ (see, e.g., Fact 5.10.3 in Bernstein 2005, p.194), we have

$$\begin{aligned}
\left\|S\left(\bar{\Phi}_{NT}^{-1}-\Phi_0^{-1}\right)V_{NT}\right\|^2 &= \left\|S\bar{\Phi}_{NT}^{-1}\left(\bar{\Phi}_{NT}-\Phi_0\right)\Phi_0^{-1}V_{NT}\right\|^2 \\
&= \text{tr}\left\{\bar{\Phi}_{NT}^{-1}\left(\bar{\Phi}_{NT}-\Phi_0\right)\Phi_0^{-1}V_{NT}V_{NT}'\Phi_0^{-1}\left(\bar{\Phi}_{NT}-\Phi_0\right)\bar{\Phi}_{NT}^{-1}S'S\right\} \\
&\le \left[\mu_{\min}\left(\bar{\Phi}_{NT}\right)\right]^{-1}\text{tr}\left\{\left(\bar{\Phi}_{NT}-\Phi_0\right)\Phi_0^{-1}V_{NT}V_{NT}'\Phi_0^{-1}\left(\bar{\Phi}_{NT}-\Phi_0\right)\bar{\Phi}_{NT}^{-1}S'S\right\} \\
&= \left[\mu_{\min}\left(\bar{\Phi}_{NT}\right)\right]^{-1}\text{tr}\left\{\bar{\Phi}_{NT}^{-1}S'S\bar{\Phi}_{NT}\Phi_0^{-1}V_{NT}V_{NT}'\Phi_0^{-1}\left(\bar{\Phi}_{NT}-\Phi_0\right)\right\} \\
&\quad -\left[\mu_{\min}\left(\bar{\Phi}_{NT}\right)\right]^{-1}\text{tr}\left\{\bar{\Phi}_{NT}^{-1}S'S\Phi_0\Phi_0^{-1}V_{NT}V_{NT}'\Phi_0^{-1}\left(\bar{\Phi}_{NT}-\Phi_0\right)\right\},
\end{aligned}$$

where

$$\begin{aligned}
&\text{tr}\left\{\bar{\Phi}_{NT}^{-1}S'S\bar{\Phi}_{NT}\Phi_0^{-1}V_{NT}V_{NT}'\Phi_0^{-1}\left(\bar{\Phi}_{NT}-\Phi_0\right)\right\} \\
\le\ & \mu_{\max}\left(\bar{\Phi}_{NT}-\Phi_0\right)\text{tr}\left\{\bar{\Phi}_{NT}^{-1}S'S\bar{\Phi}_{NT}\Phi_0^{-1}V_{NT}V_{NT}'\Phi_0^{-1}\right\} \\
\le\ & \left\|\bar{\Phi}_{NT}-\Phi_0\right\|_{\text{sp}}\left[\mu_{\min}\left(\bar{\Phi}_{NT}\right)\right]^{-1}\text{tr}\left\{\bar{\Phi}_{NT}\Phi_0^{-1}V_{NT}V_{NT}'\Phi_0^{-1}S'S\right\} \\
\le\ & \left\|\bar{\Phi}_{NT}-\Phi_0\right\|_{\text{sp}}\left[\mu_{\min}\left(\bar{\Phi}_{NT}\right)\right]^{-1}\mu_{\max}\left(\bar{\Phi}_{NT}\right)\left\|S\Phi_0^{-1}V_{NT}\right\|^2 \\
=\ & o_P(1)\,O_P(1)\,O_P(1)\,O_P(1) = o_P(1)
\end{aligned}$$

and similarly

$$\begin{aligned}
&-\text{tr}\left\{\bar{\Phi}_{NT}^{-1}S'S\Phi_0\Phi_0^{-1}V_{NT}V_{NT}'\Phi_0^{-1}\left(\bar{\Phi}_{NT}-\Phi_0\right)\right\} \\
\le\ & -\mu_{\min}\left(\bar{\Phi}_{NT}-\Phi_0\right)\text{tr}\left\{\bar{\Phi}_{NT}^{-1}S'S\Phi_0\Phi_0^{-1}V_{NT}V_{NT}'\Phi_0^{-1}\right\} \\
\le\ & \left\|\bar{\Phi}_{NT}-\Phi_0\right\|_{\text{sp}}\left[\mu_{\min}\left(\bar{\Phi}_{NT}\right)\right]^{-1}\text{tr}\left\{\Phi_0\Phi_0^{-1}V_{NT}V_{NT}'\Phi_0^{-1}S'S\right\} \\
\le\ & \left\|\bar{\Phi}_{NT}-\Phi_0\right\|_{\text{sp}}\left[\mu_{\min}\left(\bar{\Phi}_{NT}\right)\right]^{-1}\mu_{\max}(\Phi_0)\left\|S\Phi_0^{-1}V_{NT}\right\|^2 \\
=\ & o_P(1)\,O_P(1)\,O_P(1)\,O_P(1) = o_P(1).
\end{aligned}$$

Hence $\left\|S\left(\bar{\Phi}_{NT}^{-1}-\Phi_0^{-1}\right)V_{NT}\right\| = o_P(1)$ and $\sqrt{N}S\mathbb{D}_{m^0+1}\left(\tilde{\boldsymbol{\alpha}}_{m^0}^p\left(\mathcal{T}_{m^0}\right)-\boldsymbol{\alpha}^0\right) \xrightarrow{D} N\left(0, S\Phi_0^{-1}\Omega_0\Phi_0^{-1}S'\right)$. ■

**Proof of Theorem 3.6.** Recall $\tilde{\boldsymbol{\alpha}}_{\tilde{m}_{\lambda_1}}^p(\tilde{\mathcal{T}}_{\tilde{m}_{\lambda_1}}) = (\tilde{\alpha}_1^p(\tilde{\mathcal{T}}_{\tilde{m}_{\lambda_1}})', ..., \tilde{\alpha}_{\tilde{m}_\lambda+1}^p(\tilde{\mathcal{T}}_{\tilde{m}_{\lambda_1}})')'$ denotes the set of post-Lasso OLS estimates of the regression coefficients based on the break dates in $\tilde{\mathcal{T}}_{\tilde{m}_{\lambda_1}} = \{\tilde{T}_1(\lambda_1), ..., \tilde{T}_{\tilde{m}_{\lambda_1}}(\lambda_1)\}$, where we make the dependence of various estimates on $\lambda_1$ explicit. Let $\tilde{\sigma}^2_{\tilde{\mathcal{T}}_{\tilde{m}_{\lambda_1}}} \equiv \frac{1}{T-1} Q_{1NT}(\tilde{\boldsymbol{\alpha}}^p_{\tilde{m}_{\lambda_1}}(\tilde{\mathcal{T}}_{\tilde{m}_{\lambda_1}}); \tilde{\mathcal{T}}_{\tilde{m}_{\lambda_1}})$. For any $\lambda^0_{1NT} \in \Omega_0$, we have $\lim_{N\to\infty} P(\tilde{m}_{\lambda^0_{1NT}} = m^0) = 1$ and $\lim_{N\to\infty} P(\tilde{T}_j \left(\lambda^0_{1NT}\right) = T^0_j,$ $j = 1, ..., m^0) = 1$ by Corollary 3.4 as $\lambda^0_{1NT}$ also satisfies Assumptions A.2(ii)-(iii). It follows that w.p.a.1 $\tilde{\sigma}^2_{\tilde{\mathcal{T}}_{\tilde{m}_{\lambda_1}}} = \tilde{\sigma}^2_{\mathcal{T}_{m^0}}$. Using the $\sqrt{NI^0_j}$-consistency of $\tilde{\alpha}^p_j\left(\mathcal{T}_{m^0}\right)$ and the expression $\Delta y_{it} = \alpha^{0\prime}_j \Delta x_{it} + \Delta u_{it}$ if $t \in [T^0_{j-1}+1, T^0_j - 1]$ and $\Delta y_{it} = \alpha^{0\prime}_{j+1} x_{it} - \alpha^{0\prime}_j x_{i,t-1} + \Delta u_{it}$ if $t = T^0_j$, we can readily show that

$$\begin{aligned}
\tilde{\sigma}^2_{\mathcal{T}_{m^0}} &= \frac{1}{N(T-1)} \sum_{j=1}^{m^0+1} \sum_{t=T^0_{j-1}+1}^{T^0_j - 1} \sum_{i=1}^{N} \left(\Delta y_{it} - \tilde{\alpha}'_{j,\mathcal{T}_{m^0}} \Delta x_{it}\right)^2 \\
&+ \frac{1}{N(T-1)} \sum_{j=1}^{m^0} \sum_{i=1}^{N} \left(\Delta y_{iT^0_j} - \tilde{\alpha}'_{j+1,\mathcal{T}_{m^0}} x_{iT^0_j} + \tilde{\alpha}'_{j,\mathcal{T}_{m^0}} x_{i,T^0_j - 1}\right)^2 \\
&= \bar{\sigma}^2_{NT} + O_P[(NI_{\min})^{-1}],
\end{aligned}$$

where $\bar{\sigma}^2_{NT} \equiv \frac{1}{N(T-1)} \sum_{t=2}^{T} \sum_{i=1}^{N} \Delta u^2_{it} \xrightarrow{P} \sigma^2_0 \equiv \lim_{(N,T)\to\infty} \frac{1}{N(T-1)} \sum_{t=2}^{T} \sum_{i=1}^{N} E\left(\Delta u^2_{it}\right)$ under Assumptions A.1(i)-(ii). Then by Assumption A.5 and Slutsky lemma, $IC_1\left(\lambda^0_{1NT}\right) = \tilde{\sigma}^2_{\mathcal{T}_{m^0}} + \rho_{1NT} p \left(m^0 + 1\right) \xrightarrow{P} \sigma^2_0$. We consider the case of under- and over-fitted models separately.

*Case 1: Under-fitted model:* $\tilde{m}_{\lambda_1} < m^0$. By Lemma B.2, $\inf_{\lambda_1 \in \Omega_-} \tilde{\sigma}^2_{\tilde{\mathcal{T}}_{\tilde{m}_{\lambda_1}}} - \tilde{\sigma}^2_{\mathcal{T}^0_{m^0}} \geq c_0$ where $c_0 = \frac{I_{\min} J^2_{\min}}{T-1}\left[c + o_P(1)\right]$ for some $c > 0$. Then by Assumption A5,

$$\begin{aligned}
P\left(\inf_{\lambda_1 \in \Omega_-} IC_1(\lambda_1) > IC_1\left(\lambda^0_{1NT}\right)\right) &= P\left(\inf_{\lambda_1 \in \Omega_-} \left[\left(\tilde{\sigma}^2_{\tilde{\mathcal{T}}_{\tilde{m}_{\lambda_1}}} - \tilde{\sigma}^2_{\mathcal{T}_{m^0}}\right) + \rho_{1NT} p \left(\tilde{m}_{\lambda_1} - m^0\right)\right] > 0\right) \\
&\geq P\left(\frac{I_{\min} J^2_{\min}}{\rho_{1NT}(T-1)}\left[c + o_P(1)\right] + O_P(1) > 0\right) \to 1.
\end{aligned}$$

*Case 2: Over-fitted model:* $\tilde{m}_{\lambda_1} > m^0$. Let $\mathbb{T}_m \equiv \{\mathcal{T}_m = \{T_1, ..., T_m\} : 2 \leq T_1 < ... < T_m \leq T\}$. Given $\mathcal{T}_m = \{T_1, ..., T_m\} \in \mathbb{T}_m$, let $\bar{\mathcal{T}}_{m^*+m^0} = \{\bar{T}_1, \bar{T}_2, ..., \bar{T}_{m^*+m^0}\}$ denote the union of $\mathcal{T}_m$ and $\mathcal{T}^0_{m^0}$ with elements ordered in non-descending order: $2 \leq \bar{T}_1 < \bar{T}_2 < \cdots < \bar{T}_{m^*+m^0} \leq T$ for some $m^* \in \{0, 1, ..., m\}$. Let $\tilde{\boldsymbol{\alpha}}^p_m(\mathcal{T}_m) \equiv \left(\tilde{\alpha}^p_1(\mathcal{T}_m)', ..., \tilde{\alpha}^p_{m+1}(\mathcal{T}_m)'\right)' = \arg\min_{\boldsymbol{\alpha}_m} Q_{1NT}\left(\boldsymbol{\alpha}_m; \mathcal{T}_m\right)$ and $\tilde{\sigma}^2_{\mathcal{T}_m} \equiv Q_{1NT}\left(\tilde{\boldsymbol{\alpha}}^p_m(\mathcal{T}_m); \mathcal{T}_m\right)$. $\tilde{\sigma}^2_{\bar{\mathcal{T}}_{m^*+m^0}}$ is analogously defined. In view of the fact that $\tilde{\sigma}^2_{\bar{\mathcal{T}}_{m^*+m^0}} \leq \tilde{\sigma}^2_{\mathcal{T}_m}$ for all $\mathcal{T}_m \in \mathbb{T}_m$, $N(\tilde{\sigma}^2_{\bar{\mathcal{T}}_{m^*+m^0}} - \bar{\sigma}^2_{NT}) = O_P(1)$ uniformly in $\mathcal{T}_m \in \mathbb{T}_m$ by Lemma B.3, and $N\rho_{1NT} \to \infty$ by Assumption A.5, we have

$$\begin{aligned}
&P\left(\inf_{\lambda_1 \in \Omega_+} IC_1(\lambda_1) > IC_1\left(\lambda^0_{1NT}\right)\right) \\
&\geq P\left(\min_{m^0 < m \leq m_{\max}} \inf_{\mathcal{T}_m \in \mathbb{T}_m} \left[N\left(\tilde{\sigma}^2_{\mathcal{T}_m} - \tilde{\sigma}^2_{\mathcal{T}_{m^0}}\right) + N\rho_{1NT} p \left(m - m^0\right)\right] > 0\right) \\
&\geq P\left(\min_{m^0 < m \leq m_{\max}} \inf_{\mathcal{T}_m \in \mathbb{T}_m} \left[N\left(\tilde{\sigma}^2_{\bar{\mathcal{T}}_{m^*+m^0}} - \tilde{\sigma}^2_{\mathcal{T}_{m^0}}\right) + N\rho_{1NT} p \left(m - m^0\right)\right] > 0\right) \\
&\to 1 \text{ as } N \to \infty. \blacksquare
\end{aligned}$$

# C Proof of the results in Section 4

**Proof of Lemma 4.1.** Recall that $\ddot{Q}_{NT}$ =TriD$(\dot{Q},\ddot{Q})_T$ by (A.6), $\ddot{Q}_{zx,t,s}=\phi'_{zx,t,s}W_t\phi_{zx,t,s}$, $\ddot{Q}_{zx,t}=\ddot{Q}_{zx,t,t}$ for $t,s=1,2,...,T$, and $\dot{Q}_{zx,t,t-1}=\phi'_{zx,t}W_t\phi_{zx,t,t-1}$ for $t=2,...,T$. Define

$$\begin{aligned}
\bar{\Lambda}_1 &= \ddot{Q}_{zx,2,1},\\
\bar{\Lambda}_t &= \left(\ddot{Q}_{zx,t}+\ddot{Q}_{zx,t+1,t}\right)-\dot{Q}_{zx,t,t-1}\bar{\Lambda}_{t-1}^{-1}\dot{Q}'_{zx,t,t-1} \text{ for } t=2,...,T-1,\\
\bar{\Lambda}_T &= \ddot{Q}_{zx,T}-\dot{Q}_{zx,T,T-1}\bar{\Lambda}_{T-1}^{-1}\dot{Q}'_{zx,T,T-1}.
\end{aligned} \tag{C.1}$$

We first argue that the above notations are well defined under Assumptions B.1(iii)-(iv) and B.2(iv) and that

$$0<\frac{1}{2}\min(\underline{c}_{zx}\underline{c}_w,\underline{c}_{z\eta})\leq\min_{1\leq t\leq T}\mu_{\min}\left(\bar{\Lambda}_t\right)\leq\max_{1\leq t\leq T}\mu_{\max}\left(\bar{\Lambda}_t\right)\leq 4\bar{c}_{zx}\bar{c}_w<\infty. \tag{C.2}$$

By Assumption B.1(iii) and B.2(iv), we can readily show that $\bar{\Lambda}_1$ is p.d. w.p.a.1. To study the behavior of $\bar{\Lambda}_t$ for $t=2,...,T$, we consider the auxiliary GMM estimation of the model

$$x_{it}=\alpha_t x_{i,t-1}+\eta_{it},\ i=1,...,N, \tag{C.3}$$

by using $z_{it}$ as the IV for $x_{i,t-1}$ and $W_t$ as the weighting function. The GMM estimator of $\alpha_t$ is given by

$$\hat{\alpha}_t=\left(\phi'_{zx,t,t}W_t\phi_{zx,t,t-1}\right)\left(\phi'_{zx,t,t-1}W_t\phi_{zx,t,t-1}\right)^{-1}=\dot{Q}_{zx,t,t-1}\ddot{Q}_{zx,t,t-1}^{-1}. \tag{C.4}$$

Let $\hat{\eta}_{it}=x_{it}-\hat{\alpha}_t x_{i,t-1}$, $\hat{\eta}_t=(\hat{\eta}_{1t},...,\hat{\eta}_{Nt})'$, $x_t=(x_{1t},...,x_{Nt})'$ and $z_t=(z_{1t},...,z_{Nt})'$. The first order conditions for the above GMM estimation imply that

$$\phi'_{z\hat{\eta},t,t}W_t\phi_{zx,t,t-1}=\frac{1}{N}\hat{\eta}'_t z_t W_t\frac{1}{N}z'_t x_{t-1}=0 \tag{C.5}$$

where $\phi_{z\hat{\eta},t,t}=\frac{1}{N}\sum_{i=1}^N z_{it}\hat{\eta}'_{it}$. Using (C.5), (C.4), and the equality $x_t=x_{t-1}\hat{\alpha}'_t+\hat{\eta}_t$, we can readily show that

$$\begin{aligned}
\ddot{Q}_{zx,t} &= \frac{1}{N^2}x'_t z_t W_t z'_t x_t=\frac{1}{N^2}\left(\hat{\alpha}_t x'_{t-1}+\hat{\eta}'_t\right)z_t W_t z'_t\left(x_{t-1}\hat{\alpha}'_t+\hat{\eta}_t\right)\\
&= \frac{1}{N^2}\left(\hat{\alpha}_t x'_{t-1}z_t W_t z'_t x_{t-1}\hat{\alpha}'_t+\hat{\eta}'_t z_t W_t z'_t\hat{\eta}_t+2\hat{\eta}'_t z_t W_t z'_t x_{t-1}\hat{\alpha}'_t\right)\\
&= \dot{Q}_{zx,t,t-1}\ddot{Q}_{zx,t,t-1}^{-1}\dot{Q}'_{zx,t,t-1}+\phi'_{z\hat{\eta},t,t}W_t\phi'_{z\hat{\eta},t,t}.
\end{aligned} \tag{C.6}$$

It follows that

$$\ddot{Q}_{zx,t}\geq\dot{Q}_{zx,t,t-1}\ddot{Q}_{zx,t,t-1}^{-1}\dot{Q}'_{zx,t,t-1} \text{ for } t=2,...,T,$$

which further implies that

$$\bar{\Lambda}_2=\left(\ddot{Q}_{zx,2}+\ddot{Q}_{zx,3,2}\right)-\dot{Q}_{zx,2,1}\bar{\Lambda}_1^{-1}\dot{Q}'_{zx,t,t-1}\geq\ddot{Q}_{zx,3,2}>0 \tag{C.7}$$

and by induction that

$$\bar{\Lambda}_t\geq\left(\ddot{Q}_{zx,t}+\ddot{Q}_{zx,t+1,t}\right)-\dot{Q}_{zx,t,t-1}\ddot{Q}_{zx,t,t-1}^{-1}\dot{Q}'_{zx,t,t-1}\geq\ddot{Q}_{zx,t+1,t}>0 \text{ for } t=2,...,T-1. \tag{C.8}$$

In addition, by (C.6) and (C.1)

$$\bar{\Lambda}_T \geq \ddot{Q}_{zx,T} - \dot{Q}_{zx,T,T-1}\ddot{Q}_{zx,T,T-1}^{-1}\dot{Q}'_{zx,T,T-1} = \phi'_{z\hat{\eta},T,T}W_t\phi'_{z\hat{\eta},T,T} > 0. \tag{C.9}$$

Consequently, $\min_{1\leq t\leq T}\mu_{\min}\left(\bar{\Lambda}_t\right) \geq \min\{\min_{1\leq t\leq T-1}\mu_{\min}(\ddot{Q}_{zx,t+1,t}), \mu_{\min}\left(\phi'_{z\hat{\eta},T,T}W_t\phi'_{z\hat{\eta},T,T}\right)\}$. In view of the fact that $\dot{Q}_{zx,t,t-1}\bar{\Lambda}_{t-1}^{-1}\dot{Q}'_{zx,t,t-1}$ is p.s.d. for $t = 2, ..., T$, we have $\bar{\Lambda}_t \leq \ddot{Q}_{zx,t} + \ddot{Q}_{zx,t+1,t}$ for $t = 1, ..., T-1$ and $\bar{\Lambda}_T \leq \ddot{Q}_{zx,T}$. It follows that $\max_{1\leq t\leq T}\mu_{\max}\left(\bar{\Lambda}_t\right) \leq \max_{1\leq t\leq T-1}\mu_{\max}(\ddot{Q}_{zx,t+1,t}) + \max_{1\leq t\leq T}\mu_{\max}(\ddot{Q}_{zx,t})$.

By Assumptions B.1 and B.2(iv) and using arguments as used in the proofs of Lemma 3.1(ii)-(iii), we can readily show that $\frac{1}{2}\underline{c}_{zx}\underline{c}_w \leq \min_{1\leq t\leq T-1}\mu_{\min}(\ddot{Q}_{zx,t+1,t}) \leq \max_{1\leq t\leq T-1}\mu_{\max}(\ddot{Q}_{zx,t+1,t}) \leq 2\bar{c}_{zx}\bar{c}_w$ and $\frac{1}{2}\underline{c}_{zx}\underline{c}_w \leq \min_{1\leq t\leq T-1}\mu_{\min}(\ddot{Q}_{zx,t}) \leq \max_{1\leq t\leq T-1}\mu_{\max}(\ddot{Q}_{zx,t}) \leq 2\bar{c}_{zx}\bar{c}_w$ w.p.a.1. Then (C.2) follows.

Let $\bar{\Lambda}$ denote a block diagonal matrix whose diagonal blocks are denoted by $\bar{\Lambda}_t$ for $t = 1, ..., T$. Let $\bar{L}$ denote the block lower part of $\ddot{Q}_{NT}$. By (B.1), the inverse $\bar{\Lambda}^{-1}$ of $\bar{\Lambda}$ exists asymptotically and we can consider the block LU factorization of the SBTM $\ddot{Q}_{NT}$ : $\ddot{Q}_{NT} = \left(\bar{\Lambda} + \bar{L}\right)\bar{\Lambda}^{-1}\left(\bar{\Lambda} + \bar{L}'\right)$. Following the proof of Lemma 3.1(i), we can readily show that w.p.a.1,

$$\mu_{\max}\left(\ddot{Q}_{NT}\right) \leq \left[\mu_{\max}\left(\bar{\Lambda} + \bar{L}\right)\right]^2\mu_{\max}\left(\bar{\Lambda}^{-1}\right) \leq (4\bar{c}_{zx}\bar{c}_w)^2\left[\frac{1}{2}\min(\underline{c}_{zx}\underline{c}_w, \underline{c}_{z\eta})\right]^{-1},$$

and

$$\mu_{\min}\left(\ddot{Q}_{NT}\right) \geq \left[\mu_{\min}\left(\bar{\Lambda} + \bar{L}\right)\right]^2\mu_{\min}\left(\bar{\Lambda}^{-1}\right) \geq \left[\frac{1}{2}\min(\underline{c}_{zx}\underline{c}_w, \underline{c}_{z\eta})\right]^2(4\bar{c}_{zx}\bar{c}_w)^{-1}.$$

The lemma holds with $\underline{c}_{\ddot{Q}} = [\frac{1}{2}\min(\underline{c}_{zx}\underline{c}_w, \underline{c}_{z\eta})]^2(4\bar{c}_{zx}\bar{c}_w)^{-1}$ and $\bar{c}_{\ddot{Q}} = (4\bar{c}_{zx}\bar{c}_w)^2[\frac{1}{2}\min(\underline{c}_{zx}\underline{c}_w, \underline{c}_{z\eta})]^{-1}$. ■

Next, we state a technical lemma whose proof is given in the supplemental appendix.

**Lemma C.1** *Suppose Assumption B.1 holds. Then* $\ddot{\beta}_t - \beta_t^0 = O_P\left(N^{-1/2}\right)$ *for each* $t = 1, 2, ..., T$.

**Proof of Theorem 4.2.** (i) The proof parallels that of Theorem 3.2 and we only sketch it. Let $\hat{b}_t = N^{1/2}(\hat{\beta}_t - \beta_t^0)$ and $\hat{\mathbf{b}} = N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)$. Noting that $\Delta y_{it} - x'_{it}\beta_t + x'_{i,t-1}\beta_{t-1} = \Delta u_{it} - N^{-1/2}\xi_{it}$ where $\xi_{it} = x'_{it}b_t - x'_{i,t-1}b_{t-1}$, we have

$$\begin{aligned}
&N\left[V_{2NT,\lambda_2}\left(\boldsymbol{\beta}\right) - V_{2NT,\lambda_2}\left(\boldsymbol{\beta}^0\right)\right] \\
&= \mathbf{b}'\ddot{Q}_{NT}\mathbf{b} - 2\mathbf{b}'\sqrt{N}\ddot{R}_{NT}^u + N\lambda_2\sum_{t\in\mathcal{T}_{m^0}^0}\ddot{w}_t\left[\left\|\beta_t^0 - \beta_{t-1}^0 + N^{-1/2}(b_t - b_{t-1})\right\| - \left\|\beta_t^0 - \beta_{t-1}^0\right\|\right] \\
&+ N\lambda_2\sum_{t\in\mathcal{T}_{m^0}^{0c}}\ddot{w}_t\left\|N^{-1/2}(b_t - b_{t-1})\right\| \\
&\equiv B_1\left(\mathbf{b}\right) - 2B_2\left(\mathbf{b}\right) + B_3\left(\mathbf{b}\right) + B_4\left(\mathbf{b}\right), \text{ say.}
\end{aligned}$$

As in the proof of Theorem 3.2, we can show that $\left|T^{-1}B_3\left(\mathbf{b}\right)\right| = O_P\left((m^0N)^{1/2}\lambda_2T^{-1/2}J_{\min}^{-\kappa_2}\right)T^{-1/2}\left\|\mathbf{b}\right\| = O_P\left(1\right)T^{-1/2}\left\|\mathbf{b}\right\|$ and w.p.a.1

$$\left[B_1\left(\mathbf{b}\right) - 2B_2\left(\mathbf{b}\right) + B_3\left(\mathbf{b}\right)\right]/T \geq \mu_{\min}\left(\ddot{Q}_{NT}\right)T^{-1}\left\|\mathbf{b}\right\|^2 - O_P\left(1\right)T^{-1/2}\left\|\mathbf{b}\right\| > 0$$

if $T^{-1/2}\left\|\mathbf{b}\right\| = L$ is sufficiently large. Consequently, $N\left[V_{2NT,\lambda_2}\left(\boldsymbol{\beta}\right) - V_{2NT,\lambda_2}\left(\boldsymbol{\beta}^0\right)\right] > 0$ w.p.a.1 for large $L$ and $V_{2NT,\lambda_2}\left(\boldsymbol{\beta}\right)$ cannot be minimized in this case. This further implies that $T^{-1/2}\left\|\hat{\mathbf{b}}\right\|$ has to be stochastically bounded.

(ii) The proof is analogous to that of the second part of Theorem 3.2 by utilizing the fact that $\ddot{Q}_{NT}$ is an asymptotically nonsingular symmetric block tridiagonal matrix. ■

**Proof of Theorem 4.3.** We want to demonstrate that

$$P\left(\left\|\hat{\theta}_t\right\| = 0 \text{ for all } t \in \mathcal{T}^{0c}_{m^0}\right) \to 1 \text{ as } N \to \infty. \tag{C.10}$$

Suppose that to the contrary, $\hat{\theta}_t = \hat{\beta}_t - \hat{\beta}_{t-1} \neq 0$ for some $t \in \mathcal{T}^{0c}_{m^0}$ for sufficiently large $N$. To consider the optimization conditions wrt $\beta_t$, $t \geq 2$, based on subdifferential calculus (e.g., Bersekas 1995, Appendix B.5), we distinguish two cases: (a) $2 \leq t \leq T-1$ and (b) $t = T$ and $T \in \mathcal{T}^{0c}_{m^0}$.

In case (a), we consider two subcases: (a1) $t+1 = T^0_j \in \mathcal{T}^0_{m^0}$ for some $j = 1, ..., m^0$, and (a2) $t+1 \in \mathcal{T}^{0c}_{m^0}$. In either case, we can apply the FOC wrt $\beta_t$ and the equality $\Delta y_{it} = \beta^{0\prime}_t x_{it} - \beta^{0\prime}_{t-1} x_{i,t-1} + \Delta u_{it}$ to obtain

$$\begin{aligned}
\mathbf{0}_{p\times 1} &= -\frac{2}{N}\sum_{i=1}^{N} x_{it} z_{it}' W_t \frac{1}{\sqrt{N}} \sum_{i=1}^{N} z_{it}\left[\Delta y_{it} - \hat{\beta}_t' x_{it} + \hat{\beta}_{t-1}' x_{i,t-1}\right] \\
&\quad + \frac{2}{N}\sum_{i=1}^{N} x_{it}' z_{i,t+1} W_{t+1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} z_{i,t+1}\left[\Delta y_{i,t+1} - \hat{\beta}_{t+1}' x_{i,t+1} + \hat{\beta}_t' x_{it}\right] \\
&\quad + \sqrt{N}\lambda_2 \ddot{w}_t \frac{\hat{\theta}_t}{\left\|\hat{\theta}_t\right\|} - \sqrt{N}\lambda_2 \ddot{w}_{t+1} e_{t+1} \\
&= -2\phi_{zx,t}' W_t \frac{1}{\sqrt{N}} \sum_{i=1}^{N} z_{it}\left[\Delta u_{it} - \left(\hat{\beta}_t - \beta_t^0\right)' x_{it} + \left(\hat{\beta}_{t-1} - \beta_{t-1}^0\right)' x_{i,t-1}\right] \\
&\quad + 2\phi_{zx,t+1,t}' W_{t+1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} z_{i,t+1}\left[\Delta u_{i,t+1} - \left(\hat{\beta}_{t+1} - \beta_{t+1}^0\right)' x_{i,t+1} + \left(\hat{\beta}_t - \beta_t^0\right)' x_{it}\right] \\
&\quad + \sqrt{N}\lambda_2 \ddot{w}_t \frac{\hat{\theta}_t}{\left\|\hat{\theta}_t\right\|} - \sqrt{N}\lambda_2 \ddot{w}_{t+1} e_{t+1} \\
&= -2\sqrt{N}\left[\phi_{zx,t+1,t}' W_{t+1} \phi_{zx,t+1}\left(\hat{\beta}_{t+1} - \beta_{t+1}^0\right) - \phi_{zx,t}' W_{t+1} \phi_{zx,t}\left(\hat{\beta}_t - \beta_t^0\right)\right. \\
&\quad \left. -\phi_{zx,t+1,t}' W_{t+1} \phi_{zx,t+1,t}\left(\hat{\beta}_t - \beta_t^0\right) + \phi_{zx,t}' W_t \phi_{zx,t,t-1}\left(\hat{\beta}_{t-1} - \beta_{t-1}^0\right)'\right] \\
&\quad + 2\sqrt{N}\left(\phi_{zx,t+1,t}' W_{t+1} \phi_{z\Delta u,t+1} - \phi_{zx,t} W_t \phi_{z\Delta u,t}\right) + \sqrt{N}\lambda_2 \ddot{w}_t \frac{\hat{\theta}_t}{\left\|\hat{\theta}_t\right\|} - \sqrt{N}\lambda_2 \ddot{w}_{t+1} e_{t+1} \\
&\equiv B_{1t} + B_{2t} + B_{3t} - B_{4t}, \text{ say,}
\end{aligned} \tag{C.11}$$

where $\hat{e}_{t+1} = \hat{\theta}_{t+1} / \left\|\hat{\theta}_{t+1}\right\|$ if $\left\|\hat{\theta}_{t+1}\right\| \neq 0$ and $\left\|\hat{e}_{t+1}\right\| \leq 1$ otherwise.

Since $\hat{\theta}_t \neq 0$, there exists $r \in \{1, ..., p\}$ such that $\left|\hat{\theta}_{t,r}\right| = \max\left\{\left|\hat{\theta}_{t,l}\right|,\ l = 1, ..., p\right\}$, where for any $p \times 1$ vector $a_t$, $a_{t,l}$ denotes its $l$th element. Wlog assume that $r = p$, implying that $\left|\hat{\theta}_{t,p}\right| / \left\|\hat{\theta}_t\right\| \geq 1/\sqrt{p}$. By Assumptions B.1(i)-(ii) and Theorem 4.2, $B_{1t,p} = O_P(1)$ and $B_{2t,p} = O_P(1)$. In view of the fact that $\ddot{w}_t^{-1} = O_P(N^{-\kappa_2/2})$ for $t \in \mathcal{T}_{m^0}^0$, $|B_{3t,p}| \geq \sqrt{N}\lambda_2 \ddot{w}_t/\sqrt{p}$, which is explosive in probability under Assumption B.2(iii) ($N^{(\kappa_2+1)/2}\lambda_2 \to \infty$). To bound the probability order of $B_{4t,p}$, we distinguish two subcases. In subcase (a1), noting that $\dot{\beta}_{t+1} - \dot{\beta}_t \xrightarrow{P} \theta_{t+1}^0 \neq 0$ by Theorem 4.2, we have $\ddot{w}_{t+1} = \left\|\theta_{t+1}^0 + O_P(N^{-1/2})\right\|^{-\kappa_2} = O_P\left(J_{\min}^{-\kappa_2}\right)$ and $B_{4t} = \sqrt{N}\lambda_2 \ddot{w}_{t+1}\hat{e}_{t+1,p} = O_P(\sqrt{N}\lambda_2 J_{\min}^{-\kappa_2}) = O_P\left(1\right).$ Consequently, $|B_{3t,p}| \gg |B_{1t,p} + B_{2t,p} + B_{4t,p}|$ so that (C.11) cannot be true for sufficiently large $N$ or $(N, T)$. Then we conclude that w.p.a.1, $\hat{\theta}_t$ must be in a position where $\left\|\hat{\theta}_t\right\|$ is not differentiable in subcase (a1). In addition, a direct application of this result is that if $T_j^0 - 1 \in \mathcal{T}_{m^0}^{0c}$ for some $j = 1, ..., m^0$, then $P\left(\left\|\hat{\theta}_{T_j^0-1}\right\| = 0\right) \to 1$ as $N \to \infty$ and $\sqrt{N}\lambda_2 \ddot{w}_{T_j^0-1} e_{T_j^0-1} = O_P\left(1\right)$ in order for the FOC to hold for $t = T_j^0 - 1$.

In subcase (a2), we apply deductive arguments as used in the proof of Theorem 3.3 and the result in subcase (a1) so show that $\hat{\theta}_t$ must be in a position that $\left\|\hat{\theta}_t\right\|$ is not differentiable for all $t \in \mathcal{T}_{m^0}^{0c}$ and $t \neq T$.

In case (b), noting that only one term in the penalty term ($\lambda_2 \sum_{t=2}^T \ddot{w}_t \left\|\beta_t - \beta_{t-1}\right\|$) is involved with $\beta_T$, it is easy to show that $\hat{\theta}_T = \hat{\beta}_T - \hat{\beta}_{T-1}$ must be in a position where $\left\|\hat{\theta}_T\right\|$ is not differentiable if $T \in \mathcal{T}_{m^0}^{0c}$. Consequently (C.10) follows. ■

**Proof of Corollary 4.4.** The proof is analogous to that of Corollary 3.4 by using Theorems 4.2-4.3 instead. ■

**Proof of Theorem 4.5.** Note that $\hat{\boldsymbol{\alpha}}_{\hat{m}}^p(\hat{\mathcal{T}}_{\hat{m}}) = (\hat{\alpha}_1^p(\hat{\mathcal{T}}_{\hat{m}})', ..., \hat{\alpha}_{\hat{m}+1}^p(\hat{\mathcal{T}}_{\hat{m}})')' = \arg\min_{\boldsymbol{\alpha}_m} Q_{2NT}\left(\boldsymbol{\alpha}_m; \hat{\mathcal{T}}_{\hat{m}}\right).$ The first order conditions for this minimization problem are

$$\begin{aligned}
\mathbf{0}_{p\times 1} &= \frac{-2}{N}\sum_{t=2}^{\hat{T}_1-1}\sum_{i=1}^{N}\Delta x_{it}z_{it}'W_1^p\frac{1}{N}\sum_{t=2}^{\hat{T}_1-1}\sum_{i=1}^{N}z_{it}\left(\Delta y_{it} - \hat{\alpha}_1^{p\prime}\Delta x_{it}\right)\\
&\quad+\frac{2}{N}\sum_{i=1}^{N}x_{i,\hat{T}_1-1}z_{i\hat{T}_1}'\ W_{\hat{T}_1}\frac{1}{N}\sum_{i=1}^{N}z_{i\hat{T}_1}\left(\Delta y_{i\hat{T}_1} - \hat{\alpha}_2^{p\prime}x_{i\hat{T}_1} + \hat{\alpha}_1^{p\prime}x_{i,\hat{T}_1-1}\right),\\
\mathbf{0}_{p\times 1} &= \frac{-2}{N}\sum_{t=\hat{T}_{j-1}+1}^{\hat{T}_j-1}\ \sum_{i=1}^{N}\Delta x_{it}z_{it}'W_j^p\frac{1}{N}\sum_{t=\hat{T}_{j-1}+1}^{\hat{T}_j-1}\ \sum_{i=1}^{N}z_{it}\left(\Delta y_{it} - \hat{\alpha}_j^{p\prime}\Delta x_{it}\right)\\
&\quad+\frac{2}{N}\sum_{i=1}^{N}x_{i,\hat{T}_j-1}z_{i\hat{T}_j}'W_{\hat{T}_j}\frac{1}{N}\sum_{i=1}^{N}z_{i\hat{T}_j}\left(\Delta y_{i\hat{T}_j} - \hat{\alpha}_{j+1}^{p\prime}x_{i\hat{T}_j} + \hat{\alpha}_j^{p\prime}x_{i,\hat{T}_j-1}\right)\\
&\quad-\frac{2}{N}\sum_{i=1}^{N}x_{i,\hat{T}_{j-1}}z_{i\hat{T}_{j-1}}'W_{\hat{T}_{j-1}}\frac{1}{N}\sum_{i=1}^{N}z_{i,\hat{T}_{j-1}}\left(\Delta y_{i\hat{T}_{j-1}} - \hat{\alpha}_j^{p\prime}x_{i\hat{T}_{j-1}} + \hat{\alpha}_{j-1}^{p\prime}x_{i,\hat{T}_{j-1}-1}\right)\text{ for } j{=}2, ..., \hat{m},\\
\mathbf{0}_{p\times 1} &= \frac{-2}{N}\sum_{t=\hat{T}_{\hat{m}}+1}^{T}\ \sum_{i=1}^{N}\Delta x_{it}z_{it}'W_{\hat{m}+1}^p\frac{1}{N}\sum_{t=\hat{T}_{\hat{m}}+1}^{T}\ \sum_{i=1}^{N}z_{it}\left(\Delta y_{it} - \hat{\alpha}_{\hat{m}+1}^{p\prime}\Delta x_{it}\right)\\
&\quad-\frac{2}{N}\sum_{i=1}^{N}x_{i\hat{T}_{\hat{m}}}z_{i\hat{T}_{\hat{m}}}'W_{\hat{T}_{\hat{m}}}\frac{1}{N}\sum_{i=1}^{N}z_{i\hat{T}_{\hat{m}}}\left(\Delta y_{i\hat{T}_{\hat{m}}} - \hat{\alpha}_{\hat{m}+1}^{p\prime}x_{i\hat{T}_{\hat{m}}} + \hat{\alpha}_{\hat{m}}^{p\prime}x_{i,\hat{T}_{\hat{m}}-1}\right),
\end{aligned}$$

where we suppress the dependence of $\hat{\alpha}_j^p$'s on $\hat{\mathcal{T}}_{\hat{m}}$. We can readily verify that $\hat{\boldsymbol{\alpha}}_{\hat{m}}^p = \hat{\boldsymbol{\alpha}}_{\hat{m}}^p\left(\hat{\mathcal{T}}_{\hat{m}}\right) = \Upsilon_{NT}\left(\hat{\mathcal{T}}_{\hat{m}}\right)^{-1}\Xi_{NT}^y\left(\hat{\mathcal{T}}_{\hat{m}}\right)$, where $\Upsilon_{NT}\left(\cdot\right)$ and $\Xi_{NT}^y$ are defined in (A.8) and (A.9) in Appendix A.2, respectively.

By Corollary 4.4, $\hat{\boldsymbol{\alpha}}^p_{\hat{m}}(\hat{\mathcal{T}}_{\hat{m}}) = \hat{\boldsymbol{\alpha}}^p_{m^0}\left(\mathcal{T}_{m^0}\right)$ w.p.a.1. Therefore we can study the asymptotic distribution of $\hat{\boldsymbol{\alpha}}^p_{\hat{m}}(\hat{\mathcal{T}}_{\hat{m}})$ by studying that of $\hat{\boldsymbol{\alpha}}_{m^0}\left(\mathcal{T}^0_{m^0}\right)$. Note that $\hat{\boldsymbol{\alpha}}_{m^0}\left(\mathcal{T}^0_{m^0}\right) = \Upsilon^{-1}_{NT}\Xi_{NT}$, where $\Upsilon_{NT}$ and $\Xi_{NT}$ are defined in (4.1). It is easy to verify that

$$\begin{aligned}\sqrt{N}S\mathbb{D}_{m^0+1}\left(\hat{\boldsymbol{\alpha}}^p_{m^0}\left(\mathcal{T}^0_{m^0}\right) - \boldsymbol{\alpha}^0\right) &= S\left(\mathbb{D}^{-3}_{m^0+1}\Upsilon_{NT}\mathbb{D}^{-1}_{m^0+1}\right)^{-1}\sqrt{N}\mathbb{D}^{-3}_{m^0+1}\Xi^u_{NT} \\ &= S\Upsilon_0^{-1}\check{V}_{NT} + S(\check{\Upsilon}^{-1}_{NT} - \Upsilon_0^{-1})\check{V}_{NT},\end{aligned}$$

where $\Xi^u_{NT} = \Xi^u_{NT}\left(\mathcal{T}^0_{m^0}\right)$ and $\Xi^u_{NT}\left(\cdot\right)$ is defined in (A.9), $\check{V}_{NT} = \sqrt{N}\mathbb{D}^{-3}_{m^0+1}\Xi^u_{NT}$, and $\check{\Upsilon}_{NT} = \mathbb{D}^{-3}_{m^0+1}\Upsilon_{NT}\mathbb{D}^{-1}_{m^0+1}$. By Assumption B.3(ii), $S\Upsilon_0^{-1}\check{V}_{NT} \overset{D}{\to} N\left(0, S\Upsilon_0^{-1}\Sigma_0\Upsilon_0^{-1}S'\right)$. Using arguments as used in the proof of Theorem 3.5, we can show that $\left\|S(\check{\Upsilon}^{-1}_{NT} - \Upsilon_0^{-1})\check{V}_{NT}\right\| = o_P\left(1\right)$. Then the result follows by the Slutsky lemma. ■

**Proof of Theorem 4.6.** The proof is analogous to that of Theorem 3.6 and thus omitted. ■

*REFERENCE*

Andrews, D. W. K., 1993. Tests for parameter instability and structural change with unknown change point. Econometrica 61, 821-856.

Andrews, D. W. K., 2003. End-of-sample instability tests. Econometrica 71, 1661-1694.

Angelosante, D., Giannakis, G.B., 2012. Group Lassoing change-points in piecewise-constant AR processes. EURASIP Journal on Advances in Signal Processing 1, 1-16.

Bai, J., 1997a. Estimation of a change point in multiple regression models. Review of Economics and Statistics 79, 551-563.

Bai, J., 1997b. Estimating multiple breaks one at a time. Econometric Theory 13, 315-352.

Bai, J., 2010. Common breaks in means and variances for panel data. Journal of Econometrics 157, 78-92.

Bai, J., Lumsdaine, R. L., Stock, J., 1998. Testing and dating common breaks in multivariate time series. Review of Economic Studies 65, 395-432.

Bai, J., Perron, P., 1998. Estimating and testing liner models with multiple structural changes. Econometrica 66, 47-78.

Baltagi, B. H., Feng, Q., Kao, C., 2015. Estimation of heterogeneous panels with structural breaks. Journal of Econometrics, forthcoming.

Baltagi, B. H., Kao, C., Liu, L., 2014. Estimation and identification of change points in panel models with nonstationary or stationary regressors and error terms. Working paper, Syracuse University.

Belloni, A., Chernozhukov, V., Hansen, C., 2012. Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica 80, 2369-2429.

Ben-David, D., Papell, D. H., 1995. The great wars, the great crash, and steady state growth: some new evidence about an old stylized fact. Journal of Monetary Economics 36, 453-475.

Bernstein, D. S., 2005. Matrix Mathematics: Theory, Facts, and Formulas with Application to Linear Systems Theory. Princeton University Press, Princeton.

Bertsekas, D., 1995. Nonlinear Programming. Athena Scientific, Belmont, MA.

Breitung, J., Eickmeier, S., 2011. Testing for structural breaks in dynamic factor models. Journal of Econometrics 163, 71-84.

Caner, M., 2009. Lasso-type GMM estimator. Econometric Theory 25, 270-290.

Caner, M., Han, X., 2014. Selecting the correct number of factors in approximate factor models: the large panel case with group Bridge estimators. Journal of Business and Economics Statistics 32, 359-374.

Caner, M., Knight, K., 2013. An alternative to unit root tests: Bridge estimators differentiate between nonstationary versus stationary models and select optimal lag. Journal of Statistical Planning and Inference 143, 691-715.

Chan, F., Mancini-Griffoli, T., Pauwels, L. L., 2008. Stability tests for heterogenous panel data. Working paper, Curtin University of Technology.

Cheng, X., Liao, Z., 2015. Select the valid and relevant moments: an information-based LASSO for GMM with many moments. Journal of Econometrics 186, 443–464.

Cheng, X., Liao, Z., Schorfheide, F., 2015. Shrinkage estimation of high-dimensional factor models with structural instabilities. NBER Working Paper Series 19792, National Bureau of Economic Research.

De Watcher, S., Tzavalis, E., 2005. Monte Carlo comparison of model and moment selection and classical inference approaches to break detection in panel data models. Economics Letters 99, 91-96.

De Watcher, S., Tzavalis, E., 2012. Detection of structural breaks in linear dynamic panel data models. Computational Statistics and Data Analysis 56, 3020-3034.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96, 1348-1360.

Fan, J., Liao, Y., 2014. Ultra high dimensional variable selection with endogenous covariates. Annals of Statistics 42, 872-917.

Fan, J., Peng, H., 2004. Nonconcave penalized likelihood with a diverging number of parameters. Annals of Statistics 32, 928-961.

García, P. E. 2011. Instrumental variable estimation and selection with many weak and irrelevant instruments. Working paper, University of Wisconsin, Madison.

Harville, D. A., 1997. Matrix Algebra from a Statistician's Perspective. Springer, New York.

Hsu, C-C., Lin, C-C., 2012. Change-point estimation for nonstationary panel. Working paper, National Central University.

Islam, N., 1995. Growth empirics: a panel data approach. The Quarterly Journal of Economics 110, 1127-1170.

Jones, C. I., 2002. Sources of U.S. economic growth in a world of ideas. American Economic Review 92, 220-239.

Ke, Z., Fan, J., Wu, Y., 2015. Homogeneity pursuit. Journal of American Statistical Association 110, 175-194.

Kim, D., 2011. Estimating a common deterministic time trend break in large panels with cross sectional dependence. Journal of Econometrics 164, 310-330.

Kim, D., 2014. Common breaks in time trends for large panel data with a factor structure. Econometrics Journal, forthcoming.

Knight, K., Fu, W., 2000. Asymptotics for Lasso-type estimators. Annals of Statistics 28, 1356-1378.

Kock, A. B., 2013. Oracle efficient variable selection in random and fixed effects panel data models. Econometric Theory 29, 115-152.

Kurozumi, E., 2015. Testing for multiple structural changes with non-homogeneous regressors. Journal of Time Series Econometrics 7, 1-35.

Lam, C. Fan, J., 2008. Profile-kernel likelihood inference with diverging number of parameters. Annals of Statistics 36, 2232-2260.

Liao, W., Wang, P., 2012. Structural breaks in panel data models: a common distribution approach. Working paper, HKUST.

Liao, Z., 2013. Adaptive GMM shrinkage estimation with consistent moment selection. Econometric Theory 29, 857-904.

Liao, Z., Phillips, P. C. B., 2015. Automated estimation of vector error correction models. Econometric Theory 31, 581-646.

Lu, X., Su, L., 2015a. Jackknife model averaging for quantile regressions. Journal of Econometrics 188, 40-58.

Lu, X., Su, L., 2015b. Shrinkage estimation of dynamic panel data models with interactive fixed effects. Journal of Econometrics, forthcoming.

Meurant, G., 1992. A review on the inverse of symmetric tridiagonal and block tridiagonal matrices. SIAM Journal of Matrix Analysis and Applications 13, 707-728.

Molinari, L. G., 2008. Determinant of block tridiagonal matrices. Linear Algebra and Its Applications 429, 2221-2226.

Pesaran, M. H., 2006. Estimation and inference in large heterogeneous panels with a multifactor error structure. Econometrica 74, 967-1012.

Pesaran, M. H., Yamagata, T., 2008. Testing slope homogeneity in large panels. Journal of Econometrics 142, 50-93.

Qian, J., Su, L., 2014. Structural change estimation in time series regressions with endogenous variables. Economics Letters 125, 415-421.

Qian, J., Su, L., 2015. Shrinkage estimation of regression models with multiple structural changes. Econometric Theory, forthcoming.

Qu, Z., Perron, P., 2007. Estimating and testing structural changes in multiple regressions. Econometrica 75, 459-502.

Ran, R-S., Huang, T-Z., 2006. The inverses of block tridiagonal matrices. Applied Mathematics and Computation 179, 243-247.

Rinaldo, A., 2009. Properties and refinement of the fused Lasso. Annals of Statistics 37, 2922-2952.

Romer, P. M., 1986. Increasing returns and long-run growth. Journal of Political Economy 94, 1002-1037.

Serfling, R. J., 1980. Approximation Theorems of Mathematical Statistics. John Wiley & Sons, New York.

Su, L., Chen, Q., 2013. Testing homogeneity in panel data models with interactive fixed effects. Econometric Theory 29, 1079-1135.

Su, L., Shi, Z., Phillips, P. C. B., 2014. Identifying latent structures in panel data. Working paper, Dept. of Economics, Yale University.

Su, L., Wang, X., 2015. On time-varying factor models: estimation and testing. Working paper, Singapore Management University.

Su, L., White, H., 2010. Testing structural change in partially linear models. Econometric Theory 26, 1761-1806.

Tibshirani, R. J., 1996. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B 58, 267-288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K., 2005. Sparsity and smoothness via the fused Lasso. Journal of the Royal Statistical Society, Series B 67, 91-108.

Wang, H., Leng, C., 2008. A note of adaptive group Lasso. Computational Statistics and Data Analysis 52, 5277-5286.

Wang, H., Li, R., Tsai, C.-L., 2007. Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika 94, 553-568.

Yamazaki, D., Kurozumi, E., 2014. Testing for parameter constancy in the time series direction in fixed-effect panel data models. Journal of Statistical Computation and Simulation 85, 2874-2902.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B 68, 49-67.

Zhang, Y., Li, R., Tsai, C-L., 2010. Regularization parameter selections via generalized information criterion. Journal of American Statistical Association 105, 312-323.

Zou, H., 2006. The adaptive Lasso and its oracle properties. Journal of the American Statistical Association 101, 1418-1429.