Estimation of panel group structure models with structural breaks in group memberships and coefficients*

Robin L. Lumsdaine† Ryo Okui‡ Wendun Wang§

January 31, 2021

Abstract

This paper considers linear panel data models with a grouped pattern of heterogeneity when the latent group membership structure and/or the values of slope coefficients change at a break point. We propose a least squares approach to jointly estimate the break point, group membership structure, and coefficients. The proposed estimators are consistent, and the asymptotic distribution of the coefficient estimators is identical to that under known break point and group structure even when the cross-sectional sample size is much larger than the length of time series. Monte Carlo simulations and an empirical example illustrate the use of the approach and associated inference.

Keywords: Panel data; grouped patterns; structural breaks; group membership change.

JEL classification: C23; C38; C51.

---

*__Views and opinions expressed are those of the authors only and do not necessarily represent the views of any institutions with which they are affiliated.__ Okui acknowledges financial support from School of Social Sciences of Seoul National University, New Faculty Startup Grant of Seoul National University and the Housing and Commercial Bank Economic Research Fund for Institute of Economic Research of Seoul National University. The authors would also like to thank Serena Ng (Editor), the associate editor, two anonymous referees, Sang Yoon (Tim) Lee, Hyungsik Roger Moon, Bent Nielsen, Liangjun Su, seminar and session participants of the Asian Meeting of the Econometric Society 2019, the International Association for Applied Econometrics Annual Conference 2019, the Korean Economic Review International Conference 2019, Queen Mary University of London, University of Oxford, Nanyang Technological University, and the University of Tokyo for their valuable comments and suggestions. All errors are ours.

†Department of Finance and Real Estate, Kogod School of Business, American University; Econometric Institute, Erasmus University Rotterdam; National Bureau of Economic Research; Tinbergen Institute. Email: robin.lumsdaine@american.edu

‡Department of Economics and the Institute of Economic Research, Seoul National University. Email: okuiryo@snu.ac.kr

§Econometric Institute, Erasmus University Rotterdam; Tinbergen Institute. Email: wang@ese.eur.nl

1 Introduction

When conducting economic analyses using panel data, it is often important to take into account the time-varying and cross-sectionally heterogeneous nature of economic relationships. The functional relationship between economic variables is frequently influenced by various exogenous shocks, such as financial crises, technology changes, policy implementation, etc. The impacts of these shocks are often modelled by structural breaks of the slope coefficients in a regression model, and they can differ remarkably across individual units. Two examples of such heterogeneity are the impacts of the implementation of the U.S. Sarbanes-Oxley (SOX) Act in 2002 and the European sovereign debt crisis. SOX created new rules and imposed more stringent requirements for US public company boards, management and public accounting firms. A number of studies have shown that the effects of SOX on corporate governance vary remarkably across firms (see, e.g., Heron and Lie, 2007; Chhaochharia and Grinstein, 2007; Banerjee et al., 2015, among others). For the European debt crisis, the economic structure of southern European countries was affected to a larger extent than that of the central European countries (Claeys and Vašíček, 2014), and its global impact also varied significantly across countries (Stracca, 2015).

A salient empirical finding is that such cross-sectional impact tends to be common within a group of units but varies across groups (see, e.g., Mitton, 2002; Linck et al., 2009; Duchin et al., 2010; Thakor, 2015, among many others). Thus, imposing a group pattern is a plausible yet parsimonious way of modelling cross-sectional heterogeneity. Many studies have considered clustering individual units in panel data (Sun, 2015; Hahn and Moon, 2010; Lin and Ng, 2012; Bonhomme and Manresa, 2015; Su et al., 2016; Ando and Bai, 2016; Vogt and Linton, 2016; Wang et al., 2018, among others), all assuming that the group structure does not change over time. On the other hand, while there is a vast volume of literature on structural breaks in panel data models (Bai, 2010; Wachter and Tzavalis, 2012; Kim, 2011; Baltagi et al., 2017; Qian and Su, 2016; Li et al., 2016, among others), most studies assume that coefficients are homogeneous across units or structural breaks are common to all units. Recently, Okui and Wang (2021) proposed a new model that allows researchers to capture a group pattern of heterogeneity in structural breaks. Although their model allows that the break point, the size of break, and slope coefficients can differ across groups in any arbitrary manner, the group structure is required to be time invariant, or in other words, the group membership structure is not affected by the shock.

The time-invariant group structure is sometimes a strong assumption. A macroeconomic shock, such as a financial crisis, may restructure the functional relations between economic variables for all units, such that some units become highly heterogeneous (homogeneous) after the shock even though their behaviour is similar (different) before the shock. For example, Germany and France shared a similar debt-to-GDP ratio before the European debt crisis (64–69% in 2007 and 2008), but the time series paths of the ratio of the two countries diverge dramatically after 2008, with that of France increasing from 83% to 98.4% until 2018 and that of Germany fluctuating around 75% in 2009–2014 and decreasing to 61.9% in 2018.[1] A similar situation may apply with the introduction of regulatory policies, such as SOX, which likely reshaped the heterogeneity pattern of firms according to their observable and unobservable time-varying management practices and firm characteristics (Engel et al., 2007; Heron and Lie, 2007; Linck et al., 2009). Hence, it is of great empirical interest to develop a method that allows the group structure to change after the break, in addition to allowing for a coefficient break. In practice, it is also possible that the shock only affects the group membership, but the functional relationship remains the same within each group. It is additionally possible that the number of groups changes. These scenarios raise several challenges. How do we model and estimate a time-varying group structure with an unknown break date? How do we determine the potentially varying number of groups? How do we identify whether a break occurs to only the group structure, the coefficient, or both?

To address these challenges, this paper proposes a new model and a new estimation approach that allows us to capture a structural break in the slope coefficients and/or group structure. We propose a least squares approach that estimates the break point, group membership, and regression coefficients simultaneously. To solve this least squares problem, we employ an iterative estimation approach that first iterates between coefficient and group estimation given a fixed break point, and then searches for the optimal break point given the estimated coefficients and group memberships. We show that the estimators of the break point and group memberships are both consistent. The consistent estimation of breaks and group structure further allows us to estimate the coefficients consistently, and the asymptotic behaviour of the coefficient estimator is equivalent to that obtained under known break point and group memberships. One of our theoretical contributions is to formally establish conditions under which the break detection error does not affect the asymptotic behaviour of the group membership and coefficient estimators.

An alternative empirical strategy is to examine each unit separately. Baltagi et al. (2016) propose detection of a *common* break point based on individual estimation of each unit. Compared with their approach, we impose a grouped pattern of heterogeneity. While group heterogeneity is more restrictive than individual heterogeneity, it brings various advantages over their approach. First, we can make use of the cross-sectional information. This allows

[1]Source of Data: Eurostat, general government gross debt - annual data (downloaded in January 2020). URL: https://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&language=en&pcode=teina225

us to estimate a model with a relatively short time dimension compared to the number of explanatory variables (or even when the time dimension is smaller than the number of explanatory variables), and yields efficiency gain. The efficiency gains further translate into a more precise detection of the break point. Such comparison is confirmed via an extensive simulation study, where we find that imposing a group pattern leads to more accurate break point estimates. Increasing both time and cross-section dimensions helps improve the accuracy of break detection for our method, but the cross-section dimension plays less role in individual-based break detection. This is because the group structure allows us to make use of cross-sectional variation, which typically leads to more efficient estimation. Second, the grouped pattern of heterogeneity enables us to investigate common patterns across units, which is difficult when looking at each unit completely separately, while maintaining heterogenous effects in our model.

A related approach in terms of motivation, yet different in terms of technique, is to combine factor models with structural breaks; see, e.g., Cheng et al. (2016). This approach captures heterogeneity by factors, while our approach employs group structure. Both factor structure (also called interactive effects) and grouped pattern are useful in modeling heterogeneity but from different perspectives. Which approach is more appropriate depends on the specific application and also the empirical objective. Our work is also related to Miao et al. (2020) which considered slope coefficients and thresholds varying over groups in a panel threshold regression. While both studies examine latent group structures and thresholding, a key difference is that Miao et al. (2020) assumed a constant group structure over the regimes (segmented by a threshold variable), but we allow the memberships to change across regimes.

We apply the proposed method to study the determinants of sales growth of US firms. We find that both the group structure of heterogeneity and the functional relationship between firms' variables and sales growth exhibit a significant break in the year of the Asian financial crisis. More groups emerge after the break, suggesting a larger degree of heterogeneity after the crisis. Our data-driven clustering results also suggest that the group structure is only moderately related to the industry classification that is widely used in corporate finance to capture heterogeneity.

The reminder of the paper is organized as follows. In Section 2, we explain the setting in Subsection 2.1 and the estimation method in Subsection 2.2. Subsection 2.3.1 explains how to choose the number of groups, and Subsection 2.3.2 provides some specification tests. Section 3 presents the asymptotic results. Section 4 discusses two extensions: models with individual fixed effects and multiple breaks. Finite sample results from Monte Carlo simulations are discussed in Section 5. An empirical example demonstrates how these tests and estimation methods can be used in practice in Section 6. Section 7 concludes. The proofs of the theorems and additional theoretical results are included in the technical appendix. An online supplement contains an additional algorithm, theoretical analyses for models with fixed effects, and additional simulation results.[2]

2 Model setup and estimation method

2.1 Model setup

Suppose that we have panel data (y_{it}, x_{it}) for $i = 1, \ldots, N$ and $t = 1, \ldots, T$, where y_{it} is a scalar outcome variable of interest, x_{it} is a vector of exogenous explanatory variables, and indexes i and t denote cross-sectional unit and time period, respectively. We are interested in the effects of x_{it} on y_{it}, and allow such effects to be (potentially) heterogeneous across units and vary over time. We model cross-sectional heterogeneity via a latent group pattern and the time-varying feature via a structural break. Importantly, we allow the structural break to change not only the effects of x_{it} but also the latent group pattern.

In particular, consider the following linear panel data model:

$$y_{it} = x_{it}' \beta_{g_{it},t} + u_{it}, \tag{1}$$

where u_{it} is an error term which may be heteroskedastic and weakly dependent in both time series and cross-sectional dimensions. We assume that x_{it} is exogenous such that $E(u_{it} \mid x_{it}) = 0$, and thus a predetermined regressor is allowed. $\beta_{g_{it},t}$ is the coefficient that depends on time period t and the group membership of i at time t, and it captures the effect of a change of its associated regressor holding other regressors constant. The current proof technique does not allow time effects and time trends to be parts of x_{it}, but their presence may be justified under alternative proof techniques. The assumptions are stated formally in Section 3.

Here the cross-sectional heterogeneity in the slope coefficients is featured by a group pattern, such that units in the same group share the same values of coefficients. Both the group membership structure (i.e. to which group each unit belongs) and the coefficient vector itself may change over time, and we consider the cases where the time-varying pattern can be characterized by structural breaks.

We assume that there is one structural break at time k^0. This implies that the group structure and the value of coefficients for each group do not change until k^0, and also remain stable after k^0 once the break happens. Let G^B and G^A denote the number of groups before and after the break, respectively. Note that superscripts B and A stand for "before the break" and "after the break", respectively. G^B and G^A may or may not be the same; that is, we

[2]The online supplement is available at https://drive.google.com/file/d/1ZvNs3Zoc6FyT0zt4Mj1RB7EvWTm0rJ-H/view?usp=sharing

allow the break to change the number of groups. Before the break, each unit belongs to one of the elements of $\mathbb{G}^B = \{1, \ldots, G^B\}$. After the break, the set of groups potentially changes and becomes $\mathbb{G}^A = \{1, \ldots, G^A\}$. Note that, because group memberships are unobserved, group labels are arbitrary and there may not be any natural correspondence between groups before and after the break. For example, group 1 before the break may not be related to group 1 after the break. The coefficient vector takes the form:

$$\beta_{g_{it},t} = \begin{cases} \beta_{g_i(B),B} & \text{if } t < k^0 \\ \beta_{g_i(A),A} & \text{if } t \geq k^0 \end{cases},$$

where $g_i(B) \in \mathbb{G}^B$ and $g_i(A) \in \mathbb{G}^A$ denote unit i's group membership before and after the break, respectively. Both the break point k^0 and group membership allocations, $g_i(B)$ and $g_i(A)$, are unknown and need to be estimated together with the slope coefficients. We first propose estimation methods for model (1) assuming that G^B and G^A are known. This knowledge is, of course, unavailable in many applications, and we discuss how to determine the number of groups in both regimes in Section 2.3.1.

As discussed in Okui and Wang (2021), a grouped pattern provides a sensible and convenient way to model individual heterogeneity, especially if there is also time instability, because it allows us to flexibly capture heterogeneity while maintaining the parsimony of the model, so that we can still take advantage of cross-sectional variation in coefficient estimation. Notably, the structural break may change the values of the coefficients, the group membership structure, or both. In some situations, it is of practical interest to identify these three cases, and we shall discuss this issue in Section 2.3.2. Also notice that our notation here implies that the group structure and slope coefficients experience a break at the same time point. This is, however, not essential as long as the method can be extended to multiple breaks since then these two different changes can be modelled as two breaks, one only changing the group structure and the other only changing the coefficients; see more detailed discussions in Section 4.2.

Several important models are nested in model (1). For example, the panel models with common structural breaks (Qian and Su, 2016; Baltagi et al., 2017) can be regarded as a special case of (1) by imposing cross-sectional homogeneity of $\beta_{g_{it},t}$. If the parameter $\beta_{g_{it},t}$ is constant over time, the model boils down to that considered in Su et al. (2016) and Lin and Ng (2012). By including a constant as one of the elements in x_{it}, the model can incorporate time-varying group specific fixed effects, and thus is related to Bonhomme and Manresa (2015). The extension of model (1) to allow for individual-specific fixed effects will be discussed in Section 4.1. Our model also extends Okui and Wang (2021) to allow not only the slope coefficients but also the group membership structure to change at the break point.

2.2 Estimation method

There are three types of parameters to estimate in model (1): the break point k, the group membership variable in pre- and post-break regimes $g_i(B)$ and $g_i(A)$ for all $i \in \{1, \ldots, N\}$, and the slope coefficients $\beta_{g(B),B}$ for all $g(B) \in \mathbb{G}^B$ and $\beta_{g(A),A}$ for all $g(A) \in \mathbb{G}^A$. We propose jointly estimating these three types of parameters by minimizing the quadratic loss function. Let β be a vector stacking $\beta_{g,B}$ for $g \in \mathbb{G}^B$ and $\beta_{g,A}$ for $g \in \mathbb{G}^A$. The parameter space for β is $\mathcal{B}$ which is a subset of $\mathbb{R}^{p(G^B+G^A)}$. Let $\Gamma = (\mathbb{G}^B \times \mathbb{G}^A)^N$ be the parameter space for group memberships. We denote an element of Γ as γ, and further denote $\gamma_B = (g_1(B), \ldots, g_N(B))$ and $\gamma_A = (g_1(A), \ldots, g_N(A))$ as the group membership vector before and after the break, respectively. Let $\mathbb{K} = \{2, \ldots, T\}$ be the parameter space for the break date k. We estimate (k, γ, β) by minimizing the least squares criterion:

$$(\hat{k}, \hat{\gamma}, \hat{\beta}) = \underset{k \in \mathbb{K}, \gamma \in \Gamma, \beta \in \mathcal{B}}{\operatorname{argmin}} \left[\sum_{t=1}^{k-1} \sum_{i=1}^{N} (y_{it} - x_{it}' \beta_{g_i(B),B})^2 + \sum_{t=k}^{T} \sum_{i=1}^{N} (y_{it} - x_{it}' \beta_{g_i(A),A})^2 \right]. \quad (2)$$

The least squares objective function offers a unified estimation framework for the three types of parameters and thus facilitates the theory. Least squares estimation of the group structure also guarantees that all units are categorized into one of the groups. Since the group membership structure may potentially change at the break point, we cannot employ the entire set of time periods to estimate this structure. The group membership structures before and after the break are estimated using only the corresponding samples of the two regimes. This is in sharp contrast to the Grouped Adaptive Group Fused Lasso (GAGFL) proposed by Okui and Wang (2021) for heterogeneous break estimation, where only coefficients have breaks and the full sample of time observations is used to estimate the time-invariant group structure. Moreover, we employ a least squares method to detect the break date, which fundamentally differs from the lasso break-detection techniques used in Okui and Wang (2021) and thus requires different theoretical analysis. Our estimation strategy also differs from Baltagi et al. (2017), who proposed estimating the break point based on the sum of squared residuals of each unit. Imposing a group structure allows us to take advantage of cross-sectional variation for coefficient estimation, and thus deliver more accurate estimated coefficients and further a more accurate break point estimate, especially when the error in each time series is large and the number of regressors is sizeable compared to the length of the time series. We shall compare these competing methods in simulation to confirm this discussion.

Clearly, a complete search of the parameter space for the three types of parameters is computationally infeasible. To solve this objective function, we propose the following algorithm that estimates the break point through concentration and estimates the group structure and coefficients through iteration.

Algorithm 1.

Let s denote the iteration number.

Step 1: Set $s = 0$. For each $k \in \{2, \ldots, T-1\}$, initialize group structures in both regimes as $\gamma_B^{(0)}$ and $\gamma_A^{(0)}$.

Step 2: For given $\gamma^{(s)}$ and k, estimate the slope coefficient $\beta^{(s)}$ in the two regimes by

$$\beta^{(s)} = \underset{\beta \in \mathcal{B}}{\operatorname{argmin}} \left[\sum_{t=1}^{k-1} \sum_{i=1}^{N} (y_{it} - x_{it}' \beta_{g_i^{(s)}(B),B})^2 + \sum_{t=k}^{T} \sum_{i=1}^{N} (y_{it} - x_{it}' \beta_{g_i^{(s)}(A),A})^2 \right].$$

Step 3: Given $\beta^{(s)}$, find the optimal group for individual i in each regime, respectively, by

$$g_i(B)^{(s+1)} = \underset{\gamma \in \Gamma}{\operatorname{argmin}} \sum_{t=1}^{k-1} (y_{it} - x_{it}' \beta_{g_i(B),B}^{(s)})^2, \quad \text{and} \quad g_i(A)^{(s+1)} = \underset{\gamma \in \Gamma}{\operatorname{argmin}} \sum_{t=k}^{T} (y_{it} - x_{it}' \beta_{g_i(A),A}^{(s)})^2.$$

Step 4: Iterate Steps 2 and 3 until numerical convergence, and obtain $\hat{\gamma}_B(k)$, $\hat{\gamma}_A(k)$, and $\hat{\beta}(k)$.

Step 5: Let k vary from 2 to T, and estimate the break point by

$$\hat{k} = \underset{k \in \mathbb{K}}{\operatorname{argmin}} \left[\sum_{t=1}^{k-1} \sum_{i=1}^{N} (y_{it} - x_{it}' \hat{\beta}_{\hat{g}_i(B,k),B}(k))^2 + \sum_{t=k}^{T} \sum_{i=1}^{N} (y_{it} - x_{it}' \hat{\beta}_{\hat{g}_i(A,k),A}(k))^2 \right],$$

where $\hat{g}_i(B,k)$ and $\hat{g}_i(A,k)$ are the i-th coordinates of $\hat{\gamma}_B(k)$ and $\hat{\gamma}_A(k)$, respectively.

Step 1 is an initialization. In Step 2, we estimate the coefficient vector given the group membership structure and the break date. It is just an OLS estimation applied to each group at each regime. Step 3 in turn estimates the group membership structure given coefficients and the break date. The iteration described in Step 4 is essentially the *Kmeans* algorithm discussed in Bonhomme and Manresa (2015). Because the objective function for a given k is non-increasing in the number of iterations, numerical convergence typically can be achieved quickly. However, the convergence is not theoretically guaranteed to be the global optimum in general, a common drawback of this type of iterative algorithms. Hence, different (random) initial values need to be tried and the one that yields the lowest objective is selected. The number of trial initial values depends on the computational capacity and the features of data, e.g., the scale of data set, the signal-to-noise ratio, the number of groups, etc. Finally, the break point is estimated in Step 5 by minimizing the sum of squared residuals over all $k \in \{2, \ldots, T\}$, a popular technique in the break detection literature (see, e.g. Bai, 1997).

The computational burden of this algorithm mainly comes from trying many initial values in the *Kmeans* part. Nonetheless, when the numbers of groups, G^B and G^A, are small, the algorithm is sufficiently fast and operational.

2.3 Empirical implementations

So far we have discussed estimation assuming that the number of groups is given and both the group structure and slope coefficients have a break. In practice, the number of groups is often unknown, and may also vary across regimes. Moreover, it is possible that the structural break affects only the slope coefficients, only the group memberships, or both, and it is often of empirical interest to identify which parameters are subject to the break. In this section, we propose to address both issues using the information criterion (IC), generally defined as

$$IC = \log \widehat{Q} + n_p f(N, T), \tag{3}$$

where $\widehat{Q}$ is the average sum of squared errors, and n_p is the total number of parameters, both of which depend on model specifications. $f(N, T)$ is a tuning parameter, and we find that $f(N, T) = 3\ln(NT)/NT$ works fairly well based on a large number of experiments with many alternatives. The following subsections will discuss how to obtain $\widehat{Q}$ and n_p in different situations.

2.3.1 Determining the number of groups

We first discuss how to determine the number of groups. Unlike Bonhomme and Manresa (2015) and Okui and Wang (2021), the complication here is that the number of groups may change after the structural break and thus we need to determine the number in each regime separately. The split-sample determination naturally requires the knowledge of the break point, which conversely depends on the number of groups. To address this challenge, we propose selecting the number of groups in the pre- and post-break regimes by minimizing the IC defined in (3), where

$$\widehat{Q}(G^B, G^A) = \frac{1}{NT}\left(\sum_{t=1}^{\hat{k}-1}\sum_{i=1}^{N}(y_{it} - x_{it}'\hat{\beta}_{\hat{g}_{i(B)},B})^2 + \sum_{t=\hat{k}}^{T}\sum_{i=1}^{N}(y_{it} - x_{it}'\hat{\beta}_{\hat{g}_{i(A)},A})^2\right),$$

with $\hat{k}$, $\hat{\gamma}$, and $\hat{\beta}$ obtained from (2) for a given set of (G^B, G^A). The total number of parameters $n_p(G^B, G^A)$ also depends on G^B and G^A, which sums the number of slope parameters over groups and regimes and the number of membership parameters to be estimated in the two regimes, namely $2N$.

To implement the proposed procedure, we first fix G^B and G^A and estimate the break point and the slope coefficients using the proposed algorithm, and then vary $G^A = 1, \ldots, G^A_{\max}$ and

$G^B = 1, \ldots, G^B_{\max}$ to compute the IC under each possible combination of G^B and G^A, where $G^A_{\max}$ and $G^B_{\max}$ are the (pre-specified) maximum numbers of groups. The selected numbers of groups thus minimize the IC defined in (3).[3] Since the standard IC can consistently select the number of groups within each regime given the true break date (Su et al., 2016; Liu et al., 2020), the proposed IC in (3) that essentially aggregates the standard IC across the two regimes is also expected to select the correct number of groups asymptotically.

2.3.2 Diagnosing the presence of structural breaks

This subsection discusses how to identify which parameters are subject to the break. We first consider selecting between the specifications of time-varying and time-invariant group structures. We propose using the IC as defined in (3), but the calculation of $\widehat{Q}$ and n_p needs to be tailored for these two specific alternatives. Let $\tilde{g}_i$ be the *time-invariant* group membership parameter of unit i. Denote by $\hat{\tilde{g}}_i$ the estimator of $\tilde{g}_i$ for $i = 1, \ldots, N$, and $\hat{\tilde{\gamma}} = (\hat{\tilde{g}}_1, \ldots, \hat{\tilde{g}}_N)$ the estimator of $\tilde{\gamma} = (\tilde{g}_1, \ldots, \tilde{g}_N) \in \mathbb{G}^N$. The associated estimator of k and β under the restriction of time-invariant group memberships is denoted by $\hat{\tilde{k}}$ and $\hat{\tilde{\beta}}$, respectively. We can obtain the estimated break point, group memberships, and slope coefficients under the restriction of time-invariant memberships by solving the following optimization:

$$(\hat{\tilde{k}}, \hat{\tilde{\gamma}}, \hat{\tilde{\beta}}) = \underset{k \in \mathbb{K}, \tilde{\gamma} \in \mathbb{G}^N, \beta \in \mathcal{B}}{\operatorname{argmin}} \left[\sum_{t=1}^{k-1} \sum_{i=1}^{N} (y_{it} - x'_{it}\beta_{\tilde{g}_i,B})^2 + \sum_{t=k}^{T} \sum_{i=1}^{N} (y_{it} - x'_{it}\beta_{\tilde{g}_i,A})^2 \right],$$

where $\beta_{\tilde{g}_i}$ is the group-specific slope parameter under the restriction of time-invariant memberships. To solve this optimization, we employ an iterative algorithm similar to the one in Section 2.2 but impose a restriction of time-invariant group memberships in initialization and the third step. The online supplement to this paper presents an algorithm for this model (Algorithm S.1).

With the group and regime specific coefficient estimates readily there, we can obtain the associated sum of squared residuals and compute the IC by (3) with the number of parameters $n_p = N + 2pG$ under time-invariant group structures and $n_p = 2N + p(G_A + G_B)$ under time-varying group structures. This IC works because if both slopes and memberships experience a break but one estimates a model assuming a time-invariant group structure, then neither group memberships nor the slope coefficients can be consistently estimated, leading to a poor fit captured by a high $\log \widehat{Q}$. In contrast, if the break only affects coefficients (but not memberships and thus of course not the number of groups), then the $\log \widehat{Q}$ estimates

[3]This method searches the optimal number of groups over all possible combinations of k, G^B, and G^A. Alternatively, one may consider incomplete search that only examines a subset of the parameter space, e.g. iterating between determination of the number of groups and break-point detection. Incomplete search can potentially reduce computational efforts, but at the cost of ending up with a local optimum.

obtained under time-varying and time-invariant group structures are close and thus one tends to choose the time-invariant group specification due to the difference in the penalty. Note that a Hausman-type of test is not applicable in this case because the coefficient estimators obtained under time-varying and time-invariant group structures share the same first order asymptotic behaviour, i.e., both are consistent and have the same asymptotic variance when the group structure is time invariant, due to the super-consistency of the group membership estimates.

Next, we consider selecting between the specifications of time-varying and time-invariant slope coefficients, i.e., whether a break only occurs in the group structures. In this case, a similar IC can be employed as above except that the number of parameters under time-invariant coefficient (but varying group structures) is $n_p = 2N + pG$. To estimate the time-invariant coefficients, we can modify Step 2 in Algorithm 1 and estimate the slope coefficient with the restriction that they are constant over time.

Finally, note that model diagnosis is implemented prior to estimation, a common procedure in frequentist econometrics. Once the model is specified, one can jointly estimate the structural break, group memberships, and slope coefficients as discussed in Section 2.2.

3 Theoretical results

This section presents the asymptotic properties of the proposed estimation method. In particular, we show that the estimated break point, group structure, and slope coefficients are consistent and that the asymptotic distribution of the coefficient estimator is identical to that under a known break point and group membership structure.

We use the following notation. Superscript 0, such as k^0, indicates the true value. β_B is the vector stacking $\beta_{g,B}$ for $g \in \mathbb{G}^B$. Similarly, β_A is the vector stacking $\beta_{g,A}$ for $g \in \mathbb{G}^A$. $\|\cdot\|$ denotes the Euclidean norm.

Assumption 1.

(i) For any $L \subseteq \{1, \ldots, N\}$ and $t'' \geq t'$, there exists M which does not depend on L, t'' nor t' such that the following equality holds

$$E\left(\left\|\frac{1}{NT}\sum_{t=t'}^{t''}\sum_{i\in L} x_{it}u_{it}\right\|^2\right) \leq M\frac{|L|(t''-t')}{N^2T^2},$$

where $|L|$ denotes the cardinality of L.

(ii) $\mathcal{B}$ is compact.

(iii) Let $\rho_{N,t}(\gamma^t, g, \tilde{g})$ *be the minimum eigenvalue of* $\sum_{i=1}^{N} \mathbf{1}\{g_{it}^0 = g\}\{g_{it} = \tilde{g}\} x_{it} x_{it}'/N$, *where* γ^t *is either* γ_B *(when* $t < k$*) or* γ_A *(when* $t \geq k$*). For any* $g \in \mathbb{G}^B$,

$$\min_{1 \leq t < k^0} \min_{\gamma_B \in (\mathbb{G}^B)^N} \max_{\tilde{g} \in \mathbb{G}^B} \rho_{N,t}(\gamma_B, g, \tilde{g}) > \hat{\rho},$$

and for any $g \in \mathbb{G}^A$,

$$\min_{k^0 < t \leq T} \min_{\gamma_A \in (\mathbb{G}^A)^N} \max_{\tilde{g} \in \mathbb{G}^A} \rho_{N,t}(\gamma_A, g, \tilde{g}) > \hat{\rho},$$

where $\hat{\rho} \to_p \rho$ *as* $N, T \to \infty$ *and* $\rho > 0$ *does not depend on* N *and* g.

(iv) There exists $\hat{\rho}^*$ *such that for any* i *and for* s *such that* s *and* $T - s$ *sufficiently large,*

$$\lambda_{\min}\left(\frac{1}{s}\sum_{t=1}^{s} x_{it} x_{it}'\right) \geq \hat{\rho}^* \qquad \text{and} \qquad \lambda_{\min}\left(\frac{1}{T-s}\sum_{t=s+1}^{T} x_{it} x_{it}'\right) \geq \hat{\rho}^*,$$

and $\hat{\rho}^* \to_p \rho^* > 0$ *as* $N, T \to \infty$, *where* $\lambda_{\min}$ *gives the minimum eigenvalue of its argument.*

(v) $\max_{1 \leq t \leq T} \sum_{i=1}^{N} \|x_{it}\|^2/N = O_p(1)$.

(vi) There exists a fixed constant $\underline{m} > 0$ *(which, in particular, does not depend on* T *and* N*) such that for any* t,

$$\frac{1}{N}\sum_{i=1}^{N} (x_{it}'(\beta^0_{g_i^0(A),A} - \beta^0_{g_i^0(B),B}))^2 > \underline{m}.$$

(vii) $k^0/T \to \tau \in (\epsilon, 1-\epsilon)$ *for* $\epsilon > 0$ *as* $T \to \infty$.

(viii) There exists a constant $c > 0$ *such that for any* $g \neq \tilde{g}$ *where* $g, \tilde{g} \in \mathbb{G}^B$ *and* $g', \breve{g} \in \mathbb{G}^A$, *it holds that* $\|\beta^0_{g,B} - \beta^0_{\tilde{g},B}\| > c$ *and* $\|\beta^0_{g',A} - \beta^0_{\breve{g},A}\| > c$.

(ix) Let z_{it} *be* $x_{it}'x_{it}$, $\|u_{it}x_{it}\|$, $2u_{it}x_{it}'(\beta^0_{g^0_{i(l)},l} - \beta^0_{g,l})$, *or* $(x_{it}'(\beta^0_{g^0_{i(l)},l} - \beta^0_{g,l}))^2$ *for* $g \in \mathbb{G}^l$ *and* $l = A, B$. *Assume the following holds for any choice of* z_{it}*: 1)* z_{it} *is a strong mixing sequence over* t *whose mixing coefficients* $a_i[t]$ *are bounded by* $a[t] \leq e^{-at^{d_1}}$ *such that* $\max_{1 \leq i \leq N} a_i[t] \leq a[t]$ *and has tail probabilities* $\max_{1 \leq i \leq N} \Pr(|z_{it}| > z) \leq e^{1-(z/b)^{d_2}}$ *for any* t, *where* a, b, d_1 *and* d_2 *are positive constants. 2) There exists* a_i, $i = 1, \ldots, N$ *such that for any* $\epsilon > 0$, *it holds that* $\max_{1 \leq i \leq N} |a_i - \sum_{t=1}^{T} E(z_{it})/T| < \epsilon$ *for sufficiently large* T.

(x) $\max_{1 \leq t \leq T} E(\|\sum_{i=1}^{N} x_{it}u_{it}/\sqrt{N}\|^{2+\delta})$ *is bounded for some* $\delta > 0$.

Assumption 1(i) concerns the degree of dependence. For example, this assumption is satisfied when (x_{it}, u_{it}) is independently and identically distributed (i.i.d.) and possesses fourth moments; it also allows for weak serial correlation and weak cross-sectional dependence. Assumption 1(ii) requires the compactness of the parameter space, a standard condition for asymptotic analysis of extremum estimators.

Assumptions 1(iii) and 1(iv) resemble the rank condition in ordinary least squares estimation that, roughly speaking, rules out cross-sectional multicollinearity in any group structure (Assumption 1(iii)) and over time periods for each unit (Assumption 1(iv)), and thus guarantees the identification of $\beta_{g,A}$ and $\beta_{g,B}$.[4] Note that Assumption 1(iii) requires that each group includes sufficiently many observations, so that $\sum_{i=1}^{N} \mathbf{1}\{g_{it}^0 = g\}/N$ does not degenerate. Empirically, this rules out the case where a group consists of only a few units. Moreover, this assumption also requires that there is no cross-sectional multicollinearity among units in each group. Assumption 1(v) excludes the presence of outliers in x_{it}.

Assumption 1(vi) ensures the identification of the break point. It implies that the value of the coefficient vector and/or the group membership structure changes sufficiently at the break point. Importantly, this assumption allows the situations in which the group membership structure changes while the coefficient vector for each group does not change, which has not been considered previously in the literature, to our knowledge. This assumption also covers the cases where only the coefficient vector changes but not group memberships as considered in Okui and Wang (2021). Assumption 1(vii) rules out the possibility that the break occurs in the very beginning or the very end of the sample period, so that there are sufficiently long (at least asymptotically) time series both before and after the break. This condition is often imposed for detecting breaks with time series data, but not with panel data (see, e.g. Bai, 2010; Qian and Su, 2016; Okui and Wang, 2021). It is required here in order to identify *time-varying* group membership structures, which requires a sufficiently large time dimension in both regimes.

Assumption 1($viii$) states that the coefficients in the two different groups are sufficiently different, often called the "group separation" condition. It is used to identify the group membership structure. Assumption 1(ix) contains technical conditions on the mixing and tail properties of various random objects appearing in the theorems. It is used to show the consistency of the group membership assignments. Technically, we can relax Assumptions 1(vi), 1($viii$) and 1(ix) and still obtain the consistency of break date, group membership, and

[4]We make the assumptions on ρ rather than on $\hat{\rho}$ (and on ρ^* rather than on $\hat{\rho}^*$) for a technical reason that there are practically relevant cases in which $\hat{\rho} > 0$ may not hold while $\rho > 0$ does. For example, when x_{it} is a binary random variable, there exists a tiny yet positive probability that x_{it}'s are identical for all i and t. Thus $\hat{\rho} > 0$ does not hold but $\rho > 0$ may be assumed.

coefficient estimators as well as the asymptotic distribution of coefficient estimator. However, the current assumptions allow us to obtain the asymptotic results under a weak and easy-to-understand condition on the relative magnitude of N and T. In Appendix A.3, we consider cases in which the break size shrinks (i.e., $\underline{m}$ tends to 0 asymptotically), group separation is not perfect asymptotically (i.e., c tends to 0 asymptotically), and the variables satisfy weaker mixing and moment conditions than those in Assumption 1(ix) but under strict stationarity of regressors and the error term (but not stationarity of the dependent variable).

Assumption 1(x) imposes a condition on the existence of the moments of $x_{it}u_{it}$, which is used to bound mixingale sequences.

Theorem 1. *Suppose that Assumption 1 holds. As $N, T \to \infty$ with $NT^{-\delta} \to 0$ for some $\delta > 0$, $\Pr(\hat{k} = k^0) \to 1$.*

This theorem establishes the consistency of the break point estimator. Note that this is a "super" consistency result in the sense that the probability of the break point estimate exactly equals the true break point with probability approaching one. Next, we examine the the properties of group membership and slope coefficient estimators.

Corollary 1. *Suppose that Assumption 1 holds. As $N, T \to \infty$ with $NT^{-\delta} \to 0$ for some $\delta > 0$,*

(1) $\Pr(\hat{\gamma} = \gamma^0) \to 1$,

(2) $\hat{\beta} = \tilde{\beta} + o_p(1/\sqrt{NT})$, *where $\tilde{\beta}$ is the estimator of β under $k = k^0$ and $\gamma = \gamma^0$.*

The first statement of this corollary shows that the group membership structure can be estimated consistently.[5] This is also a super consistency result as in the case of break point. These two super-consistency results imply the second conclusion of the corollary that the coefficient estimator asymptotically behaves as if the break point and the group membership structure were known. Because $\tilde{\beta}$ is the ordinary least squares estimator applied to each regime and each group, its asymptotic distribution is well-known and standard statistical inference applies.

[5] The super-consistency of group membership estimators is a strong result that guarantees asymptotic equivalence of estimated coefficients under unknown and true memberships. In fact, it is possible to derive the asymptotic distribution of $\hat{\beta}$ under a weaker set of assumptions. Dzemski and Okui (2020) prove the asymptotic normality of the coefficient estimator without the uniform super-consistency of group membership estimators in a simple model that contains only an intercept and no structural breaks. However, the proof becomes much more involved even in that simple setting. Thus, to avoid complicating the theories and stay focus on our goal of detecting membership breaks, we keep our assumptions strong enough to achieve the super-consistency of $\hat{\gamma}$.

Note that $T \to \infty$ and a sufficient number of time periods in each of the regimes are requisite to identify the group membership structure and to achieve the consistency of clustering, based on which the consistency of the break date and coefficient estimators is established. This implies that the estimation technique in (2) does not allow for consecutive or end-of-sample breaks, a restriction that is common to much of the structural break literature. While sufficient time series observations are required, we allow N to be much larger than T since our theoretical results hold under $NT^{-\delta} \to 0$ with δ being any arbitrary positive constant. As observed in Bonhomme and Manresa (2015), such a weak condition on the relative magnitude of N and T is an advantage of imposing a group pattern of heterogeneity. It is much weaker than those used in other papers on heterogenous panel data, such as $N/T \to 0$ and $N/T^2 \to 0$ as used in Okui and Yanagi (2019).

4 Extensions

4.1 Models with individual-specific fixed effects

So far we have considered the model with group-specific fixed effects. In many cases, it is desirable to capture individual unobserved heterogeneity, and thus we consider the following model:

$$y_{it} = \alpha_i + x_{it}'\beta_{g_{it},t} + u_{it}, \tag{4}$$

where α_i are individual effects that can be arbitrarily correlated with covariates; the specification of other quantities remains mostly identical to model (1), but we primarily consider cases of strict exogeneity $E(u_{it} \mid \alpha_i, \{x_{it}\}_{t=1}^{T}) = 0$. In this subsection, we suppose that the panel data starts at $t = 0$ for notational convenience.

4.1.1 Estimation method

To eliminate the fixed effects, we take the first difference. The transformed variables satisfy:

$$\begin{aligned}\Delta y_{it} =& x_{it}'\beta_{g_{it},t} - x_{i,t-1}'\beta_{g_{i,t-1},t-1} + \Delta u_{it} \\ =& \begin{cases} \Delta x_{it}'\beta_{g_i(B),B} + \Delta u_{it} & \text{if } t < k^0 \\ x_{it}'\beta_{g_i(A),A} - x_{i,t-1}'\beta_{g_i(B),B} + \Delta u_{it} & \text{if } t = k^0 \\ \Delta x_{it}'\beta_{g_i(A),A} + \Delta u_{it} & \text{if } t > k^0 \end{cases},\end{aligned}$$

where Δ is the first difference operator, for example, $\Delta y_{it} = y_{it} - y_{i,t-1}$. First differencing is convenient in our setting because Δy_{it} depends on the coefficients in both regimes, namely $\beta_{g_i(B),B}$ and $\beta_{g_i(A),A}$, only at $t = k^0$. In contrast, within transformation, an alternative method to remove fixed effects, yields dependent variables that depend on coefficients of both regimes at all time periods, rendering the transformed model more difficult to analyze.

As above, we estimate the coefficients, group membership structure, and the break date by minimizing the quadratic loss function. In particular, we estimate (k, γ, β) by minimizing the least squares criterion:

$$
\begin{aligned}
(\hat{k}, \hat{\gamma}, \hat{\beta}) = \underset{k \in \mathbb{K}, \gamma \in \Gamma, \beta \in \mathcal{B}}{\operatorname{argmin}} \Bigg[& \sum_{t=1}^{k-1} \sum_{i=1}^{N} (\Delta y_{it} - \Delta x_{it}' \beta_{g_i(B),B})^2 \\
& + \sum_{i=1}^{N} (\Delta y_{ik} - x_{ik}' \beta_{g_i(A),A} + x_{i,k-1}' \beta_{g_i(B),B})^2 \\
& + \sum_{t=k+1}^{T} \sum_{i=1}^{N} (\Delta y_{it} - \Delta x_{it}' \beta_{g_i(A),A})^2 \Bigg].
\end{aligned} \tag{5}
$$

To solve the optimization problem, we can employ a similar algorithm as in Section 2 but replace the objective function with (5).

4.1.2 Asymptotic properties

To show the asymptotic properties of the estimators in the presence of fixed effects, a similar set of assumptions to those in Section 3 is needed, but some of the assumptions need to be adjusted to incorporate the first-differenced data. Specifically, we keep Assumptions 1(*ii*), 1(*vii*) and 1(*viii*) and the other parts of Assumption 1 are modified as follows:

Assumption 2.

(i) For any $L \subseteq \{1, \ldots, N\}$ and $t'' \geq t'$, there exists M which does not depend on L, t'' nor t' such that the following equality holds

$$
E\left(\left\| \frac{1}{NT} \sum_{t=t'}^{t''} \sum_{i \in L} x_{it+l} u_{it+j} \right\|^2 \right) = M \frac{|L|(t'' - t')}{NT^2}. \tag{6}
$$

for $l, j = 0, -1$, where $|L|$ is the cardinality of L.

(iii) Let $\rho_{D,N,t}(\gamma^t, g, \tilde{g})$ be the minimum eigenvalue of $\sum_{i=1}^{N} \mathbf{1}\{g_{it}^0 = g\}\{g_{it} = \tilde{g}\} \Delta x_{it} \Delta x_{it}'/N$, where γ^t is either γ_B (when $t < k$) or γ_A (when $t \geq k$). For any $g \in \mathbb{G}^B$,

$$
\min_{1 \leq t < k^0} \min_{\gamma_B} \max_{\tilde{g} \in \mathbb{G}^B} \rho_{D,N,t}(\gamma_B, g, \tilde{g}) > \hat{\rho}_D,
$$

and for any $g \in \mathbb{G}^A$,

$$
\min_{k^0 < t \leq T} \min_{\gamma_A} \max_{\tilde{g} \in \mathbb{G}^A} \rho_{D,N,t}(\gamma_A, g, \tilde{g}) > \hat{\rho}_D,
$$

where $\hat{\rho}_D \to_p \rho_D$ and $\rho_D > 0$ does not depend on N and g.

(iv) There exists $\hat{\rho}_D^*$ such that for any i and for s such that s and $T-s$ sufficiently large,

$$\lambda_{\min}\left(\frac{1}{s}\sum_{t=1}^{s}\Delta x_{it}\Delta x_{it}'\right) \geq \hat{\rho}_D^* \qquad \text{and} \qquad \lambda_{\min}\left(\frac{1}{T-s}\sum_{t=s+1}^{T}\Delta x_{it}\Delta x_{it}'\right) \geq \hat{\rho}_D^*,$$

and $\hat{\rho}_D^* \to_p \rho_D^* > 0$.

(v) $\max_{1\leq t\leq T}\sum_{i=1}^N \|x_{it}\|^2/N = O_p(1)$ and $\max_{1\leq t\leq T}\sum_{i=1}^N \|\Delta x_{it}\|^2/N = O_p(1)$.

(vi) There exists a fixed constant $\underline{m} > 0$ independent on T and N, such that for any t,

$$\frac{1}{N}\sum_{i=1}^{N}(\Delta x_{it}'(\beta^0_{g_i^0(A),A} - \beta^0_{g_i^0(B),B}))^2 > \underline{m}.$$

(ix) Let z_{it} be $\Delta x_{it}'\Delta x_{it}$, $\|\Delta u_{it}\Delta x_{it}\|$, $2\Delta u_{it}\Delta x_{it}'(\beta^0_{g^0_{i(l)},l} - \beta^0_{g,l})$ or $(\Delta x_{it}'(\beta^0_{g^0_{i(l)},l} - \beta^0_{g,l}))^2$, for $g \in \mathbb{G}_l$ and $l = A, B$. Assume the following holds for any choice of z_{it}: 1) z_{it} is a strong mixing sequence over t whose mixing coefficients $a_i[t]$ are bounded by $a[t] \leq e^{-at^{d_1}}$ such that $\max_{1\leq i\leq N} a_i[t] \leq a[t]$ and has tail probabilities $\max_{1\leq i\leq N}\Pr(|z_{it}| > z) \leq e^{1-(z/b)^{d_2}}$ for any t where a, b, d_1 and d_2 are positive constants. 2) There exists a_i, $i = 1, \ldots, N$ such that for any $\epsilon > 0$, it holds that $\max_{1\leq i\leq N}|a_i - \sum_{t=1}^T E(z_{it})/T| < \epsilon$ for T sufficiently large.

(x) $\max_{1\leq t\leq T} E(\|\sum_{i=1}^N x_{it}u_{it+l}/\sqrt{N}\|^{2+\delta})$ is bounded for some $\delta > 0$ where $l = 0, -1$.

With these modified assumptions, we can establish exactly the same asymptotic properties of the estimators of the break point, group memberships, and slope coefficients as stated in Theorem 1 and Corollary 1. Note that Assumption 2(i) excludes predetermined regressors such as lagged dependent variables in model (4). If x_{it} contains predetermined regressors, then $E(x_{it}u_{i,t-1}) \neq 0$ and the order of the left side of (6) would be at least $O(1)$, violating the order required in the assumption. In contrast, this assumption would be satisfied, for example, if x_{it} is strictly exogenous.

4.2 Multiple breaks

Our framework can allow the group structure and slope coefficients to experience a break at different time points when the method is extended to multiple breaks, because these two different changes can be modelled by two breaks. When we have m structural breaks at time points $k_1^0, \ldots, k_m^0$, the coefficient vector can be written as:

$$\beta_{g_{it},t} = \begin{cases} \beta_{g_i(1),1} & \text{if } t < k_1^0, \\ \beta_{g_i(2),2} & \text{if } k_1^0 \leq t < k_2^0, \\ \vdots & \\ \beta_{g_i(m),m} & \text{if } t \geq k_m^0. \end{cases}$$

To estimate multiple breaks, we can consider either the simultaneous approach (Bai and Perron, 1998) or the sequential approach discussed by Bai (2010) and Baltagi et al. (2016) in the panel data context. Both approaches are based on the least squares objective function, with the former simultaneously estimating all break points and the latter estimating the break points one at a time. More specifically, assuming that the number of breaks is known, the simultaneous approach first computes the sums of squared residuals of the relevant segments of the time periods, and then searches for the partition that achieves a global minimization of the overall sum of squared residuals using dynamic programming. Because the estimation of latent group structures requires a sufficiently large number of time observations, some partitions with short time periods in a regime need to be excluded, rendering this simultaneous approach computationally more difficult. In contrast, the sequential approach is computationally more attractive because it proceeds as if there were only one break each time. It detects the first break point $\hat{k}_1$ (not necessarily corresponding to k_1^0) by solving (2) over the entire time period. Then it splits the time period into two regimes at $\hat{k}_1$, and estimates a break point, again by estimating a similar objective function as (2), in each of the two regimes respectively. The second break point $\hat{k}_2$ is the one (out of the two newly estimated points within the two regimes) that leads to a larger reduction in the sum of squared residuals. One can repeat such a procedure until m breaks are obtained. Our preliminary study suggests that this sequential procedure will work in detecting multiple break points and identifying the latent group structures in each regime, but a comprehensive theoretical analysis is reserved for future investigation.

4.3 Models with partially homogeneous and time-invariant coefficients

Model (1) can be extended to allow a subvector of coefficients to be time-invariant and homogeneous. Without loss of generality, let the first k_1 explanatory variables $x_{1,it}$ have time-invariant and homogeneous coefficients β_1, and the remaining covariates $x_{2,it}$ correspond to heterogeneous time-varying coefficients $\beta_{2,g_{it},t}$. The objective function in this case becomes

$$(\hat{k}, \hat{\gamma}, \hat{\beta}) = \underset{k \in \mathbb{K}, \gamma \in \Gamma, \beta \in \mathcal{B}}{\operatorname{argmin}} \left[\sum_{t=1}^{k-1} \sum_{i=1}^{N} (y_{it} - x_{1,it}' \beta_1 - x_{2,it}' \beta_{2,g_i(B),B})^2 \right.$$

$$+\sum_{t=k}^{T}\sum_{i=1}^{N}(y_{it}-x'_{1,it}\beta_1-x'_{2,it}\beta_{2,g_i(A),A})^2\Bigg]. \quad (7)$$

To find a minimizer of this objective function, we adjust the iterative algorithm by separating the estimation of β_1 and $\beta_{2,g,t}$ such that an estimate of β_1 is estimated using all the observations in Step 2 of Algorithm 1, and the group membership update in Step 3 is based on the adjusted objective function (7).

5 Simulation study

This section evaluates the finite sample performance of the proposed method. First, we compare three methods of detecting heterogeneous breaks when the number of groups in each regime is known. The first method is what we propose in this paper which simultaneously estimates the group and break points using (2). The second method first detects break points based on individual estimation as in Baltagi et al. (2016) and then classifies units given the estimated breaks. The final method is GAGFL proposed by Okui and Wang (2021). Next, we examine the performance of the proposed procedure in determining the numbers of groups, and in this part of the simulation study we explicitly allow that the number of groups and hence group membership may change after the structural break.

5.1 Estimation under a given number of groups

In this section, we first evaluate the performance of the three methods when the number of groups in each regime is given. Our benchmark designs consist of four data generation processes (DGPs), each of which contains three sub-cases that differ in the type of breaks: (1) a structural break only in the magnitude of the group-specific slope coefficients, (2) a structural break only in the group memberships, and (3) a structural break (at the same point in time) in both the coefficients and group memberships. We also consider various extensions from the benchmark to examine how the performance of the methods varies in different situations, and the details are provided in the online supplement.

5.1.1 Data generation processes

The four benchmark DGPs are as follow:

DGP 1 [Static panel]: Our baseline case considers the following model

$$y_{it} = x'_{it}\beta_{g_{it},t} + u_{it},$$

where x_{it} is a $p \times 1$ vector that contains as its first element a constant and the other $p-1$ nonconstant regressors generated from $N(0, I_{p-1})$ with I_{p-1} being an identity matrix of dimension $p-1$. We set $p = 6$ to illustrate the case where the length of time series is not dominantly larger than the number of parameters for each individual unit. We consider two groups. Let N_j, $j = 1, 2$, denote the number of units in group j with $N = N_1 + N_2$. The group membership and slope coefficients are both allowed to change after a structural break at time $k^0 = \lfloor 0.7T \rfloor$, where $\lfloor \cdot \rfloor$ takes the integer part. The error term u_{it} follows a standard normal distribution.

DGP 2 [AR error]: The same as DGP 1 except that the error term is autoregressive as $u_{it} = \rho u_{it-1} + \epsilon_{it}$, where $\rho = 0.6$ and ϵ_{it} follows a standard normal distribution.

DGP 3 [Individual FE]: We consider the model

$$y_{it} = \alpha_i + x_{it}' \beta_{g_{it},t} + u_{it},$$

where α_i follows a standard normal distribution, and $x_{it} = \alpha_i + z_{it}$ with z_{it} being a $p \times 1$ vector of standard normally distributed variables. The remaining specification is the same as DGP 1.

DGP 4 [Dynamic panel]: The same as DGP 1 except that the regressors contain a lagged dependent variable y_{it-1}.

For each of these four DGPs, we consider three sub-cases, depending on which parameters contain structural breaks.

DGP X.1 This case permits a structural break only in the slope coefficients. We fix the ratio of units among groups as $N_1 : N_2 = 0.4 : 0.6$, and the group membership does not change after the structural break. The coefficients in the first group exhibit a structural break as

$$\beta_{1,t} = \begin{cases} \beta_{1,B} = \iota_p & \text{if } t < k^0 \\ \beta_{1,A} = 2\iota_p & \text{if } t \geq k^0 \end{cases},$$

where ι_p is a $p \times 1$ vector of ones. Units in the second group do not exhibit breaks, and the slope coefficient in this group is given by $\beta_{2,t} = 0.5\iota_p$.

DGP X.2: In this case, group memberships change after the break but the slope coefficients do not. The ratio of units among groups is $N_1 : N_2 = 0.4 : 0.6$ before the break, and $N_1 : N_2 = 0.6 : 0.4$ after the break. We generate the group memberships before and after the break independently. For simplicity and clarity, we assign the first

40% of units to Group 1 before the break and the first 60% to the same group after the break, such that 20% of units change their group memberships. This way of generating memberships allows us to easily vary the proportion of units that change memberships by changing the ratio of units among groups. The slope coefficients are $\beta_{1,t} = \iota_p$ in Group 1 and $\beta_{2,t} = 0.5\iota_p$ in Group 2.

DGP X.3: In this case, both the slope coefficients of each group and the group structure change after the break. The slope coefficients are the same as in DGP X.1, while the group structure is the same as DGP X.2.

We therefore have 12 cases in total. We further consider the cross sectional dimensions $N = (100, 200)$ and the time series dimensions $T = (10, 20)$, leading to four combinations of sample sizes. In all cases, the simulation is conducted based on 1000 replications.

5.1.2 Implementation and evaluation

We compare the proposed method with GAGFL (Okui and Wang, 2021) and break detection methods based on individual estimation, denoted as BFK (Baltagi et al., 2016). We call our proposed method the Least Squares estimator for models with Group structure and structural Break (LSGB). The LSGB method detects the break point and cluster units jointly by minimizing the objective function (2).

GAGFL assumes that the group memberships do not change after structural breaks. It estimates the group memberships, break points, and slope coefficients simultaneously by minimizing the following objective function

$$(\hat{\beta}, \hat{\gamma}) = \underset{(\beta,\gamma)\in\mathcal{B}^{GT}\times\mathbb{G}^N}{\operatorname{argmin}} \left[\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - x'_{it}\beta_{g_i,t})^2 + \lambda \sum_{g\in\mathbb{G}} \sum_{t=2}^{T} \dot{w}_{g,t} \left\| \beta_{g,t} - \beta_{g,t-1} \right\| \right],$$

where λ is a tuning parameter typically chosen by an IC (see, e.g., Qian and Su, 2016) and $\dot{w}_{g,t}$ is a data-driven weight constructed based on a preliminary consistent estimate of β (see Okui and Wang (2021) for details). To apply this method when group membership is allowed to change after the break, we need to segment a homogeneous group if some of its members shift to another group in a different regime, such that each group only contains units that do not change memberships in all regimes. Particularly, in our DGP X.2 and DGP X.3, $X = 1, \ldots, 4$, we need to impose 3 groups for GAGFL to assure no units change memberships.[6]

[6] Since we generate the memberships with the first 40% of units in Group 1 before the break and the first 60% in the same group after the break, we have the first 40% and last 40% of units who do not change memberships while the middle 20% do change. The number of groups needed for GAGFL to assure time-invariant memberships depends on the minimum number of groups in each regime and how units change memberships.

Unlike the other two methods, BFK first detects the break point by estimating

$$\hat{k} = \arg\min_{k} \min_{\beta_i, \delta_i} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[y_{it} - x_{it}'\beta_i - x_{it}'\delta_i \mathbf{1}(t \geq k) \right]^2,$$

and then for each stable regime estimates the group membership structure γ_A and γ_B using least squares estimation, and obtains group-specific slope coefficients. Note that (Baltagi et al., 2016) do not consider group structure and strictly speaking, the BFK method we consider is a combination of the break detection by (Baltagi et al., 2016) and the group fixed effect method by Bonhomme and Manresa (2015).

We evaluate the three methods based on the accuracy of break point detection, group assignment, and coefficient estimates. First, we measure the accuracy of break point estimates based on their Hausdorff distance (HD), which collapses to the absolute distance between the true and estimated break points in the case of one break as

$$\mathrm{HD}(\widehat{k}, k^0) \equiv |k - k^0|.$$

We report the distance as a percentage of the sample, that is, multiplied by 100 and divided by T, i.e. $100 \times \mathrm{HD}(\widehat{k}, k^0)/T$, averaged across the replications. We also report the average estimated break points across replications, $\bar{k}$. For GAGFL, we report the HD and $\bar{k}$ conditional on the correct estimation of the number of breaks.

Second, the clustering accuracy is measured by the average of the misclustering frequency $(\widehat{g}_i \neq g_i^0)$ across replications. Let $I(\cdot)$ be the indicator function. The misclustering frequency (MF) is the ratio of misclustered units to the total number of units, i.e.

$$\mathrm{MF} = \frac{1}{N} \sum_{i=1}^{N} I(\widehat{g}_i \neq g_i^0).$$

The average misclustering frequency (across replications) before and after the break is reported separately, denoted by MF_B and MF_A for the three methods.

Finally we evaluate the accuracy of coefficient estimates by overall mean square error (MSE) averaged across replications as

$$\mathrm{MSE}(\widehat{\beta}_{it}) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (\widehat{\beta}_{it} - \beta_{it})^2,$$

where $\widehat{\beta}_{it}$ and β_{it} are the estimated and true values of β for unit i at period t. Since the MSE here is computed based on coefficient estimates for each unit at each time, this measure also reflects the error of clustering and break point detection.

5.1.3 Results

We first examine the accuracy of break point detection of the three methods. Table 1 presents the Hausdorff distance and average estimated break point. In almost all cases, the proposed LSGB produces more accurate break point estimates than GAGFL and BFK. LSGB outperforms GAGFL because the latters segmentation into homogenous groups (to ensure time invariant memberships) leads to rather inefficient coefficient estimates, and further produces less accurate break point estimates. The degree of efficiency loss depends on the amount of "unnecessary" separation. In these benchmark cases, GAGFL only segments three groups, while the true number of groups in both regimes is two. When the cross-section dimension is large, GAGFL often performs as equally well as LSGB. The advantage of LSGB is demonstrated when N is not large, especially when only group memberships change after the break (DGP X.2). We also consider the case with more groups, and the results (see the online supplement) show that the efficiency loss of GAGFL can be quite substantial if the number of groups increases. Although GAGFL performs worse than LSGB in some cases, it still has its own advantages. For example, it can handle multiple breaks more conveniently than LSGB and can incorporate end-of-sample or consecutive breaks which cannot be accommodated by LSGB since consistent group membership estimation of LSGB requires a sufficiently long time series dimension.[7] Moreover, we find that GAGFL also outperforms BFK in all DGPs since it makes use of cross-sectional variation, in line with the findings of Okui and Wang (2021).

Comparing LSGB with BFK, the former performs significantly better because it imposes a group structure, making use of cross-sectional variation, and therefore produces more efficient estimation than individual time series estimation. Moreover, the poor performance of BFK is also due to the fact that not all individuals have structural breaks (see also Baltagi et al., 2016; Okui and Wang, 2021). As N increases, the accuracy of the LSGB break point estimates improves remarkably because there are more units in each group, providing more cross-sectional variation for estimation. In contrast, the accuracy of the BFK break point estimate is hardly affected by an increase of N since it is based on separate individual estimation. As T increases, we find that the performance of both methods improves. The Hausdorff distance of the BFK break point estimates is roughly halved when T is doubled, while that of LSGB is reduced even more dramatically, confirming the super-consistency of the group membership estimators. The average break point estimates show that the BFK approach tends to estimate the break point in the middle of the time period. This is expected because individual time series estimation can be highly inefficient. Splitting the sample roughly equally provides sufficient samples for both regimes, and therefore leads to a smaller sum of

[7] GAGFL can incorporate end-of-sample breaks because it employs the entire time period for group membership estimation; see Okui and Wang (2021) for details.

squared residuals than splitting at the true break point which leads to too small a post-break sample. In contrast, LSGB makes use of the cross section samples, and thus avoids the problem caused by an insufficient number of time observations after the break. More results on how the performance of BFK and LSGB depends on the sample sizes in the two regimes are provided in the online supplement.

We further examine the performance of LSGB in each DGP in detail. In DGP 1, LSGB produces highly accurate break point estimates with the Hausdorff distance close to zero in all cases. When the error is serially correlated as in DGP 2, the performance of LSGB is slightly affected, but the method can still detect break points fairly accurately. In DGP 3 with individual fixed effects, first differencing largely reduces the accuracy of LSGB, but it improves quickly as T increases. Including predetermined lagged dependent variables does not seem to deteriorate the performance of LSGB at all. Interestingly, it appears that breaks only in group memberships are more difficult to detect than breaks in the slope coefficients. A possible reason is that breaks in group membership with 20% of the units changing their memberships correspond to a moderate size of coefficient breaks in a fixed-group context. This is partly confirmed by increasing the percentage of units that change memberships, and we indeed obtain more accurate break point estimates.

INSERT TABLE 1 HERE

Next, we compare the group membership estimates of the three methods and present the misclustering frequency in Table 2. GAGFL often produces the lowest misclustering frequency when slope coefficients exhibit a break, because it employs the entire time period for clustering, while LSGB and BFK estimate memberships only using time observations in each regime. However, when only group memberships exhibit a break (DGP X.2), GAGFL is outperformed by LSGB. Comparing LSGB with BFK, the clustering produced by LSGB is more accurate than that by BFK in most cases due to more accurate break point estimation. This is expected because the slope estimates of BFK are contaminated by incorrect break point estimation, which further affects the clustering. In DGP 1.1, LSGB leads to much lower misclustering frequency than BFK before the break, but not after the break. The poorer performance of LSGB after the break is because LSGB correctly detects the break point that lies close to the end of the time period, leaving only a short time dimension after the break for group identification. On the contrary, BFK incorrectly estimates the break point around the middle of the time period, leading to a longer post-break time dimension. In DGP 1.2, LSGB again produces more accurate clustering than BFK before the break. The performance of the two methods is similar after the break, and LSGB even slightly outperforms BFK when

$T = 20$. In DGP 1.3, LSGB dominates BFK both before and after the break for all sample sizes. This suggests that the incorrect break point estimation contaminates the group membership estimates to a larger extent when group memberships exhibit a break. When we allow for serial correlation in errors in DGP 2, all methods produce less accurate clustering as expected. Like in DGP 1, we find that when only the coefficients exhibit a break, LSGB clustering is more accurate than BFK in the pre-break regime. When the group memberships exhibit a break, LSGB performs equally well or better in both regimes. Similar results appear in DGP 3 where we allow for individual fixed effects and estimate using first-differenced data. In DGP 4 where lagged dependent variables are included, the impact of incorrect break point estimates on clustering is larger, and thus LSGB outperforms BFK in clustering in all cases, regardless of the type of break and sample sizes. To better appreciate the super-consistency of group membership estimates, we also compute the percentage of perfect clustering across replications (results available upon request). We find that the percentage indeed improves exponentially as T increases, providing strong evidence of super-consistency.

INSERT TABLE 2 HERE

Finally, we compare the average MSE of the coefficient estimates (across replications) produced by the three methods in Table 3. Despite accurate group assignment, the average MSE of GAGFL is generally larger than that of LSGB except in a few cases of DGP X.1 (where only the coefficients break). The efficiency loss of GAGFL is particularly sizeable when group memberships change after the break (DGP X.2 and X.3) and in the models with individual fixed effects (DGP 3) or lagged dependent variables (DGP 4). BFK performs even worse than GAGFL. Its MSE is at least twice as large as that of LSGB in all cases, and sometimes (i.e., DGP 4.1 and 4.3) even explodes due to incorrect break point and group membership estimation as in some cases of dynamic panels.

INSERT TABLE 3 HERE

We also consider alternative settings of coefficients to generate lower signal-to-noise ratios, less sizeable breaks, and more alike groups. Our method can still correctly detect the break point and cluster units under reasonably low signal-to-noise ratios (see the online supplement). The accuracy of break point estimates and clustering mainly depends on the size of break and the degree of group separation, while the signal-to-noise ratio mainly affects the accuracy of the coefficient estimates.

5.2 Determining the number of groups

So far, the simulation study focused on the cases when the number of groups is known. When this number is not a priori knowledge, as in most applications, we propose an IC-based method to determine the number in each regime (see Section 2.3.1). Here we investigate the performance of this approach in finite samples. We primarily focus on the case when the number of groups changes after the structural break.[8]

We consider four data generating processes similar to DGPs 1–4 as above except that we set the number of groups before the break to $G^B = 2$ and after the break to $G^A = 3$. In each DGP, we modify DGP X.2 and X.3 to allow the number of groups to change after the break as follows

DGP X.2': Only the group membership changes. The size of groups before the break is $N_1 : N_2 = 0.4 : 0.6$ as above, but after the break and $N_1 : N_2 : N_3 = 0.3 : 0.3 : 0.4$. The slope coefficients before the break are $\beta_{1,B} = \iota_p$ in Group 1 and $\beta_{2,B} = 0.5\iota_p$ in Group 2, and those after the break are $\beta_{1,A} = \iota_p$ in Group 1, $\beta_{2,A} = 0.5\iota_p$ in Group 2, and $\beta_{3,A} = 2\iota_p$ in Group 3.

DGP X.3': Both the slope coefficient of each group and the group structure change after the break. The group structure is the same as DGP X.2' above. The slope coefficients before the break are $\beta_{1,B} = 1.5\iota_p$ in Group 1 and $\beta_{2,B} = 0.5\iota_p$ in Group 2, and those after the break are $\beta_{1,A} = 2.5\iota_p$ in Group 1, $\beta_{2,A} = 0.5\iota_p$ in Group 2, and $\beta_{3,A} = 3.5\iota_p$ in Group 3.

Table 4 presents the empirical probability of selecting a particular number of groups in each regime with the possible number ranging from 1 to 4.[9] Recall that the true number is $G^B = 2$ before the break and $G^A = 3$ after the break. The proposed method generally performs well in determining the number of groups in the two regimes except when T is particularly small. Unsurprisingly, when the groups are not well separated (DGP X.2') and the sample is small ($N = 100$ and $T = 10$), the method tends to underestimate the number of groups. Otherwise, it can identify the correct number of groups with a high probability in most of cases, though it sometimes slightly overestimates the number (e.g., DGP 3.2 and

[8] The case of time invariant number of groups is a special case of the current setup, and the performance of IC when the number of groups is time invariant has been studied in other papers, e.g. Bonhomme and Manresa (2015) and Okui and Wang (2021).

[9] Allowing for a larger set of possible numbers hardly affects the results for almost all cases. The only exception occurs in the post-break regime in DGP 3, where the length of time span after the break is short and first-differencing is applied. In that case, increasing the maximum number of groups leads to more severe overspecification with the probability on each selected number being more dispersed.

3.3 after the break). Since overspecification only lowers the efficiency but does not affect the consistency of the coefficient estimates, our IC-based method provides a satisfactory tool to guarantee consistent coefficient estimates in most cases. More specifically, in DGP X.2' where only the group membership changes, our method can guarantee not to underestimate the number of groups as long as T is not too small. When the groups are relatively well separated as in DGP X.3', the IC-based method can correctly identify the group number in both regimes with a high probability in 10 out of 16 cases. In the remaining 6 cases (2 of DGP 2.3' and 4 of DGP 3.3'), it selects a slightly inflated number. As expected, the presence of serially correlated errors and first differencing (to deal with individual fixed effects) both deteriorate the performance of the method.

INSERT TABLE 4 HERE

6 Empirical Example

6.1 Sales growth determinants in US firms

Our empirical application examines the relationship between the sales growth (SG) of US firms. Sales growth is one of central interests in corporate finance as it serves as an important measure of corporate performance (Opler and Titman, 1994; Geroski et al., 1997; Beck et al., 2005; Barrot and Sauvagnat, 2016). Financial economists, investors, and decision makers have exhibited a keen interest in understanding which determinants affect sales growth and how. We are particularly interested in the relationship between leverage (LEV) and corporate performance as this association largely affects corporate investment strategy. On one side of this relationship, high leverage is of concern because it reduces a firm's ability to finance growth through a liquidity effect (Myers, 1977). On the other side are researchers who believe that the capital structure of a firm is essentially irrelevant because firms with good projects can always achieve funding and grow regardless of their leverage level. For example, Miller (1991) regarded financial leveraging as a second-order or largely self-correcting issue and thus argued that it should not be over-emphasized. We examine the relationship between sales growth and leverage, controlling for a number of determinants that are generally regarded as relevant for sales as in Barrot and Sauvagnat (2016) and Sojli et al. (2019), namely the logarithm of total assets (TA), Tobin's q (TQ), cash flow (CF), property, plant and equipment (PPE), and return on assets (ROA).

Thus, we examine the determinants of sales growth by allowing a structural break which induces time-varying group memberships:

$$y_{it} = \alpha_i + x'_{i,t-1}\beta_{g_i(B),B}\mathbf{1}(t < k) + x'_{i,t-1}\beta_{g_i(A),A}\mathbf{1}(t \geq k) + u_{it},$$

where y_{it} is sales growth, $x_{it} = (\text{LEV}_{it}, \text{TA}_{it}, \text{TQ}_{it}, \text{CF}_{it}, \text{PPE}_{it}, \text{ROA}_{it})$, $g_i(B) = 1, \ldots, G^B$, and $g_i(A) = 1, \ldots, G^A$. Lagged regressors are used to alleviate possible endogeneity following Sojli et al. (2019). Our modelling strategy of specifying time-varying coefficients and group memberships is motivated by the ample evidence of heterogenous responses to market-wide shocks in the literature. In particular, firms are exposed to various shocks from the market. Some shocks may reshape the structural relationship between firms' variables, leading to a structural change in the slope coefficients of the sales growth regression. A salient empirical finding in the literature on market-wide shocks is that firms' responses to such shocks have a group pattern of heterogeneity, i.e., firms in the same group respond similarly to the shocks, while the responses differ across groups (see, e.g., Banerjee et al., 2015; Duchin et al., 2010; Joh, 2003; Sojli et al., 2019, among others). Such heterogeneity of response is due to differences in observable and unobservable firm and managerial characteristics, such as profitability, industry, corporate culture, business strategies, managerial qualities, etc. Therefore, when firms are affected by market-wide shocks, not only do the relationships between firms' variables change, the group structure also may be reshaped because both the firm and managerial characteristics are changed.

We collect all variables from Compustat. To maximize the sample size of a balanced panel, we employ a sample of 740 firms from 1981 to 2013. Prior to our estimation, we follow the literature in dropping firms that contain outlying observations of sales growth (see, e.g. Adams et al., 2019), where the outliers are defined as sales growth at least 10 times larger than that in neighbouring years.[10] This leads to a sample of 703 firms.

To estimate the sales growth regression, we first determine the number of groups before and after the breaks, i.e. G^B and G^A, respectively. We allow G^B and G^A to potentially differ, and apply the IC proposed in Section 2.3.1 to select from the range of 1 to 10. The IC suggests $G^B = 5$ and $G^A = 7$; that is, two extra groups emerge after the structural break. Given the number of groups in both regimes, the LSGB estimation reports a break point in the year of 1997 ($\widehat{k} = 16$), which precisely matches the onset of the Asian financial crisis. Our diagnostic analysis as proposed in Section 2.3.2 confirms that both slope coefficients and group memberships are affected by the financial crisis.

Our break point estimate of 1997 is an interesting result. The Asian financial crisis was triggered by the dramatic devaluation of the Thai baht in the summer of 1997, and then

[10] We also tried other definitions of outliers, e.g. larger than 3 times standard error, and the results are qualitatively similar.

spread to Indonesia, South Korea, Malaysia, and other Asian countries, causing abrupt and severe economic slowdowns in the region. The serious turbulence also influenced US firms due to the strong ties between the US and Asia in international trade, money lending, etc., and the reverberations of the Asian crisis on US firms have been multifaceted (Emmons and Schmid, 2000). On one hand, the severe recession in Asian countries inevitably caused a shortfall in both demand and supply from this region and hence a deterioration of exports and imports of US firms. On the other hand, the uncertain financial environment of Asia brought in cash flows to the US, and together with lower interest rates and lower commodity prices, facilitated the growth of the US economy and further positively influenced the performance of local firms. Therefore, it is not surprising that the relationship between the financial variables of US firms would significantly change, and firms response to this crisis differed depending on their relations with the Asian market, the type of firms they were, their strategies under risky and uncertain environments, etc. To examine the effects of the determinants for each group in both regimes in detail, we label the groups in the two regimes according to the strength of the leverage effect.[11] The groups before the break are denoted as Groups 1.B–5.B, and groups after the break as Groups 1.A–7.A. We examine the effect of the determinants of sales growth and the group structure in turn.

INSERT TABLE 5 HERE

Table 5 reports the estimated slope coefficients of sales growth determinants for each group in both regimes. In general, we find the effect of leverage on sales growth to be highly heterogeneous both across groups and across regimes. A small group of 32 firms before the break, i.e., Group 5.B, is characterized by an insignificant effect of leverage, and most of these firms stay in the same group after the break (Group 7.A), with the relationship between leverage and sales growth still insignificant. Nonetheless, the majority of firms are characterized by a significant association between leverage and sales growth but with highly heterogeneous magnitude and differing directional effect. Among these firms, a small group in each regime (Group 1.B with 27 firms and 1.A with 10 firms) exhibits a particularly strong and positive effect of leverage (larger than 3), but the members of the two groups show little overlap. For the remaining firms with a significant effect of leverage, the magnitude is much smaller, ranging between -1.7 and 0.5. Such heterogeneous effect of leverage is in line with Lang et al. (1996) who also found that the association between leverage and various measures of firm growth can be either positive or negative, depending on firm characteristics. Interestingly,

[11] Because groups are invariant to relabeling, for ease of exposition and examination, we order the groups according to the strength of the leverage effect.

we find that the variation of the leverage effect across groups is enlarged after the structural break, suggesting that the financial crisis made firms more diversified in their investment and sales performance. To better understand these results, we further examine the dynamics of the composition and characteristics of each group.

INSERT TABLE 6 HERE

Table 6 provides the descriptive statistics of sales growth and all other regressors, by group.[12] We investigate each group in turn. As noted above, the members in Groups 1.B and 1.A with extraordinarily strong positive leverage effects show little overlap. The majority of firms in Group 1.B are in the transportation & public utilities industry, while those in Group 1.A are mostly related to housing and energy. Despite these diversified industries, firms of both groups generally require high investment in fixed assets but relatively low liquidity, and thus we find high total assets and low cash flow as common features of these two groups.

Another group with a significant and positive leverage effect before the break is Group 2.B containing 201 firms. The majority of this group are manufacturing firms, with roughly 20% and 35% in light and heavy manufacturing industry, respectively. These firms generally operate well with reasonably good sales growth. They are also characterized by a relatively high Tobin's q and low PPE, explaining the strongly positive effect of Tobin's q and negative effect of PPE on sales growth. The characteristics of this group are in line with Lang et al. (1996) that found a positive relationship between leverage and corporate growth for firms with high Tobin's q. After the structural break, a large number of manufacturing firms stay in the same group and form Group 2.A.[13] The estimated coefficients of most sales growth determinants remain similar after the break except that the positive effect of leverage and Tobin's q and the negative effect of return on asset both become much stronger. Further examination of the descriptive statistics shows that sales growth and leverage in Groups 2.B and 2.A do not differ much, while it is mainly the sizeable decrease of Tobin's q that causes these effects to change after the break. Although Groups 2.B and 2.A share largely the same members, a large portion of electronic & electric firms from Group 2.B, e.g., Advanced Micro Devices, Allied Motion Technologies, Analog Devices, and Eaton Corporation, move to other groups, mainly Groups 4.A and 5.A. As will be discussed shortly, Groups 4.A and 5.A are characterized by poor sales performance. These electronic & electric firms performed worse after the financial crisis partly because they were closely tied to related industries in East

[12] We also evaluate the appearance frequency of each industry (specified by the two-digit Standard Industrial Classification (SIC) code) in each group for both regimes;see the summary figure in the online supplement.

[13] Overall, almost 40% of units in Group 2.A come from Group 2.B.

and South Asia, such as the production of semiconductors and chips in South Korea, Japan, Taiwan, Singapore, and Malaysia, which were highly affected by the financial crisis.

Group 3.B is characterized by a significantly negative leverage effect, in contrast to Groups 1.B and 2.B. It contains 290 firms with a large proportion in the heavy industry and energy-related industries, i.e. 24 firms in chemical and allied products (SIC=28), 65 firms in industrial machinery equipment, electronic equipment, transportation equipment, and instruments (SIC=35–38), and 29 firms in electric, gas, and sanitary services (SIC=49). These firms perform moderately in terms of sales growth (median 0.067), but are characterized by low leverage (the lowest at the median and 95% quantile), explaining the slightly negative association between leverage and sales growth. Most members of Group 3.B also stay as one group after the break, Group 4.A, with most financial variables remaining at a similar level as in the pre-break regime, except for a decrease in PPE.[14] Interestingly, several energy-related firms, e.g. Prime Energy and EQT, move from Group 3.B to Group 2.A after the break. Further examination reveals that these switching firms experienced significant growth in sales, total assets, and leverage, among other firm financial variables, and their leverage effect turns from negative to strongly positive after the break.

Another group with an even more negative effect of leverage before the break is Group 4.B. This group covers a wide range of industries, and is featured by particular poor performance of sales growth and low total assets (lowest median sales growth and total assets across the five groups). Thus, we may interpret this group as poorly performing small firms in each industry. This group is further divided into Groups 5.A and 6.A after the financial crisis, both of which continue to struggle in their sales performance. Groups 5.A and 6.A differ in their cash flow, PPE and return on assets, leading to heterogeneous coefficient estimates of these determinants.

Overall, we find that the effects of the determinants of sales growth are highly heterogeneous across groups, and also vary significantly after the structural break. The group pattern is also restructured to a certain extent after the break. Importantly, our data-driven clustering suggests that the group structure is related but clearly does not coincide with the industry classification delineations. Hence, the industry-based grouping that most financial studies use cannot fully capture cross-sectional heterogeneity, at least in sales growth regressions. In contrast, our method allows us to capture both observed and unobserved heterogeneity and at the same time detect potential structural changes in the slope coefficients and/or the group structure.

[14] More than 42% of units in Group 4.A are from Group 3.B.

7 Conclusion

This paper proposes an estimation method that enables detection of a structural break in either group membership structure or slope coefficients, or both. We establish the consistency of estimated break dates, time-varying group memberships and slope coefficients. We derive the asymptotic distribution of our slope coefficient estimates, which is equivalent to those obtained under the known break date and group memberships. Compared with individual time series estimation, we show that the proposed method provides more accurate break point estimates, because our approach makes use of cross-sectional variation. The more accurate break point estimates further lead to more accurate clustering.

There are at least three possible directions for future research. First, a comprehensive analysis of multiple break points is of practical interest. Second, how to estimate break points and time-varying group memberships in nonlinear models remains an interesting but challenging question. Finally, extending the current approach to non-stationary data is desirable in certain applications.

References

J. Adams, D. Hayunga, S. Mansi, D. Reeb, and V. Verardi. Identifying and treating outliers in finance. *Financial Management*, 48:345–384, 2019.

T. Ando and J. Bai. Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics*, 31:163–191, 2016.

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

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

J. Bai and P. Perron. Estimating and testing linear models with multiple structural changes. *Econometrica*, 66:47–78, 1998.

B. H. Baltagi, Q. Feng, and C. Kao. Estimation of heterogeneous panels with structural breaks. *Journal of Econometrics*, 191:176–195, 2016.

B. H. Baltagi, C. Kao, and L. Liu. Estimation and identification of change points in panel models with nonstationary or stationary regressors and error term. *Econometric Reviews*, 36:85–102, 2017.

S. Banerjee, M. H. Jenner, and V. Nanda. Restraining overconfident CEOs through improved governance: Evidence from the Sarbanes-Oxley Act. *Review of Financial Studies*, 28:2812–2858, 2015.

J.-N. Barrot and J. Sauvagnat. Input specificity and the propagation of idiosyncratic shocks in production networks. *The Quarterly Journal of Economics*, 131:1543–1592, 2016.

T. Beck, A. Demirgüc-Kunt, and M. Vojislav. Financial and legal constraints to growth: Does firm size matter? *The Journal of Finance*, 60:137–177, 2005.

S. Bonhomme and E. Manresa. Grouped patterns of heterogeneity in panel data. *Econometrica*, 83:1147–1184, 2015.

X. Cheng, Z. Liao, and F. Schorfheide. Shrinkage estimation of high-dimensional factor models with structural instabilities. *The Review of Economic Studies*, 83:1511–1543, 2016.

V. Chhaochharia and Y. Grinstein. Corporate governance and firm value: The impact of the 2002 governance rules. *The Journal of Finance*, 62:1789–1825, 2007.

P. Claeys and B. Vašíček. Measuring bilateral spillover and testing contagion on sovereign bond markets in Europe. *Journal of Banking & Finance*, 46:151–165, 2014.

J. Davidson. *Stochastic limit theory: An introduction for econometricians*. Oxford University Press, Oxford, UK, 1994.

R. Duchin, J. G. Matsusaka, and O. Ozbas. When are outside directors effective? *Journal of Financial Economics*, 96:195–214, 2010.

A. Dzemski and R. Okui. Convergence rate of estimators of clustered panel models with misclassification. mimeo, 2020.

W. R. Emmons and F. A. Schmid. The Asian crisis and the exposure of large U.S. firms. *Federal Reserve Bank of St. Louis Review*, January/February:15–34, 2000.

E. Engel, R. M. Hayes, and X. Wang. The Sarbanes-Oxley Act and firms' going-private decisions. *Journal of Accounting and Economics*, 44:116–145, 2007.

P. A. Geroski, S. J. Machin, and C. F. Walters. Corporate growth and profitability. *The Journal of Industrial Economics*, 45:171–189, 1997.

J. Hahn and H. R. Moon. Panel data models with finite number of multiple equilibria. *Econometric Theory*, 26:863–881, 2010.

J. Hájek and A. Rényi. Generalization of an inequality of Kolmogorov. *Acta Mathematica Hungarica*, 6:281–283, 1955.

R. A. Heron and E. Lie. Does backdating explain the stock price pattern around executive stock option grants? *Journal of Financial Economics*, 83:271–295, 2007.

S. W. Joh. Corporate governance and firm profitability: Evidence from Korea before the economic crisis. *Journal of Financial Economics*, 68:287–322, 2003.

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

L. Lang, E. Ofek, and R. Stulz. Leverage, investment, and firm growth. *Journal of Financial Economics*, 40:3–29, 1996.

D. Li, J. Qian, and L. Su. Panel data models with interactive fixed effects and multiple structural breaks. *Journal of the American Statistical Association*, 111:1804–1819, 2016.

C.-C. Lin and S. Ng. Estimation of panel data models with parameter heterogeneity when group membership is unknown. *Journal of Econometric Methods*, 1:42–55, 2012.

J. S. Linck, J. M. Netter, and T. Yang. The effects and unintended consequences of the Sarbanes-Oxley Act on the supply and demand for directors. *The Review of Financial Studies*, 22:3287–3328, 2009.

R. Liu, Z. Shang, Y. Zhang, and Q. Zhou. Identification and estimation in panel models with overspecified number of groups. *Journal of Econometrics*, 215(2):574–590, 2020.

F. Merlevède, M. Peligrad, and E. Rio. A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151:435–474, 2011.

K. Miao, L. Su, and W. Wang. Panel threshold regressions with latent group structures. *Journal of Econometrics*, 214:451–481, 2020.

M. H. Miller. Leverage. *The Journal of Finance*, 46:479–488, 1991.

T. Mitton. A cross-firm analysis of the impact of corporate governance on the East Asian financial crisis. *Journal of Financial Economics*, 64:215–241, 2002.

S. C. Myers. Determinants of corporate borrowing. *Journal of Financial Economics*, 5: 147–175, 1977.

R. Okui and W. Wang. Heterogeneous structural breaks in panel data models. *Journal of Econometrics*, 220:447–473, 2021.

R. Okui and T. Yanagi. Panel data analysis with heterogeneous dynamics. *Journal of Econometrics*, 212:451–475, 2019.

T. C. Opler and S. Titman. Financial distress and corporate performance. *The Journal of Finance*, 49:1015–1040, 1994.

J. Qian and L. Su. Shrinkage estimation of common breaks in panel data models via adaptive group fused lasso. *Journal of Econometrics*, 191:86–109, 2016.

E. Rio. *Asymptotic theory of weakly dependent random processes*. Springer, Berlin, Heidelberg, Germany, 2017.

E. Sojli, W. W. Tham, and W. Wang. Market-wide events and time fixed effects. SSRN, 2019. URL https://ssrn.com/abstract=3258048.

L. Stracca. Our currency, your problem? The global effects of the Euro debt crisis. *European Economic Review*, 74:1–13, 2015.

L. Su, Z. Shi, and P. C. B. Phillips. Identifying latent structures in panel data. *Econometrica*, 84:2215–2264, 2016.

Y. Sun. Estimation and inference in panel strcutural models. Working paper, Department of Economics, UCSD, 2015.

A. V. Thakor. The financial crisis of 2007-2009: Why did it happen and what did we learn? *The Review of Corporate Finance Studies*, 4:155–205, 2015.

M. Vogt and O. Linton. Classification of non-parametric regression functions in longitudinal data models. *Journal of the Royal Statistical Society: Series B*, 79:5–27, 2016.

S. D. Wachter and E. Tzavalis. Detection of structural breaks in linear dynamic panel data models. *Computational Statistics & Data Analysis*, 56:3020–3034, 2012.

W. Wang, P. C. B. Phillips, and L. Su. Homogeneity pursuit in panel data models: Theory and application. *Journal of Applied Econometrics*, 33:797–815, 2018.

R. Yokoyama. Moment bounds for stationary mixing sequences. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 52:45–57, 1980.

Table 1: Accuracy of break point estimates

		$k^0 = 7$				$k^0 = 14$			
		$N = 100, T = 10$		$N = 200, T = 10$		$N = 100, T = 20$		$N = 200, T = 20$	
		HD	$\bar{k}$	HD	$\bar{k}$	HD	$\bar{k}$	HD	$\bar{k}$
DGP 1.1	LSGB	0.000	7.000	0.000	7.000	0.000	14.000	0.000	14.000
	BFK	0.247	4.523	0.247	4.528	0.125	11.490	0.112	11.747
	GAGFL	0.020	6.820	0.005	6.989	0.029	13.876	0.004	14.000
DGP 1.2	LSGB	0.019	6.915	0.003	6.977	0.002	14.012	0.000	14.002
	BFK	0.252	4.480	0.250	4.499	0.158	10.840	0.141	11.172
	GAGFL	0.071	6.657	0.062	6.825	0.061	14.000	0.047	13.648
DGP 1.3	LSGB	0.000	7.000	0.000	7.000	0.000	14.000	0.000	14.000
	BFK	0.247	4.529	0.246	4.539	0.100	11.998	0.100	12.000
	GAGFL	0.010	7.005	0.001	7.010	0.008	13.969	0.002	13.970
DGP 2.1	LSGB	0.002	6.993	0.000	7.000	0.000	14.000	0.000	14.000
	BFK	0.253	4.465	0.248	4.513	0.151	10.966	0.145	11.092
	GAGFL	0.011	6.907	0.001	7.010	0.036	13.448	0.002	14.000
DGP 2.2	LSGB	0.087	6.173	0.071	6.302	0.030	13.867	0.009	14.004
	BFK	0.246	4.535	0.248	4.517	0.178	10.438	0.172	10.547
	GAGFL	0.126	6.340	0.178	5.382	0.137	11.926	0.128	11.875
DGP 2.3	LSGB	0.000	7.000	0.000	7.000	0.000	14.000	0.000	14.000
	BFK	0.241	4.587	0.245	4.547	0.102	11.950	0.100	11.995
	GAGFL	0.014	6.963	0.000	7.000	0.005	13.926	0.000	14.000
DGP 3.1	LSGB	0.012	7.125	0.007	7.079	0.019	14.390	0.018	14.369
	BFK	0.240	4.595	0.242	4.571	0.100	12.000	0.100	12.000
	GAGFL	0.008	6.911	0.000	7.000	0.000	13.984	0.000	14.000
DGP 3.2	LSGB	0.035	6.848	0.015	6.975	0.021	14.258	0.013	14.257
	BFK	0.247	4.521	0.243	4.565	0.101	11.963	0.100	11.995
	GAGFL	0.169	5.384	0.022	6.877	0.014	13.844	0.007	13.852
DGP 3.3	LSGB	0.008	7.080	0.004	7.046	0.016	14.326	0.012	14.247
	BFK	0.226	4.736	0.223	4.761	0.100	12.000	0.100	12.000
	GAGFL	0.021	6.793	0.000	7.000	0.000	14.000	0.000	14.000
DGP 4.1	LSGB	0.000	7.000	0.000	7.000	0.000	14.000	0.000	14.000
	BFK	0.214	4.857	0.222	4.772	0.100	12.000	0.100	12.000
	GAGFL	0.000	7.000	0.000	7.000	0.000	14.000	0.000	14.000
DGP 4.2	LSGB	0.015	6.936	0.000	6.993	0.001	13.974	0.000	13.997
	BFK	0.241	4.585	0.240	4.600	0.128	11.433	0.114	11.704
	GAGFL	0.072	6.795	0.036	7.027	0.092	13.113	0.036	14.118
DGP 4.3	LSGB	0.000	7.000	0.000	7.000	0.000	14.000	0.000	14.000
	BFK	0.212	4.873	0.216	4.840	0.100	12.000	0.100	12.000
	GAGFL	0.003	7.037	0.000	7.000	0.013	14.272	0.000	14.000

Notes: HD denotes Hausdorff distance and $\bar{k}$ is the average break point estimate. LSGB is the proposed Least Squares estimator for models with Group structure and structural Break. BFK stands for the method by Baltagi et al. (2016), and GAGFL is the method by Okui and Wang (2021). Each simulation is based on 1000 replications.

Table 2: Misclustering frequency before and after the structural break

		$N = 100, T = 10$		$N = 200, T = 10$		$N = 100, T = 20$		$N = 200, T = 20$	
		MF_B	MF_A	MF_B	MF_A	MF_B	MF_A	MF_B	MF_A
DGP 1.1	LSGB	0.006	0.022	0.005	0.021	0.000	0.002	0.000	0.001
	BFK	0.025	0.006	0.024	0.006	0.001	0.001	0.000	0.001
	GAGFL	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000
DGP 1.2	LSGB	0.071	0.195	0.066	0.180	0.014	0.088	0.014	0.083
	BFK	0.130	0.170	0.124	0.168	0.031	0.098	0.027	0.094
	GAGFL	0.268	0.268	0.238	0.238	0.157	0.157	0.123	0.123
DGP 1.3	LSGB	0.005	0.022	0.005	0.021	0.000	0.002	0.000	0.001
	BFK	0.026	0.074	0.024	0.074	0.000	0.022	0.000	0.021
	GAGFL	0.031	0.031	0.017	0.017	0.002	0.002	0.001	0.001
DGP 2.1	LSGB	0.018	0.040	0.017	0.039	0.002	0.006	0.002	0.006
	BFK	0.045	0.018	0.043	0.018	0.006	0.003	0.005	0.003
	GAGFL	0.003	0.003	0.004	0.004	0.000	0.000	0.000	0.000
DGP 2.2	LSGB	0.239	0.375	0.236	0.379	0.075	0.298	0.064	0.291
	BFK	0.287	0.351	0.293	0.356	0.123	0.248	0.112	0.244
	GAGFL	0.397	0.397	0.388	0.388	0.297	0.297	0.286	0.286
DGP 2.3	LSGB	0.017	0.040	0.017	0.039	0.002	0.006	0.002	0.006
	BFK	0.043	0.085	0.042	0.086	0.004	0.028	0.004	0.027
	GAGFL	0.094	0.094	0.038	0.038	0.007	0.007	0.006	0.006
DGP 3.1	LSGB	0.020	0.064	0.020	0.062	0.001	0.010	0.001	0.010
	BFK	0.039	0.039	0.039	0.037	0.002	0.008	0.002	0.008
	GAGFL	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
DGP 3.2	LSGB	0.132	0.270	0.120	0.248	0.041	0.149	0.038	0.141
	BFK	0.169	0.239	0.156	0.224	0.047	0.144	0.044	0.137
	GAGFL	0.170	0.170	0.186	0.186	0.090	0.090	0.060	0.060
DGP 3.3	LSGB	0.021	0.123	0.020	0.119	0.004	0.027	0.003	0.029
	BFK	0.037	0.123	0.036	0.120	0.002	0.045	0.004	0.046
	GAGFL	0.008	0.008	0.005	0.005	0.000	0.000	0.000	0.000
DGP 4.1	LSGB	0.002	0.012	0.002	0.009	0.000	0.000	0.000	0.000
	BFK	0.014	0.245	0.015	0.247	0.000	0.225	0.000	0.216
	GAGFL	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
DGP 4.2	LSGB	0.034	0.101	0.032	0.090	0.003	0.027	0.003	0.025
	BFK	0.087	0.106	0.081	0.104	0.007	0.039	0.007	0.035
	GAGFL	0.198	0.198	0.124	0.124	0.075	0.075	0.040	0.040
DGP 4.3	LSGB	0.002	0.012	0.002	0.013	0.000	0.000	0.000	0.000
	BFK	0.014	0.438	0.014	0.444	0.000	0.423	0.000	0.414
	GAGFL	0.031	0.031	0.009	0.009	0.025	0.025	0.000	0.000

Notes: MF_B is the misclustering frequency before the break, and MF_A denotes the frequency after the break. LSGB is the proposed Least Squares estimator for models with Group structure and structural Break. BFK stands for the method by Baltagi et al. (2016). GAGFL is the method by Okui and Wang (2021) which assumes time-invariant group structures, and thus its $\text{MF}_B = \text{MF}_A$. Each simulation is based on 1000 replications.

Table 3: Average mean squared error of coefficient estimates

	$N=100, T=10$			$N=200, T=10$			$N=100, T=20$			$N=200, T=20$		
	LSGB	BFK	GAGFL	LSGB	BFK	GAGFL	LSGB	BFK	GAGFL	LSGB	BFK	GAGFL
DGP 1.1	0.005	0.029	0.007	0.002	0.027	0.003	0.002	0.011	0.004	0.001	0.010	0.001
DGP 1.2	0.014	0.024	0.035	0.004	0.019	0.030	0.003	0.009	0.017	0.001	0.008	0.017
DGP 1.3	0.004	0.075	0.152	0.002	0.072	0.143	0.002	0.043	0.140	0.001	0.042	0.138
DGP 2.1	0.018	0.050	0.020	0.009	0.042	0.011	0.008	0.020	0.009	0.005	0.017	0.004
DGP 2.2	0.312	0.327	0.329	0.268	0.301	0.296	0.108	0.112	0.134	0.081	0.093	0.133
DGP 2.3	0.018	0.096	0.209	0.007	0.084	0.167	0.010	0.053	0.155	0.005	0.046	0.149
DGP 3.1	0.013	0.047	0.009	0.006	0.035	0.003	0.006	0.012	0.005	0.004	0.011	0.001
DGP 3.2	0.025	0.036	0.868	0.012	0.028	0.057	0.006	0.013	0.069	0.004	0.009	0.020
DGP 3.3	0.028	0.103	0.161	0.012	0.091	0.141	0.012	0.048	0.142	0.009	0.044	0.138
DGP 4.1	0.029	26.994	0.012	0.016	28.169	0.004	0.002	2256.7	0.019	0.007	2300.8	0.008
DGP 4.2	0.016	0.021	0.053	0.006	0.015	0.031	0.005	0.013	0.025	0.002	0.009	0.018
DGP 4.3	0.094	24.579	0.190	0.073	27.772	0.142	0.002	2255.5	0.210	0.001	2229.8	0.146

Notes: LSGB is the proposed Least Squares estimator for models with Group structure and structural Break. BFK stands for the method by Baltagi et al. (2016), and GAGFL is the method by Okui and Wang (2021). Each simulation is based on 1000 replications.

Table 4: Determine the number of groups

		$N = 100, T = 10$				$N = 200, T = 10$				$N = 100, T = 20$				$N = 200, T = 20$			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
DGP 1.2	G^B	0.000	**0.990**	0.001	0.000	0.000	**1.000**	0.000	0.000	0.000	**0.981**	0.001	0.000	0.000	**1.000**	0.000	0.000
	G^A	0.000	**0.967**	0.033	0.000	0.000	0.000	**1.000**	0.000	0.000	0.383	**0.617**	0.000	0.000	0.050	**0.950**	0.000
DGP 1.3	G^B	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000	0.000
	G^A	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000
DGP 2.2	G^B	0.000	**0.989**	0.011	0.000	0.000	0.017	**0.833**	0.150	0.000	**0.992**	0.008	0.000	0.000	0.008	**0.825**	0.167
	G^A	0.000	**0.983**	0.017	0.000	0.000	0.000	**0.925**	0.075	0.000	0.283	**0.717**	0.000	0.000	0.008	**0.967**	0.025
DGP 2.3	G^B	0.000	**0.983**	0.017	0.000	0.000	0.050	**0.683**	0.267	0.000	**0.992**	0.008	0.000	0.000	0.017	**0.650**	0.333
	G^A	0.000	0.000	**1.000**	0.000	0.000	0.000	**0.975**	0.025	0.000	0.000	**1.000**	0.000	0.000	0.000	**0.958**	0.042
DGP 3.2	G^B	0.345	**0.655**	0.000	0.000	0.000	**0.947**	0.053	0.000	0.000	**0.876**	0.124	0.000	0.026	**0.974**	0.000	0.000
	G^A	0.000	0.000	0.368	**0.632**	0.000	0.079	0.026	**0.895**	0.000	0.158	**0.679**	0.163	0.000	0.000	0.123	**0.877**
DGP 3.3	G^B	0.342	**0.632**	0.026	0.000	0.000	**1.000**	0.000	0.000	0.000	**0.974**	0.026	0.000	0.000	**1.000**	0.000	0.000
	G^A	0.000	0.000	0.105	**0.895**	0.000	0.000	0.102	**0.898**	0.000	0.053	0.053	**0.895**	0.000	0.000	0.084	**0.916**
DGP 4.2	G^B	0.000	**0.978**	0.022	0.000	0.000	**0.986**	0.014	0.000	0.000	**0.996**	0.004	0.000	0.000	**0.994**	0.006	0.000
	G^A	0.000	0.010	**0.990**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000
DGP 4.3	G^B	0.000	**1.000**	0.000	0.000	0.000	**0.998**	0.002	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000	0.000
	G^A	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	**1.000**	0.000

Notes: This table presents the empirical probability of choosing a specific number of groups in both regimes. The true number of groups before the break, G^B, is 2, and the true number after the break, G^A, is 3.

Table 5: Estimates of sales growth regression

	Pre-break						
	Group 1.B	Group 2.B	Group 3.B	Group 4.B	Group 5.B		
LEV	3.249***	0.193***	−0.107***	−0.961***	−0.004		
	(0.719)	(0.073)	(0.025)	(0.110)	(0.272)		
TA	−0.243*	−0.343***	−0.083***	−0.446***	−0.265***		
	(0.142)	(0.036)	(0.029)	(0.044)	(0.107)		
TQ	0.313*	0.152***	−0.007**	0.022***	0.396***		
	(0.167)	(0.020)	(0.003)	(0.004)	(0.096)		
CF	−0.594	−0.485***	−0.107	0.108	7.077***		
	(0.553)	(0.144)	(0.108)	(0.146)	(0.923)		
PPE	−0.485	−2.062***	0.180**	0.484***	1.774***		
	(0.507)	(0.165)	(0.087)	(0.150)	(0.401)		
ROA	−10.79***	−1.583***	−2.112***	−0.286***	0.472		
	(1.442)	(0.140)	(0.113)	(0.058)	(0.430)		
No. firms	27	201	290	153	32		
	Post-break						
	Group 1.A	Group 2.A	Group 3.A	Group 4.A	Group 5.A	Group 6.A	Group 7.A
LEV	3.483**	0.435***	0.420**	−0.117***	−0.794***	−1.695***	−0.021
	(1.604)	(0.105)	(0.194)	(0.038)	(0.126)	(0.225)	(0.046)
TA	−0.730***	−0.348***	0.080	−0.147***	0.044	−0.710***	−1.403***
	(0.273)	(0.032)	(0.082)	(0.033)	(0.049)	(0.070	(0.119)
TQ	1.362***	0.286***	0.002	0.025***	0.110***	0.105***	0.013
	(0.471)	(0.018)	(0.011)	(0.004)	(0.015)	(0.023)	(0.048)
CF	1.967	0.138	1.446***	0.076	−0.543**	−0.087	−2.355***
	(1.769)	(0.098)	(0.232)	(0.070)	(0.242)	(0.246)	(0.689)
PPE	−11.001***	−0.878***	2.646***	−0.316***	−1.440***	1.023***	−2.561***
	(2.735)	(0.117)	(0.427)	(0.124)	(0.359)	(0.313)	(0.446)
ROA	−17.781***	−2.739***	−1.861***	−0.350***	−5.313***	−2.987***	0.466***
	(3.990)	(0.130)	(0.154)	(0.049)	(0.295)	(0.330)	(0.118)
No. firms	10	167	77	247	105	67	30

Notes: LEV is leverage, TA is logarithm of total assets, TQ is Tobin's q, CF is cash flow, PPE is the ratio of property plant and equipment over total assets, ROA is return on assets.

Table 6: Descriptive statistics of firm variables

		Before the break					After the break						
	Quantile	1.B	2.B	3.B	4.B	5.B	1.A	2.A	3.A	4.A	5.A	6.A	7.A
SG	5%	−0.268	−0.224	−0.172	−0.218	−0.289	−0.074	−0.227	−0.258	−0.183	−0.173	−0.235	−0.266
	50%	0.077	0.063	0.067	0.054	0.077	0.069	0.066	0.065	0.062	0.068	0.064	0.055
	95%	0.564	0.429	0.377	0.426	0.689	0.760	0.406	0.489	0.363	0.440	0.466	0.475
LEV	5%	0.041	0.214	0.195	0.162	0.304	0.323	0.229	0.217	0.169	0.151	0.184	0.333
	50%	0.656	0.558	0.550	0.574	0.658	0.725	0.558	0.601	0.539	0.570	0.596	0.643
	95%	0.946	0.916	0.901	0.919	0.967	0.981	0.887	0.924	0.908	0.899	0.942	0.952
TA	5%	3.894	2.875	3.329	2.498	3.741	1.682	3.350	3.399	2.591	3.433	3.406	2.642
	50%	8.523	7.130	7.075	6.671	7.495	8.465	7.106	6.894	6.804	7.477	7.519	7.554
	95%	11.799	10.520	10.385	10.990	12.012	13.646	10.527	10.507	10.418	10.727	10.855	11.717
TQ	5%	0.192	0.398	0.319	0.252	0.105	0.098	0.356	0.281	0.378	0.250	0.196	0.229
	50%	0.834	0.952	1.052	0.899	0.788	0.747	0.918	0.975	1.066	0.950	0.950	0.824
	95%	2.193	2.771	3.142	2.748	2.036	1.328	2.019	3.270	3.513	2.691	2.952	2.047
CF	5%	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	50%	0.018	0.032	0.024	0.025	0.023	0.009	0.030	0.020	0.031	0.019	0.022	0.022
	95%	0.154	0.244	0.200	0.248	0.151	0.139	0.213	0.201	0.244	0.181	0.227	0.182
PPE	5%	0.000	0.016	0.011	0.009	0.000	0.000	0.026	0.000	0.014	0.000	0.007	0.012
	50%	0.239	0.238	0.307	0.321	0.276	0.015	0.279	0.281	0.277	0.276	0.334	0.385
	95%	0.841	0.658	0.826	0.837	0.836	0.814	0.771	0.854	0.753	0.817	0.843	0.846
ROA	5%	0.000	−0.034	−0.001	−0.037	−0.055	0.004	−0.020	−0.014	−0.032	0.000	−0.027	−0.069
	50%	0.069	0.087	0.095	0.080	0.062	0.055	0.083	0.081	0.104	0.083	0.075	0.062
	95%	0.176	0.223	0.230	0.235	0.157	0.128	0.189	0.235	0.262	0.203	0.198	0.178

Notes: SG is sales growth, LEV is leverage, TA is logarithm of total assets, TQ is Tobin's q, CF is cash flow, PPE is the ratio of property plant and equipment over total assets, ROA is return on assets.

Supplementary appendix to

Estimation of panel group structure models with structural breaks in group memberships and coefficients

This technical appendix includes the proof of the theorem. First, Section A.1 presents the lemmas. The proofs of the theorem and the corollary are included in Section A.2. Section A.3 discusses possible relaxation of some assumptions in the paper and provides an alternative proof of the theorems under a different set of conditions.

A.1 Lemmas

Let

$$Q(k,\gamma,\beta) = \frac{1}{NT}\left(\sum_{t=1}^{k-1}\sum_{i=1}^{N}(y_{it} - x_{it}'\beta_{g_i(B),B})^2 + \sum_{t=k}^{T}\sum_{i=1}^{N}(y_{it} - x_{it}'\beta_{g_i(A),A})^2\right),$$

and

$$\tilde{Q}(k,\gamma,\beta) = \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}(x_{it}'(\beta^0_{g^0_{it},t} - \beta_{g_{it},t}))^2 + \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}u_{it}^2.$$

Lemma 1. *Suppose that Assumptions 1(i) and 1(ii) hold. Then we have that*

$$\sup_{k\in\mathbb{K},\gamma\in\mathbb{G},\beta\in\mathbb{B}}\left|\tilde{Q}(k,\gamma,\beta) - Q(k,\gamma,\beta)\right| = O_p\left(\frac{1}{\sqrt{T}}\right).$$

Proof. The proof is almost identical to the proof of Lemma S.3 of Bonhomme and Manresa (2015), and thus we keep it brief here. First, we consider the case in which $k \geq k^0$, and we have that

$$\begin{aligned}\tilde{Q}(k,\gamma,\beta) - Q(k,\gamma,\beta) =& -2\frac{1}{NT}\sum_{t=1}^{k^0-1}\sum_{i=1}^{N}x_{it}'(\beta^0_{g_i^0(B),B} - \beta_{g_i(B),B})u_{it}\\ &-2\frac{1}{NT}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}x_{it}'(\beta^0_{g_i^0(A),A} - \beta_{g_i(B),B})u_{it}\\ &-2\frac{1}{NT}\sum_{t=k}^{T}\sum_{i=1}^{N}x_{it}'(\beta^0_{g_i^0(A),A} - \beta_{g_i(A),A})u_{it}.\end{aligned}$$

Observe that

$$\frac{1}{NT}\sum_{t=1}^{k^0-1}\sum_{i=1}^{N}x_{it}'\beta^0_{g_i^0(B),B}u_{it} = \frac{1}{NT}\sum_{g\in\mathbb{G}^B}\sum_{t=1}^{k^0-1}\sum_{i=1}^{N}\mathbf{1}(g_i(B)=g)x_{it}'\beta^0_{g_i^0(B),B}u_{it}.$$

For each $g \in \mathbb{G}^B$, by the Cauchy-Schwarz inequality we have that

$$E\left(\frac{1}{NT}\sum_{t=1}^{k^0-1}\sum_{i=1}^{N}\mathbf{1}(g_i(B)=g)x_{it}'\beta^0_{g_i^0(B),B}u_{it}\right)^2 \le CE\left\|\frac{1}{NT}\sum_{t=1}^{k^0-1}\sum_{g_i(B)=g}x_{it}u_{it}\right\|^2 = O\left(\frac{k^0}{NT^2}\right),$$

where C satisfies $\|\beta_{g_{it},t}\|^2 < C$ for any $\beta \in \mathcal{B}$ and the existence of such C is guaranteed by Assumption 1(*ii*), the inequality follows by the definition of C, and the equality follows by Assumption 1(*i*). Next, we consider

$$\begin{aligned}
\left(\frac{1}{NT}\sum_{t=1}^{k^0-1}\sum_{i=1}^{N}x_{it}'\beta_{g_i(B),B}u_{it}\right)^2 &\le \left(\frac{1}{NT}\sum_{i=1}^{N}\beta_{g_i(B),B}\sum_{t=1}^{k^0-1}x_{it}u_{it}\right)^2 \\
&\le \left(\frac{1}{N}\sum_{i=1}^{N}||\beta_{g_i(B),B}||^2\right)\left(\frac{1}{NT^2}\sum_{i=1}^{N}\left\|\sum_{t=1}^{k^0-1}x_{it}u_{it}\right\|^2\right) \\
&= O_p\left(\frac{k^0}{T^2}\right),
\end{aligned}$$

where the first inequality uses the Cauchy-Schwarz inequality and the second inequality follows by that Assumption 1(*ii*) implies $\sum_{i=1}^N ||\beta_{g_i(B),B}||^2/N < C$ for some C, and that Assumption 1(*i*) together with the Markov inequality implies $\sum_{i=1}^N \left\|\sum_{t=1}^{k^0-1} x_{it}u_{it}\right\|^2/(NT^2) = O_p(k^0/T^2)$. The other terms in the expression for $\tilde{Q}(k,\gamma,\beta) - Q(k,\gamma,\beta)$ can be analyzed similarly. It therefore holds that

$$\begin{aligned}
\tilde{Q}(k,\gamma,\beta) - Q(k,\gamma,\beta) =& O\left(\frac{\sqrt{k^0}}{\sqrt{NT}}\right) + O\left(\frac{\sqrt{k^0}}{T}\right) + O\left(\frac{\sqrt{k-k^0}}{\sqrt{NT}}\right) + O\left(\frac{\sqrt{k-k^0}}{T}\right) \\
&+ O\left(\frac{\sqrt{T-k}}{\sqrt{NT}}\right) + O\left(\frac{\sqrt{T-k}}{T}\right),
\end{aligned}$$

uniformly over β and γ. The argument for $k < k^0$ is similar. Because $k \le T$ by construction, we have that

$$\sup_{k\in\mathbb{K},\gamma\in\mathbb{G},\beta\in\mathbb{B}}\left|\tilde{Q}(k,\gamma,\beta) - Q(k,\gamma,\beta)\right| = O_p\left(\frac{1}{\sqrt{T}}\right).$$

□

Lemma 2. *Suppose that Assumptions 1(i)–1(vii) hold. Then we have that*

(1) $\max_{g\in\mathbb{G}^B}\min_{\tilde{g}\in\mathbb{G}^B}\left\|\beta^0_{g,B} - \hat{\beta}_{\tilde{g},B}\right\|^2 = O_p(1/\sqrt{T})$,

(2) $\max_{g\in\mathbb{G}^A}\min_{\tilde{g}\in\mathbb{G}^A}\left\|\beta^0_{g,A} - \hat{\beta}_{\tilde{g},A}\right\|^2 = O_p(1/\sqrt{T})$,

(3) $(\hat{k} - k^0)/T = O_p(1/\sqrt{T})$.

Proof. From Lemma 1, we have that

$$
\begin{aligned}
\tilde{Q}(\hat{k},\hat{\gamma},\hat{\beta}) =& Q(\hat{k},\hat{\gamma},\hat{\beta}) + O_p\left(\frac{1}{\sqrt{T}}\right)\\
\leq& Q(k^0,\gamma^0,\beta^0) + O_p\left(\frac{1}{\sqrt{T}}\right) = \tilde{Q}(k^0,\gamma^0,\beta^0) + O_p\left(\frac{1}{\sqrt{T}}\right).
\end{aligned}
$$

Because $\tilde{Q}(k,\gamma,\beta)$ is minimized at (k^0,γ^0,β^0), we have that

$$
\tilde{Q}(\hat{k},\hat{\gamma},\hat{\beta}) - \tilde{Q}(k^0,\gamma^0,\beta^0) = O_p\left(\frac{1}{\sqrt{T}}\right).
$$

Let $a_{NT} = \tilde{Q}(\hat{k},\hat{\gamma},\hat{\beta}) - \tilde{Q}(k^0,\gamma^0,\beta^0)$, and note that $a_{NT} = O_p\left(1/\sqrt{T}\right)$.

We first consider the case in which $k \geq k^0$, and observe that

$$
\begin{aligned}
\tilde{Q}(k,\gamma,\beta) - \tilde{Q}(k^0,\gamma^0,\beta^0) =& \frac{1}{NT}\sum_{t=1}^{k^0-1}\sum_{i=1}^{N}(x_{it}'(\beta^0_{g_i^0(B),B} - \beta_{g_i(B),B}))^2\\
&+ \frac{1}{NT}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}(x_{it}'(\beta^0_{g_i^0(A),A} - \beta_{g_i(B),B}))^2\\
&+ \frac{1}{NT}\sum_{t=k}^{T}\sum_{i=1}^{N}(x_{it}'(\beta^0_{g_i^0(A),A} - \beta_{g_i(A),A}))^2. \qquad \text{(A.8)}
\end{aligned}
$$

We study the three terms on the right side of (A.8) separately. For first term, it holds that

$$
\begin{aligned}
&\frac{1}{NT}\sum_{t=1}^{k^0-1}\sum_{i=1}^{N}(x_{it}'(\beta^0_{g_i^0(B),B} - \beta_{g_i(B),B}))^2\\
=&\frac{1}{NT}\sum_{t=1}^{k^0-1}\sum_{g=1}^{G^B}\sum_{\tilde{g}=1}^{G^B}\sum_{i=1}^{N}\mathbf{1}\{g_i^0(B)=g\}\{g_i(B)=\tilde{g}\}(x_{it}'(\beta^0_{g,B} - \beta_{\tilde{g},B}))^2\\
\geq&\frac{1}{T}\sum_{t=1}^{k^0-1}\sum_{g=1}^{G^B}\sum_{\tilde{g}=1}^{G^B}\rho_{N,t}(\gamma,g,\tilde{g})\left\|\beta^0_{g,B} - \beta_{\tilde{g},B}\right\|^2 \geq \frac{k^0-1}{T}\hat{\rho}\max_{g\in\mathbb{G}^B}\min_{\tilde{g}\in\mathbb{G}^B}\left\|\beta^0_{g,B} - \beta_{\tilde{g},B}\right\|^2,
\end{aligned}
$$

where the last inequality follows by Assumption 1(*iv*). This further implies $(k^0-1)/T\hat{\rho}\max_{g\in\mathbb{G}^B}$ $\min_{\tilde{g}\in\mathbb{G}^B}\left\|\beta^0_{g,B} - \hat{\beta}_{\tilde{g},B}\right\|^2 \leq a_{NT}$. Moreover, Assumption 1(*iv*) implies that

$$
\frac{1}{NT}\sum_{t=1}^{k^0-1}\sum_{i=1}^{N}(x_{it}'(\beta^0_{g_i^0(B),B} - \beta_{g_i(B),B}))^2 \geq \frac{k^0-1}{T}\hat{\rho}^*\frac{1}{N}\sum_{i=1}^{N}\|\beta^0_{g_i^0(B),B} - \beta_{g_i(B),B}\|^2. \qquad \text{(A.9)}
$$

Thus we have

$$
\frac{1}{N}\sum_{i=1}^{N}\|\beta^0_{g_i^0(B),B} - \hat{\beta}_{\hat{g}_i(B),B}\|^2 < Ca_{NT}.
$$

With similar reasoning, we have, for the third term, that

$$\frac{1}{NT}\sum_{t=k}^{T}\sum_{i=1}^{N}(x_{it}'(\beta^0_{g_i^0(A),A}-\beta_{g_i(A),A}))^2 \geq \frac{T-k}{T}\hat{\rho}\max_{g\in\mathbb{G}^A}\min_{\tilde{g}\in\mathbb{G}^A}\left\|\beta^0_{g,A}-\beta_{\tilde{g},A}\right\|^2,$$

and that $(T-\hat{k})/T\hat{\rho}\max_{g\in\mathbb{G}^A}\min_{\tilde{g}\in\mathbb{G}^A}\left\|\beta^0_{g,A}-\hat{\beta}_{\tilde{g},A}\right\|^2 \leq a_{NT}$. Finally, for the second term, we observe that

$$\begin{aligned}
&\frac{1}{NT}\sum_{t=k^0}^{\hat{k}-1}\sum_{i=1}^{N}(x_{it}'(\beta^0_{g_i^0(A),A}-\hat{\beta}_{\hat{g}_i(B),B}))^2\\
=&\frac{1}{NT}\sum_{t=k^0}^{\hat{k}-1}\sum_{i=1}^{N}(x_{it}'(\beta^0_{g_i^0(A),A}-\beta^0_{g_i^0(B),B}+\beta^0_{g_i^0(B),B}-\hat{\beta}_{\hat{g}_i(B),B}))^2\\
\geq&\frac{1}{NT}\sum_{t=k^0}^{\hat{k}-1}\sum_{i=1}^{N}(x_{it}'(\beta^0_{g_i^0(A),A}-\beta^0_{g_i^0(B),B}))^2+\frac{1}{NT}\sum_{t=k^0}^{\hat{k}-1}\sum_{i=1}^{N}(x_{it}'(\beta^0_{g_i^0(B),B}-\hat{\beta}_{\hat{g}_i(B),B}))^2\\
&-2\frac{1}{NT}\sum_{t=k^0}^{\hat{k}-1}\sum_{i=1}^{N}\left|x_{it}'(\beta^0_{g_i^0(A),A}-\beta^0_{g_i^0(B),B})\right|\cdot\left|x_{it}'(\beta^0_{g_i^0(B),B}-\hat{\beta}_{\hat{g}_i(B),B}))\right|\\
\geq&\frac{1}{NT}\sum_{t=k^0}^{\hat{k}-1}\sum_{i=1}^{N}(x_{it}'(\beta^0_{g_i^0(A),A}-\beta^0_{g_i^0(B),B}))^2+\frac{1}{NT}\sum_{t=k^0}^{\hat{k}-1}\sum_{i=1}^{N}(x_{it}'(\beta^0_{g_i^0(B),B}-\hat{\beta}_{\hat{g}_i(B),B}))^2\\
&-2\frac{1}{T}\sum_{t=k^0}^{\hat{k}-1}\left(\frac{1}{N}\sum_{i=1}^{N}(x_{it}'(\beta^0_{g_i^0(A),A}-\beta^0_{g_i^0(B),B}))^2\right)^{1/2}\left(\frac{1}{N}\sum_{i=1}^{N}(x_{it}'(\beta^0_{g_i^0(B),B}-\hat{\beta}_{\hat{g}_i(B),B}))^2\right)^{1/2}.
\end{aligned}$$

Assumptions 1(ii) and 1(v) imply that $\sum_{i=1}^{N}(x_{it}'(\beta^0_{g_i^0(A),A}-\beta^0_{g_i^0(B),B}))^2/N<C$. Assumptions 1($v$) and (A.9) imply that $\frac{1}{N}\sum_{i=1}^{N}(x_{it}'(\beta^0_{g_i^0(B),B}-\hat{\beta}_{\hat{g}_i(B),B}))^2<Ca_{NT}$. Hence, we have, by Assumption 1(vi),

$$\frac{1}{NT}\sum_{t=k^0}^{\hat{k}-1}\sum_{i=1}^{N}(x_{it}'(\beta^0_{g_i^0(A),A}-\hat{\beta}_{\hat{g}_i(B),B}))^2 \geq \frac{\hat{k}-k}{T}\left(\underline{m}-C\sqrt{a_{NT}}\right),$$

which further implies that

$$\frac{\hat{k}-k}{T}\left(\underline{m}-C\sqrt{a_{NT}}\right)<a_{NT}.$$

It therefore follows that for $\hat{k}\geq k^0$,

$$\begin{aligned}
\frac{k^0-1}{T}\hat{\rho}\max_{g\in\mathbb{G}^B}\min_{\tilde{g}\in\mathbb{G}^B}\left\|\beta^0_{g,B}-\hat{\beta}_{\tilde{g},B}\right\|^2 \leq& a_{NT},\\
\frac{\hat{k}-k^0}{T}\left(\underline{m}-C\sqrt{a_{NT}}\right) \leq& a_{NT},\\
\frac{T-\hat{k}}{T}\hat{\rho}\max_{g\in\mathbb{G}^A}\min_{\tilde{g}\in\mathbb{G}^A}\left\|\beta^0_{g,A}-\hat{\beta}_{\tilde{g},A}\right\|^2 \leq& a_{NT}.
\end{aligned}$$

Next, we consider the case in which $\hat{k} \leq k^0$. We can follow similar arguments to those in the case of $\hat{k} \geq k^0$ and obtain that

$$\begin{aligned}\frac{\hat{k}-1}{T}\hat{\rho}\max_{g\in\mathbb{G}^B}\min_{\tilde{g}\in\mathbb{G}^B}\left\|\beta^0_{g,B}-\hat{\beta}_{\tilde{g},B}\right\|^2 \leq& a_{NT},\\ \frac{k^0-\hat{k}}{T}\left(\underline{m}-C\sqrt{a_{NT}}\right) \leq& a_{NT},\\ \frac{T-k^0}{T}\hat{\rho}\max_{g\in\mathbb{G}^A}\min_{\tilde{g}\in\mathbb{G}^A}\left\|\beta^0_{g,A}-\hat{\beta}_{\tilde{g},A}\right\|^2 \leq& a_{NT}.\end{aligned}$$

In either case, we must have that

$$\frac{\hat{k}-k^0}{T} = O_p(a_{NT}) = O_p\left(\frac{1}{\sqrt{T}}\right).$$

because of Assumption 1(*vii*). It also follows that

$$\hat{\rho}\max_{g\in\mathbb{G}^B}\min_{\tilde{g}\in\mathbb{G}^B}\left\|\beta^0_{g,B}-\hat{\beta}_{\tilde{g},B}\right\|^2 = O_p(a_{NT}) = O_p\left(\frac{1}{\sqrt{T}}\right),$$

and that

$$\hat{\rho}\max_{g\in\mathbb{G}^A}\min_{\tilde{g}\in\mathbb{G}^A}\left\|\beta^0_{g,A}-\hat{\beta}_{\tilde{g},A}\right\|^2 = O_p\left(\frac{1}{\sqrt{T}}\right).$$

The desired result holds by Assumption 1(*iv*).

□

Lemma 3. *Suppose that Assumptions 1(i)–1(viii) are satisfied. Then there exist permutations* $\sigma_B : \mathbb{G}^B \mapsto \mathbb{G}^B$ *and* $\sigma_A : \mathbb{G}^A \mapsto \mathbb{G}^A$ *such that* $\left\|\beta^0_{g,B}-\hat{\beta}_{\sigma_B(g),B}\right\|^2 = O_p(1/\sqrt{T})$ *for any* $g \in \mathbb{G}^B$ *and* $\left\|\beta^0_{g,A}-\hat{\beta}_{\sigma_A(g),A}\right\|^2 = O_p(1/\sqrt{T})$ *for any* $g \in \mathbb{G}^A$.

Proof. The proof is constructive. Let

$$\sigma_B(g) = \min_{\tilde{g}\in\mathbb{G}^B}\left\|\beta^0_{g,B}-\hat{\beta}_{\tilde{g},B}\right\|^2, \quad \text{and} \quad \sigma_A(g) = \min_{\tilde{g}\in\mathbb{G}^A}\left\|\beta^0_{g,B}-\hat{\beta}_{\tilde{g},A}\right\|^2.$$

We show that $\sigma_B(g)$ is a permutation that satisfies the requirement in the statement of the lemma. We only present the argument for σ_B and omit that for σ_A because the arguments in the two cases are very similar.

By Lemma 2, it follows that

$$\max_{g\in\mathbb{G}^B}\min_{\tilde{g}\in\mathbb{G}^B}\left\|\beta^0_{g,B}-\hat{\beta}_{\tilde{g},B}\right\|^2 = O_p(1/\sqrt{T}).$$

Thus, by the definition of σ_B, it holds that $\left\|\beta^0_{g,B}-\hat{\beta}_{\sigma_B(g),B}\right\|^2 = O_p(1/\sqrt{T})$ for any $g \in \mathbb{G}^B$. Next, we show that σ_B is a permutation. Let $g \neq \tilde{g}$, and then by the triangular inequality, we have that

$$\left\|\hat{\beta}_{\sigma_B(g),B}-\hat{\beta}_{\sigma_B(\tilde{g}),B}\right\| \geq \left\|\beta^0_{g,B}-\beta^0_{\tilde{g},B}\right\| - \left\|\beta^0_{g,B}-\hat{\beta}_{\sigma_B(g),B}\right\| - \left\|\beta^0_{\tilde{g},B}-\hat{\beta}_{\sigma_B(\tilde{g}),B}\right\|.$$

Recall that we have already shown that $\left\|\beta^0_{g,B} - \hat{\beta}_{\sigma_B(g),B}\right\| = o_p(1)$ and $\left\|\beta^0_{\tilde{g},B} - \hat{\beta}_{\sigma_B(\tilde{g}),B}\right\| = o_p(1)$. Besides, Assumption 1(*viii*) states that $\left\|\beta^0_{g,B} - \beta^0_{\tilde{g},B}\right\| > c$. This means that $\sigma_B(g) \neq \sigma_B(\tilde{g})$ for $g \neq \tilde{g}$ with probability approaching one, which further implies that σ_B admits a well defined inverse and is bijective. Hence, σ_B is a permutation that satisfies the requirement in the statement of Lemma 3, implying that $\min_{g \in \mathbb{G}^B} \left\|\beta^0_{g,B} - \hat{\beta}_{\tilde{g},B}\right\|^2 = O_p(1/\sqrt{T})$ for any $\tilde{g}$. Thus the Hausdorff distance between β^0_B and $\hat{\beta}_B$ is of order $O_p(1/\sqrt{T})$. $\square$

Define the Hausdorff distance between β^0_B and $\hat{\beta}_B$ to be

$$\max\left(\max_{g \in \mathbb{G}^B} \min_{\tilde{g} \in \mathbb{G}^B} \left\|\beta^0_{g,B} - \hat{\beta}_{\tilde{g},B}\right\|^2, \max_{\tilde{g} \in \mathbb{G}^B} \min_{g \in \mathbb{G}^B} \left\|\beta^0_{g,B} - \hat{\beta}_{\tilde{g},B}\right\|^2\right).$$

By Lemmas 2 and 3, this Hausdorff distance converges to 0 at the rate of $\sqrt{T}$. Using the similar arguments, we can show that $\hat{\beta}$ is consistent under the Hausdorff distance and its rate of convergence is $\sqrt{T}$. By relabeling, we can set $\sigma_B(g) = g$ and $\sigma_A(g) = g$, the convention that we adopt throughout the paper, such that $\left\|\beta^0_{g,B} - \hat{\beta}_{g,B}\right\|^2 = O_p(1/\sqrt{T})$ for any $g \in \mathbb{G}^B$ and $\left\|\beta^0_{g,A} - \hat{\beta}_{g,A}\right\|^2 = O_p(1/\sqrt{T})$ for any $g \in \mathbb{G}^A$.

Let $\mathcal{N}$ be a neighborhood of β^0 such that $\left\|\beta^0_{g,C} - \beta_{g,C}\right\| < \eta$ for $\eta > 0$ for any $g \in \mathbb{G}^C$ and $C = B, A$. Note that we will take η small enough by considering large N and T by Lemma 3. Let $\bar{k} = \sqrt{T}\log T + k^0$ and $\underline{k} = -\sqrt{T}\log T + k^0$. Define $K = \{k : \underline{k} \leq k \leq \bar{k}\}$.

Lemma 4. *Suppose that Assumptions 1(ii), 1(iv), 1(vii), 1(viii), and 1(ix) hold. As $N, T \to \infty$ with $NT^{-\delta} \to 0$, it holds that*

$$\Pr\left\{\hat{\gamma}(k, \beta) \neq \gamma^0 \text{ for some } k \in K \text{ and } \beta \in \mathcal{N}\right\} \to 0.$$

Proof. To show this probability converges to zero, it is equivalent to show that

$$\max_{1 \leq i \leq N} \sup_{\beta \in \mathcal{N}} \max_{k \in K} \mathbf{1}\{\hat{g}_i(B)(k, \beta) \neq g^0_i(B)\} + \max_{1 \leq i \leq N} \sup_{\beta \in \mathcal{N}} \max_{k \in K} \mathbf{1}\{\hat{g}_i(A)(k, \beta) \neq g^0_i(A)\} = o_p(1),$$

where we observe that

$$\mathbf{1}\left\{\hat{g}_i(B)(k, \beta) \neq g^0_i(B)\right\} = \max_{g \in \mathbb{G}^B \setminus \{g^0_i(B)\}} \mathbf{1}\left(\sum_{t=1}^{k-1}(y_{it} - x'_{it}\beta_{g,B})^2 < \sum_{t=1}^{k-1}(y_{it} - x'_{it}\beta_{g^0_i(B),B})^2\right), \tag{A.10}$$

and a similar equality holds for $\mathbf{1}\{\hat{g}_i(A)(k, \beta) \neq g^0_i(A)\}$. We analyze the probability of each of these two indicators being one. To this end, we first evaluate how the deviation of k from k^0 plays a role, while the situation of $k = k^0$ can be analysed using the same arguments as in Bonhomme and Manresa (2015) and Okui and Wang (2021).

We first examine the difference between the two summations in the argument of the indicator function in (A.10), and show that this difference evaluated at any $k \in K$ and that evaluated at $k = k^0$ are not very different. Let

$$D = \sum_{t=1}^{k-1} \left((y_{it} - x_{it}'\beta_{g,B})^2 - (y_{it} - x_{it}'\beta_{g_i^0(B),B})^2 \right) - \sum_{t=1}^{k^0-1} \left((y_{it} - x_{it}'\beta_{g,B})^2 - (y_{it} - x_{it}'\beta_{g_i^0(B),B})^2 \right).$$

First, considering the case of $k < k^0$, we have that

$$\begin{aligned}
|D| =& \left| \sum_{t=k}^{k^0-1} 2u_{it}x_{it}(\beta_{g_i^0(B),B} - \beta_{g,B}) + \sum_{t=k}^{k^0-1} (\beta_{g_i^0(B),B} - \beta_{g,B})' x_{it}x_{it}' (2\beta^0_{g_i^0(B),B} - \beta_{g_i^0(B),B} - \beta_{g,B}) \right| \\
\leq & M_1 \left\| \sum_{t=k}^{k^0-1} u_{it}x_{it} \right\| + M_2 \left\| \sum_{t=k}^{k^0-1} x_{it}x_{it}' \right\| \\
\leq & M_1(k^0 - \underline{k}) \frac{1}{k^0 - \underline{k}} \sum_{t=\underline{k}}^{k^0-1} \|u_{it}x_{it}\| + M_2(k^0 - \underline{k}) \left\| \frac{1}{k^0 - \underline{k}} \sum_{t=\underline{k}}^{k^0-1} x_{it}x_{it}' \right\|,
\end{aligned}$$

where M_1 and M_2 are constants independent of (i, g, k, β). Let $M_T = T^{1/4}/\log T$. Under Assumption 1(*ix*), we can apply inequality (1.8) in Merlevède et al. (2011) which is based on Theorem 6.2 of Rio (2017), translated from a French version published in 2000, with $\lambda = (k^0 - \underline{k})M_T = T^{3/4}$ and obtain that

$$\begin{aligned}
&\Pr\left(\frac{1}{k^0 - \underline{k}} \left| \sum_{t=\underline{k}}^{k^0-1} (\|u_{it}x_{it}\| - E(\|u_{it}x_{it}\|) \right| > M_T \right) \\
\leq & 4\exp\left(-\frac{\lambda^{d/(d+1)}\log 2}{2} \right) + 16CM_T^{-1}\exp\left(-a\frac{\lambda^{d/(d+1)}}{b^d} \right) = o(T^{-\delta}),
\end{aligned}$$

where $d = d_1d_2/(d_1 + d_2)$, (a, b, d_1, d_2) are defined in Assumption 1(*ix*). Noting that $(k^0 - \underline{k})^{-1}\sum_{t=\underline{k}}^{k^0-1} E(\|u_{it}x_{it}\|)$ converges and $M_T \to \infty$, we have that $\Pr\left((k^0 - \underline{k})^{-1}\sum_{t=\underline{k}}^{k^0-1} \|u_{it}x_{it}\| > M_T\right) = o(T^{-\delta})$. Similarly, it holds $\Pr(\|(k^0 - \underline{k})^{-1}\sum_{t=\underline{k}}^{k^0-1} x_{it}x_{it}'\| > M_T) = o(T^{-\delta})$. These imply that there exists a sequence that satisfies $C_T = O(M_T)$ and $C_T \to \infty$ as $T \to \infty$, such that

$$\Pr\left(\frac{1}{k^0}|D| > \frac{k^0 - \underline{k}}{k^0} C_T \right) = o(T^{-\delta}).$$

Using a similar argument, we can show that for $k \geq k^0$,

$$\Pr\left(\frac{1}{k^0}|D| > \frac{\bar{k} - k^0}{k^0} C_T \right) = o(T^{-\delta}).$$

Next, we consider $\sum_{t=1}^{k^0-1} \left((y_{it} - x_{it}'\beta_{g,B})^2 - (y_{it} - x_{it}'\beta_{g_i^0(B),B})^2 \right)$. This term can be considered in a similar way to Bonhomme and Manresa (2015) and Okui and Wang (2021). We have

$$
\begin{aligned}
&\sum_{t=1}^{k^0-1}\left((y_{it}-x_{it}'\beta_{g,B})^2-(y_{it}-x_{it}'\beta_{g_i^0(B),B})^2\right)\\
=&\sum_{t=1}^{k^0-1}2u_{it}x_{it}(\beta_{g_i^0(B),B}-\beta_{g,B})+\sum_{t=1}^{k^0-1}(\beta_{g_i^0(B),B}-\beta_{g,B})'x_{it}x_{it}'(2\beta^0_{g_i^0(B),B}-\beta_{g_i^0(B),B}-\beta_{g,B})\\
=&\sum_{t=1}^{k^0-1}2u_{it}x_{it}(\beta^0_{g_i^0(B),B}-\beta^0_{g,B})+\sum_{t=1}^{k^0-1}(x_{it}'(\beta^0_{g_i^0(B),B}-\beta^0_{g,B}))^2+\Psi,
\end{aligned}
$$

where

$$
\begin{aligned}
\Psi=&\sum_{t=1}^{k^0-1}2u_{it}x_{it}(\beta_{g_i^0(B),B}-\beta_{g,B}-\beta^0_{g_i^0(B),B}+\beta^0_{g,B})\\
&+\sum_{t=1}^{k^0-1}(\beta_{g_i^0(B),B}-\beta_{g,B}-\beta^0_{g_i^0(B),B}+\beta^0_{g,B})'x_{it}x_{it}'(2\beta^0_{g_i^0(B),B}-\beta_{g_i^0(B),B}-\beta_{g,B})\\
&+\sum_{t=1}^{k^0-1}(\beta^0_{g_i^0(B),B}-\beta^0_{g,B})'x_{it}x_{it}'(\beta^0_{g_i^0(B),B}-\beta_{g_i^0(B),B}-\beta_{g,B}+\beta^0_{g,B}).
\end{aligned}
$$

By the Cauchy-Schwarz inequality, Assumption 1(ii) and the definition of $\mathcal{N}$ imply that

$$
|\Psi|\le\eta C_1\left\|\sum_{t=1}^{k^0-1}u_{it}x_{it}\right\|+\eta C_2\left\|\sum_{t=1}^{k^0-1}x_{it}x_{it}'\right\|,
$$

where C_1 and C_2 are constants independent of η and T. We then have that

$$
\begin{aligned}
&\mathbf{1}\left(\sum_{t=1}^{k-1}(y_{it}-x_{it}'\beta_{g,B})^2<\sum_{t=1}^{k-1}(y_{it}-x_{it}'\beta_{g_i^0(B),B})^2\right)\\
\le&\mathbf{1}\Bigg(\sum_{t=1}^{k^0-1}2u_{it}x_{it}'(\beta^0_{g_i^0(B),B}-\beta^0_{g,B})\\
&\qquad-\sum_{t=1}^{k^0-1}(x_{it}'(\beta^0_{g_i^0(B),B}-\beta^0_{g,B}))^2+\eta C_1\left\|\sum_{t=1}^{k^0-1}u_{it}x_{it}\right\|+\eta C_2\left\|\sum_{t=1}^{k^0-1}x_{it}x_{it}'\right\|+|D|\Bigg).
\end{aligned}
$$

Note that the right hand side does not depend on β. Thus, we have

$$
\begin{aligned}
&\Pr\left(\sup_{\beta\in\mathcal{N}}\max_{k\in K}\mathbf{1}(\hat{g}_i(B)(k,\beta)\neq g_i^0(B))\neq 0\right)\\
=&\Pr\left(\sup_{\beta\in\mathcal{N}}\max_{k\in K}\max_{g\in\mathbb{G}^B\setminus\{g_i^0(B)\}}\mathbf{1}\left(\sum_{t=1}^{k-1}(y_{it}-x_{it}'\beta_{g,B})^2<\sum_{t=1}^{k-1}(y_{it}-x_{it}'\beta_{g_i^0(B),B})^2\right)\neq 0\right)\\
\le&\sum_{g\in\mathbb{G}^B\setminus\{g_i^0(B)\}}\Pr\left(\sum_{t=1}^{k^0-1}2u_{it}x_{it}'(\beta^0_{g_i^0(B),B}-\beta^0_{g,B})\right.
\end{aligned}
$$

$$< -\sum_{t=1}^{k^0-1}(x'_{it}(\beta^0_{g_i^0(B),B}-\beta^0_{g,B}))^2+\eta C_1\left\|\sum_{t=1}^{k^0-1}u_{it}x_{it}\right\|+\eta C_2\left\|\sum_{t=1}^{k^0-1}x_{it}x'_{it}\right\|+|D|\Bigg)$$

$$\leq \sum_{g\in\mathbb{G}^B\setminus\{g_i^0(B)\}}\Bigg(\Pr\left(\frac{1}{k^0}\sum_{t=1}^{k^0-1}(x'_{it}(\beta^0_{g_i^0(B),B}-\beta^0_{g,B}))^2\leq\frac{c''}{2}\right)+\Pr\left(\left\|\frac{1}{k^0}\sum_{t=1}^{k^0-1}u_{it}x_{it}\right\|\geq M\right)$$

$$+\Pr\left(\left\|\frac{1}{k^0}\sum_{t=1}^{k^0-1}x_{it}x'_{it}\right\|\geq M\right)+\Pr\left(\frac{1}{k^0}|D|>\frac{k^0-\underline{k}}{k^0}C_T\right)$$

$$+\Pr\left(\frac{1}{k^0}\sum_{t=1}^{k^0-1}2u_{it}x'_{it}(\beta^0_{g_i^0(B),B}-\beta^0_{g,B})<-\frac{c''}{2}+\eta C_1M+\eta C_2M+\frac{k^0-\underline{k}}{k^0}C_T\right)\Bigg),$$

where we take $c''=c\times\rho^*$ for c in Assumption 1$(viii)$ and ρ^* in Assumption 1(iv).

We use the following lemma by Bonhomme and Manresa (2015) which is based on Rio (2017).

Lemma 5 (Lemma B.5 in Bonhomme and Manresa (2015))**.** *Let z_t be a strongly mixing process with zero mean, with strong mixing coefficients $a[t]\leq e^{-at^{d_1}}$ and with tail probabilities $\Pr(|z_t|>z)\leq e^{1-(z/b)^{d_2}}$, where a, b, d_1, and d_2 are positive constants. Then for all $z>0$, we have for all $\delta>0$, as $T\to\infty$,*

$$T^{\delta}\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}z_t\right|\geq z\right)\to 0.$$

Note that this lemma holds uniformly over i as long as the bounds for mixing coefficients and tail probabilities hold uniformly over i.

We observe that

$$\Pr\left(\left\|\frac{1}{k^0}\sum_{t=1}^{k^0-1}x_{it}x'_{it}\right\|\geq M\right)\leq\Pr\left(\frac{1}{k^0}\sum_{t=1}^{k^0-1}\|x_{it}x'_{it}\|\geq M\right)=\Pr\left(\frac{1}{k^0}\sum_{t=1}^{k^0-1}x'_{it}x_{it}\geq M\right).$$

We then apply Lemma 5, regarding $x'_{it}x_{it}-E(x'_{it}x_{it})$ as z_t in the lemma, and Assumption 1(ix) yields that $\Pr\left(\left\|(k^0)^{-1}\sum_{t=1}^{k^0-1}x_{it}x'_{it}\right\|\geq M\right)=o\left((k^0)^{-\delta}\right)=o\left(T^{-\delta}\right)$, where the last equality holds by Assumption 1(vii). Similarly, Assumption 1(ix) also implies that $\Pr\left(\left\|(k^0)^{-1}\sum_{t=1}^{k^0-1}u_{it}x_{it}\right\|\geq M\right)=o\left(T^{-\delta}\right)$. Moreover, a similar argument shows that under Assumption 1(ix), Lemma 5 implies that

$$\Pr\left(\left|\frac{1}{k^0}\sum_{t=1}^{k^0-1}(x'_{it}(\beta^0_{g_i^0(B),B}-\beta^0_{g,B}))^2-\frac{1}{k^0}\sum_{t=1}^{k^0-1}E((x'_{it}(\beta^0_{g_i^0(B),B}-\beta^0_{g,B}))^2)\right|\geq\frac{c''}{2}\right)=o(T^{-\delta}),$$

which in turn implies that under Assumptions 1(iv) and 1$(viii)$,

$$\Pr\left(\frac{1}{k^0}\sum_{t=1}^{k^0-1}(x'_{it}(\beta^0_{g_i^0(B),B}-\beta^0_{g,B}))^2\leq\frac{c''}{2}\right)=o(T^{-\delta})$$

uniformly over g. Now we have shown that $\Pr\{(k^0)^{-1}|D| > ((k^0-\underline{k})/k^0)C_T\} = o(T^{-\delta})$. Note that $((k^0-\underline{k})/k^0)C_T \to 0$ because $M_T = o(\sqrt{T}/\log T)$, $k^0 = O(T)$ and $k^0-\underline{k} = O(\sqrt{T}\log T)$. Moreover, by similar arguments as above that use Lemma 5, we can take η small enough and also T large enough such that

$$\begin{aligned}
&\Pr\left(\frac{1}{k^0}\sum_{t=1}^{k^0-1} 2u_{it}x_{it}'(\beta^0_{g_i^0(B),B} - \beta^0_{g,B}) < -\frac{c''}{2} + \eta C_1 M + \eta C_2 M + \frac{k^0-\underline{k}}{k^0}C_T\right) \\
\leq &\Pr\left(\frac{1}{k^0}\sum_{t=1}^{k^0-1} 2u_{it}x_{it}'(\beta^0_{g_i^0(B),B} - \beta^0_{g,B}) < -\frac{c''}{4}\right) = o(T^{-\delta}).
\end{aligned}$$

uniformly over g under Assumption 1(*ix*). It thus follows that

$$\begin{aligned}
&\Pr\left(\max_{1\leq i\leq N}\sup_{\beta\in\mathcal{N}}\max_{k\in K}\mathbf{1}(\hat{g}_i(B)(k,\beta)\neq g_i^0(B))\neq 0\right) \\
\leq &\sum_{i=1}^{N}\Pr\left(\sup_{\beta\in\mathcal{N}}\max_{k\in K}\mathbf{1}(\hat{g}_i(B)(k,\beta)\neq g_i^0(B))\neq 0\right) = o(NT^{-\delta}).
\end{aligned}$$

Similarly, we can show that $\Pr\left(\max_{1\leq i\leq N}\sup_{\beta\in\mathcal{N}}\max_{k\in K}\mathbf{1}(\hat{g}_i(A)(k,\beta)\neq g_i^0(A))\neq 0\right) = o(NT^{-\delta})$. This completes the proof.

□

A.2 Proof of theorem and corollary

Proof of Theorem 1

Proof. We observe that

$$\begin{aligned}
\Pr(\hat{k}\neq k^0) &\leq \Pr(\hat{k}\neq k^0, \hat{\beta}\in\mathcal{N}) + \Pr(\hat{\beta}\notin\mathcal{N}) \\
&\leq \Pr(\hat{k}\neq k^0, \hat{\gamma}=\gamma^0, \hat{\beta}\in\mathcal{N}) + \Pr(\hat{\gamma}\neq\gamma^0, \hat{\beta}\in\mathcal{N}) + \Pr(\hat{\beta}\notin\mathcal{N}).
\end{aligned}$$

We analyze the three terms in the right hand side of the above display. First, for the third term, Lemma 3 and the discussion below it imply that $\Pr(\hat{\beta}\notin\mathcal{N})\to 0$. For the second term, by Lemmas 2, 3 and 4, we have that

$$\begin{aligned}
\Pr(\hat{\gamma}\neq\gamma^0, \hat{\beta}\in\mathcal{N}) &\leq \Pr\{\hat{\gamma}(k,\beta)\neq\gamma^0 \text{ for some } k\in K \text{ and } \beta\in\mathcal{N}, \hat{\beta}\in\mathcal{N}\} + \Pr(\hat{k}\notin K) \\
&\leq \Pr\{\hat{\gamma}(k,\beta)\neq\gamma^0 \text{ for some } k\in K \text{ and } \beta\in\mathcal{N}\} + \Pr(\hat{k}\notin K)\to 0.
\end{aligned}$$

Finally, we consider the first term. We observe that

$$\Pr(\hat{k}\neq k^0, \hat{\gamma}=\gamma^0, \hat{\beta}\in\mathcal{N}) \leq \Pr(\hat{k}\neq k^0, \hat{\gamma}=\gamma^0, \hat{\beta}\in\mathcal{N}, \hat{k}\in K) + \Pr(\hat{k}\notin K)$$

$$\leq \Pr\{\hat{k}(\gamma^0,\beta) \neq k^0 \text{ for some } \beta \in \mathcal{N}, \hat{\gamma} = \gamma^0, \hat{\beta} \in \mathcal{N}\} + \Pr(\hat{k} \notin K)$$
$$\leq \Pr\{\hat{k}(\gamma^0,\beta) \neq k^0 \text{ for some } \beta \in \mathcal{N}\} + \Pr(\hat{k} \notin K),$$

where $\hat{k}(\gamma^0,\beta) = \text{argmin}_{k\in K} Q(k,\gamma,\beta)$. Note that $\Pr(\hat{k} \notin K) \to 0$ by Lemma 2, and also that $\hat{k}(\gamma^0,\beta) \neq k^0$ is equivalent to

$$Q(k^0,\gamma^0,\beta) > \min_{k\in K\setminus\{k^0\}} Q(k,\gamma^0,\beta) = \min\left(\min_{k>k^0} Q(k,\gamma^0,\beta), \min_{k<k^0} Q(k,\gamma^0,\beta)\right).$$

Thus, we have

$$\begin{aligned}
&\Pr(\hat{k}(\gamma^0,\beta) \neq k^0 \text{ for some } \beta \in \mathcal{N})\\
\leq &\Pr\left(Q(k^0,\gamma^0,\beta) > \min_{k^0<k\leq\bar{k}} Q(k,\gamma^0,\beta) \text{ for some } \beta \in \mathcal{N}\right)\\
&+ \Pr\left(Q(k^0,\gamma^0,\beta) > \min_{\underline{k}\leq k<k^0} Q(k,\gamma^0,\beta) \text{ for some } \beta \in \mathcal{N}\right).
\end{aligned}$$

Suppose for the moment that $k^0 < k \leq \bar{k}$. Using $y_{it} = x_{it}'\beta^0_{g^0_{i(C)},C} + u_{it}$ where $C = B$ if $t < k^0$ and $C = A$ if $t \geq k^0$, then we have that

$$\begin{aligned}
&Q(k^0,\gamma^0,\beta) - Q(k,\gamma^0,\beta)\\
=& -\frac{1}{NT}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}(x_{it}'(\beta_{g^0_i(A),A} - \beta_{g^0_i(B),B}))^2 + \frac{2}{NT}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}(x_{it}'(\beta^0_{g^0_i(A),A} - \beta_{g^0_i(A),A}))^2\\
&-\frac{2}{NT}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}x_{it}'(\beta_{g^0_i(A),A} - \beta_{g^0_i(B),B})u_{it}.
\end{aligned}$$

Let $d^0_i = \beta^0_{g^0_i(A),A} - \beta^0_{g^0_i(B),B}$ and $d_i = \beta_{g^0_i(A),A} - \beta_{g^0_i(B),B}$. It holds that

$$\begin{aligned}
&\frac{1}{NT}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}(x_{it}'(\beta_{g^0_i(A),A} - \beta_{g^0_i(B),B}))^2 = \frac{1}{NT}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}(x_{it}'(d^0_i + d_i - d^0_i))^2\\
\geq&\frac{1}{NT}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}(x_{it}'d^0_i)^2 + \frac{1}{NT}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}(x_{it}'(d_i - d^0_i))^2 - 2\frac{1}{NT}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}\left|x_{it}'d^0_i\right|\cdot\left|x_{it}'(d_i - d^0_i)\right|\\
\geq&\frac{1}{NT}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}(x_{it}'d^0_i)^2 + \frac{1}{NT}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}(x_{it}'(d_i - d^0_i))^2\\
&-2\frac{1}{T}\sum_{t=k^0}^{k-1}\left(\frac{1}{N}\sum_{i=1}^{N}(x_{it}'d^0_i)^2\right)^{1/2}\left(\frac{1}{N}\sum_{i=1}^{N}(x_{it}'(d_i - d^0_i))^2\right)^{1/2}.
\end{aligned}$$

Assumptions 1(*ii*) and 1(*v*) imply that $\sum_{i=1}^{N}(x_{it}'d^0_i)^2/N < (k-k^0)C/T$. Similarly, Assumption 1(*v*) and the condition that $\beta \in \mathcal{N}_\eta$ imply that $\sum_{i=1}^{N}(x_{it}'(d_i - d^0_i))^2/N < (k-k^0)\eta^2 C/T$.

We thus have that $\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}(x_{it}'(\beta_{g_i^0(A),A}-\beta_{g_i^0(B),B}))^2/(NT)\geq(k-k^0)\left(\underline{m}-C\eta\right)/T$, by Assumption 1(*vi*). Then by taking η small enough, we have that

$$\frac{1}{NT}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}(x_{it}'(\beta_{g_i^0(A),A}-\beta_{g_i^0(B),B}))^2\geq\frac{k-k^0}{2T}\underline{m}.$$

And therefore

$$\frac{2}{NT}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}(x_{it}'(\beta^0_{g_i^0(A),A}-\beta_{g_i^0(A),A}))^2>0.$$

It follows that

$$\begin{aligned}
&\Pr\left(Q(k^0,\gamma^0,\beta)>\min_{k^0<k\leq\bar{k}}Q(k,\gamma^0,\beta)\text{ for some }\beta\in\mathcal{N}\right)\\
=&\Pr\left(\sup_{\beta\in\mathcal{N}}\max_{k^0<k\leq\bar{k}}\left(Q(k^0,\gamma^0,\beta)-Q(k,\gamma^0,\beta)\right)>0\right)\\
\leq&\Pr\left(\sup_{\beta\in\mathcal{N}}\max_{k^0<k\leq\bar{k}}\left(-2\frac{1}{NT}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}x_{it}'(\beta_{g_i^0(A),A}-\beta_{g_i^0(B),B})u_{it}-\frac{k-k^0}{2T}\underline{m}\right)>0\right)\\
=&\Pr\left(\sup_{\beta\in\mathcal{N}}\max_{k^0<k\leq\bar{k}}\left(-2\frac{1}{N}\frac{1}{k-k^0}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}x_{it}'(\beta_{g_i^0(A),A}-\beta_{g_i^0(B),B})u_{it}-\frac{\underline{m}}{2}\right)>0\right)\\
\leq&\Pr\left(\sup_{\beta\in\mathcal{N}}\max_{k^0<k\leq\bar{k}}\left(-2\frac{1}{N}\frac{1}{k-k^0}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}x_{it}'(\beta_{g_i^0(A),A}-\beta_{g_i^0(B),B})u_{it}\right)>\frac{\underline{m}}{2}\right).
\end{aligned}$$

Observing that

$$\begin{aligned}
&-2\frac{1}{N}\frac{1}{k-k^0}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}x_{it}'(\beta_{g_i^0(A),A}-\beta_{g_i^0(B),B})u_{it}\\
=&-2\frac{1}{N}\frac{1}{k-k^0}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}x_{it}'(\beta^0_{g_i^0(A),A}-\beta^0_{g_i^0(B),B})u_{it}\\
&+2\frac{1}{N}\frac{1}{k-k^0}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}x_{it}'(\beta^0_{g_i^0(A),A}-\beta_{g_i^0(A),A}-\beta_{g_i^0(B),B}+\beta^0_{g_i^0(B),B})u_{it},
\end{aligned}$$

and

$$\left|2\frac{1}{N}\frac{1}{k-k^0}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}x_{it}'(\beta^0_{g_i^0(A),A}-\beta_{g_i^0(A),A}-\beta_{g_i^0(B),B}+\beta^0_{g_i^0(B),B})u_{it}\right|\leq\eta C\left\|\frac{1}{N}\frac{1}{k-k^0}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}x_{it}u_{it}\right\|,$$

we thus have that

$$\Pr\left(\sup_{\beta\in\mathcal{N}}\max_{k^0<k\leq\bar{k}}\left(-2\frac{1}{N}\frac{1}{k-k^0}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}x_{it}'(\beta_{g_i^0(A),A}-\beta_{g_i^0(B),B})u_{it}\right)>\frac{\underline{m}}{2}\right)$$

$$\leq \Pr\left(\max_{k^0<k\leq\bar{k}}\left(-2\frac{1}{N}\frac{1}{k-k^0}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}x_{it}'(\beta^0_{g_i^0(A),A}-\beta^0_{g_i^0(B),B})u_{it}\right)>\frac{m}{4}\right)$$
$$+\Pr\left(\eta C\max_{k^0<k\leq\bar{k}}\left\|\frac{1}{N}\frac{1}{k-k^0}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}x_{it}u_{it}\right\|>\frac{m}{4}\right)=O\left(\frac{1}{N}\right),$$

where the last equality follows by applying Bai and Perron (1998, Lemma A.6) which is an extension of Hájek and Rényi (1955). Here we use the observation that an L_r-bounded mixing sequence is an L_p mixingale sequence for $1\leq p<r$ as discussed in (Davidson, 1994, page 248). Thus, under Assumptions 1(*ix*) and 1(*x*), $x_{it}u_{it}$ is an L_2 mixingale and we can apply Bai and Perron (1998, Lemma A.6).

A similar argument shows that

$$\Pr\left(Q(k^0,\gamma^0,\beta)>\min_{\underline{k}<k<k^0}Q(k,\gamma^0,\beta)\text{ for some }\beta\in\mathcal{N}\right)=O\left(\frac{1}{N}\right).$$

To sum up, we have that $\Pr(\hat{k}\neq k^0,\hat{\gamma}=\gamma^0,\hat{\beta}\in\mathcal{N})\to 0$.

□

Proof of Corollary 1

Proof. We first show result (1) of the corollary. We observe that

$$\Pr(\hat{\gamma}\neq\gamma^0)\leq\Pr(\hat{\gamma}\neq\gamma^0,\hat{\beta}\in\mathcal{N})+\Pr(\hat{\beta}\notin\mathcal{N}).$$

The second paragraph of the proof of Theorem 1 shows that $\Pr(\hat{\gamma}\neq\gamma^0,\hat{\beta}\in\mathcal{N})\to 0$. Lemma 3 and the discussion below imply that $\Pr(\hat{\beta}\notin\mathcal{N})\to 0$. Hence, the desired result holds.

Next, we show result (2) of the corollary. We have that

$$\begin{aligned}\Pr\left(\left\|\hat{\beta}-\tilde{\beta}\right\|>a/\sqrt{NT}\right)\leq&\Pr\left(\left\|\hat{\beta}-\tilde{\beta}\right\|>a\sqrt{NT},\hat{\gamma}=\gamma^0,\hat{k}=k^0\right)+\Pr(\gamma\neq\gamma^0)+\Pr(\hat{k}\neq k^0)\\ \leq&0+\Pr(\gamma\neq\gamma^0)+\Pr(\hat{k}\neq k^0)\to 0\end{aligned}$$

for any $a>0$, where the second inequality follows because $\hat{\beta}=\tilde{\beta}$ holds under $\gamma=\gamma^0$ and $\hat{k}=k^0$, and the third inequality holds by result (1) of this corollary and Theorem 1. We thus have the desired result. □

A.3 Proofs under alternative conditions

In this section, we relax the break size condition in Assumption 1(*vi*), the group separation assumption in Assumption 1(*viii*) as well as the mixing and moment conditions in Assumption 1(*ix*). In exchange for this, however, we have to impose strict stationarity. Formally, we make the following assumption as an alternative to Assumptions 1(*vi*), 1(*viii*) and 1(*ix*).

Assumption 3.

(vi) There exists a fixed constant $\underline{M} > 0$ *such that for any* t, $\underline{m} = \underline{M}T^{-q}$ *for some* $q > 0$ *satisfies*

$$\frac{1}{N}\sum_{i=1}^{N}(x_{it}'(\beta^0_{g_i^0(A),A} - \beta^0_{g_i^0(B),B}))^2 > \underline{m}.$$

(viii) There exists a nonrandom sequence $c_T > T^{-1/2+e}$ *for some* $e > 0$ *such that for any* $g \neq \tilde{g}$ *where* $g, \tilde{g} \in \mathbb{G}^B$ *and* $g', \check{g} \in \mathbb{G}^A$, *it holds that* $\|\beta^0_{g,B} - \beta^0_{\tilde{g},B}\| > c_T$ *and* $\|\beta^0_{g',A} - \beta^0_{\check{g},A}\| > c_T$.

(ix) Let z_{it} *be* $x_{it}'x_{it}$, $\|u_{it}x_{it}\|$, $2u_{it}x_{it}'(\beta^0_{g^0_{i(l)},l} - \beta^0_{g,l})$, *or* $(x_{it}'(\beta^0_{g^0_{i(l)},l} - \beta^0_{g,l}))^2$ *for* $g \in \mathbb{G}^l$ *and* $l = A, B$. *Assume the following holds for any choice of* z_{it}: *1)* z_{it} *is a strictly stationary and strong mixing sequence over* t *whose mixing coefficients* $a_i[t]$ *are bounded by* $a[t]$ *such that* $\max_{1\leq i\leq N} a_i[t] \leq a[t]$ *and* $\sum_{t=0}^{\infty}(t+1)^{r/2-1}a[t]^{b/r+b} < \infty$ *for some* $b > 0$, *and* $\max_{1\leq i\leq N} E(|z_{it}|^{r+b}) < \infty$ *for some* $b > 0$; *2) There exists* a_i, $i = 1, \ldots, N$, *such that for any* $\epsilon > 0$, *it holds that* $\max_{1\leq i\leq N} |a_i - \sum_{t=1}^{T} E(z_{it})/T| < \epsilon$ *for* T *sufficiently large.*

Assumption 3(*vi*) allows the break size to shrink as $T \to \infty$. In Theorem A.1 below, we impose a concrete condition on q. A fixed break size corresponds to $q = 0$.

Assumption 3(*viii*) allows a pair of groups to be identical in the limit. However, in finite samples, they are separated and the magnitude of the difference between coefficients of two groups is bounded from below by an order slower than $1/\sqrt{T}$. This specification captures situations in which groups are different but the difference is not large compared with the length of time series.

Assumption 3(*ix*) imposes weaker mixing and moment conditions than those in Assumption 1(*ix*). In exchange for this relaxation on the mixing and moment conditions, we need to impose strict stationarity, which is a relatively strong assumption here given that we consider structural breaks. Note that Assumption 3(*ix*) only imposes stationarity on the regressors and error terms but not the dependent variable, and thus it can still be satisfied in some applications.

Under this alternative set of assumptions, we obtain the consistency of the estimator and asymptotic distribution of the coefficients estimator. However, the required condition on the relative magnitude of N and T becomes stronger.

Theorem A.1. *Suppose that Assumptions 1(i)-1(v), 3(vi), 1(vii), 3(viii), 3(ix) and 1(x) hold. For* q *defined in Assumption 3(vi) and* e *defined in Assumption 3(viii), suppose that* $q < \min(1/4, e)$. *As* $N, T \to \infty$ *with* $NT^{-er} \to 0$ *and* $T^{2q}/N \to 0$, *where* r *is defined in Assumption 3(ix),* $\Pr(\hat{k} = k^0) \to 1$.

Corollary A.1. *Suppose that Assumptions 1(i)-1(v), 3(vi), 1(vii), 3(viii), 3(ix) and 1(x) hold. For q defined in Assumption 3(vi) and e defined in Assumption 3(viii), suppose that $q < \min(1/4, e)$. As $N, T \to \infty$ with $NT^{-er} \to 0$ and $T^{2q}/N \to 0$, where r is defined in Assumption 3(ix),*

(1) $\Pr(\hat{\gamma} = \gamma^0) \to 1$,

(2) $\hat{\beta} = \tilde{\beta} + o_p(1/\sqrt{NT})$, *where $\tilde{\beta}$ is the estimator of β under $k = k^0$ and $\gamma = \gamma^0$.*

We now prove Theorem A.1 and Corollary A.1. We note that Lemma 1 holds under Assumptions 1(*i*) and 1(*ii*) and it can be used without modification even under the alternative set of assumptions. However, Lemmas 2, 3, and 4 and their proofs need to be modified as follows.

Lemma 6. *Suppose that Assumptions 1(i)-1(v), 3(vi) and 1(vii) hold. Also assume that $q < 1/4$ where q is defined in Assumption 3(vi). Then we have that*

(1) $\max_{g\in\mathbb{G}^B} \min_{\tilde{g}\in\mathbb{G}^B} \left\|\beta^0_{g,B} - \hat{\beta}_{\tilde{g},B}\right\|^2 = O_p(1/\sqrt{T})$,

(2) $\max_{g\in\mathbb{G}^A} \min_{\tilde{g}\in\mathbb{G}^A} \left\|\beta^0_{g,A} - \hat{\beta}_{\tilde{g},A}\right\|^2 = O_p(1/\sqrt{T})$,

(3) $(\hat{k} - k^0)/T = O_p(T^{-1/2+q})$.

Proof. The proof is identical to that of Lemma 2 until we arrive at the following step. It follows that for $\hat{k} \geq k^0$, we have

$$\frac{k^0-1}{T}\hat{\rho}\max_{g\in\mathbb{G}^B}\min_{\tilde{g}\in\mathbb{G}^B}\left\|\beta^0_{g,B} - \hat{\beta}_{\tilde{g},B}\right\|^2 \leq a_{NT}, \quad \frac{\hat{k}-k^0}{T}\left(\underline{m} - C\sqrt{a_{NT}}\right) \leq a_{NT},$$
$$\frac{T-\hat{k}}{T}\hat{\rho}\max_{g\in\mathbb{G}^A}\min_{\tilde{g}\in\mathbb{G}^A}\left\|\beta^0_{g,A} - \hat{\beta}_{\tilde{g},A}\right\|^2 \leq a_{NT}.$$

Similarly, for $\hat{k} \leq k^0$, we have

$$\frac{\hat{k}-1}{T}\hat{\rho}\max_{g\in\mathbb{G}^B}\min_{\tilde{g}\in\mathbb{G}^B}\left\|\beta^0_{g,B} - \hat{\beta}_{\tilde{g},B}\right\|^2 \leq a_{NT}, \quad \frac{k^0-\hat{k}}{T}\left(\underline{m} - C\sqrt{a_{NT}}\right) \leq a_{NT},$$
$$\frac{T-k^0}{T}\hat{\rho}\max_{g\in\mathbb{G}^A}\min_{\tilde{g}\in\mathbb{G}^A}\left\|\beta^0_{g,A} - \hat{\beta}_{\tilde{g},A}\right\|^2 \leq a_{NT}.$$

Noting that $\underline{m} > C\sqrt{a_{NT}}$ for sufficiently large T because $q < 1/4$, in either case ($\hat{k} \geq k^0$ or $\hat{k} \leq k^0$), we must have that $(\hat{k} - k^0)/T = O_p(a_{NT}) = O_p\left(T^{-1/2+q}\right)$ because of Assumption 1(*vii*). This result and Assumption 1(*iv*) imply that

$$\max_{g\in\mathbb{G}^B}\min_{\tilde{g}\in\mathbb{G}^B}\left\|\beta^0_{g,B} - \hat{\beta}_{\tilde{g},B}\right\|^2 = O_p(a_{NT}) = O_p\left(\frac{1}{\sqrt{T}}\right), \text{ and } \max_{g\in\mathbb{G}^A}\min_{\tilde{g}\in\mathbb{G}^A}\left\|\beta^0_{g,A} - \hat{\beta}_{\tilde{g},A}\right\|^2 = O_p\left(\frac{1}{\sqrt{T}}\right).$$

□

Lemma 7. *Suppose that Assumptions 1(i)-1(v), 3(vi), 1(vii) and 3(viii) are satisfied. Suppose that $q < 1/4$ where q is defined in Assumption 3(vi). Then there exist permutations σ_B : $\mathbb{G}^B \mapsto \mathbb{G}^B$ and $\sigma_A : \mathbb{G}^A \mapsto \mathbb{G}^A$ such that $\left\|\beta^0_{g,B} - \hat{\beta}_{\sigma_B(g),B}\right\|^2 = O_p(1/\sqrt{T})$ for any $g \in \mathbb{G}^B$ and $\left\|\beta^0_{g,A} - \hat{\beta}_{\sigma_A(g),A}\right\|^2 = O_p(1/\sqrt{T})$ for any $g \in \mathbb{G}^A$.*

Proof. The proof is identical to that of Lemma 3 until we observe that

$$\left\|\hat{\beta}_{\sigma_B(g),B} - \hat{\beta}_{\sigma_B(\tilde{g}),B}\right\| \geq \left\|\beta^0_{g,B} - \beta^0_{\tilde{g},B}\right\| - \left\|\beta^0_{g,B} - \hat{\beta}_{\sigma_B(g),B}\right\| - \left\|\beta^0_{\tilde{g},B} - \hat{\beta}_{\sigma_B(\tilde{g}),B}\right\|.$$

Recall that we have already shown that $\left\|\beta^0_{g,B} - \hat{\beta}_{\sigma_B(g),B}\right\| = O_p(1/\sqrt{T})$ and $\left\|\beta^0_{\tilde{g},B} - \hat{\beta}_{\sigma_B(\tilde{g}),B}\right\| = O_p(1/\sqrt{T})$. Besides, Assumption 3(*viii*) states that $\left\|\beta^0_{g,B} - \beta^0_{\tilde{g},B}\right\| > c_T$. Because $c_T > T^{-1/2+e}$, the right hand side of the above inequality is strictly positive with probability approaching one. This means that $\sigma_B(g) \neq \sigma_B(\tilde{g})$ for $g \neq \tilde{g}$ with probability approaching one, which further implies that σ_B admits a well defined inverse and is bijective. The rest of the proof is identical to the corresponding part of the proof of Lemma 3.

□

Lemmas 6 and 7 imply that the Hausdorff distance between β^0_B and $\hat{\beta}_B$ and the distance between β^0_A and $\hat{\beta}_A$ both converge to 0 at the rate of $\sqrt{T}$. We also relabel the groups such that $\sigma_B(g) = g$ and $\sigma_A(g) = g$ and we have $\left\|\beta^0_{g,B} - \hat{\beta}_{g,B}\right\|^2 = O_p(1/\sqrt{T})$ for any $g \in \mathbb{G}^B$ and $\left\|\beta^0_{g,A} - \hat{\beta}_{g,A}\right\|^2 = O_p(1/\sqrt{T})$ for any $g \in \mathbb{G}^A$.

Let $\mathcal{N}$ be a neighborhood of β^0 such that $\left\|\beta^0_{g,C} - \beta_{g,C}\right\| < \eta = T^{-1/2+f}$ for $0 < f < e$, where e is defined in Assumption 3(*viii*), for any $g \in \mathbb{G}^C$ and $C = B, A$. Note that we can take η in the above range by considering large N and T by Lemma 7. Let $\bar{k} = k^0 + T^{1/2+q}\log T$ and $\underline{k} = k^0 - T^{1/2+q}\log T$. Define $K = \{k : \underline{k} \leq k \leq \bar{k}\}$.

Lemma 8. *Suppose that Assumptions 1(ii), 1(iv), 1(vii), 3(viii), and 3(ix) hold. We take q and f in the definitions of K and $\mathcal{N}$, respectively, such that $q < e$ and $e - 2q - 1/4 < f < e - q$ where e is defined in Assumption 3(viii). As $N, T \to \infty$ with $NT^{-er} \to 0$, where r is defined in Assumption 3(ix), it holds that*

$$\Pr\left\{\hat{\gamma}(k,\beta) \neq \gamma^0 \text{ for some } k \in K \text{ and } \beta \in \mathcal{N}\right\} \to 0.$$

Proof. To show this probability converges to zero, it is equivalent to show that

$$\max_{1\leq i\leq N} \sup_{\beta\in\mathcal{N}} \max_{k\in K} \mathbf{1}\{\hat{g}_i(B)(k,\beta) \neq g^0_i(B)\} + \max_{1\leq i\leq N} \sup_{\beta\in\mathcal{N}} \max_{k\in K} \mathbf{1}\{\hat{g}_i(A)(k,\beta) \neq g^0_i(A)\} = o_p(1),$$

where we observe that

$$\mathbf{1}\left\{\hat{g}_i(B)(k,\beta) \neq g^0_i(B)\right\} = \max_{g\in\mathbb{G}^B\setminus\{g^0_i(B)\}} \mathbf{1}\left(\sum_{t=1}^{k-1}(y_{it} - x'_{it}\beta_{g,B})^2 < \sum_{t=1}^{k-1}(y_{it} - x'_{it}\beta_{g^0_i(B),B})^2\right), \tag{A.11}$$

and a similar equality holds for $\mathbf{1}\{\hat{g}_i(A)(k,\beta) \neq g_i^0(A)\}$. We analyze the probability of each of these two indicators being one. To this end, we first evaluate how the deviation of k from k^0 plays a role, while the situation of $k = k^0$ can be analysed using the same arguments as in Bonhomme and Manresa (2015) and Okui and Wang (2021).

We first examine the difference between the two summations in the argument of the indicator function in (A.11), and show that this difference evaluated at any $k \in K$ and that evaluated at $k = k^0$ are not very different. Let

$$D = \sum_{t=1}^{k-1} \left((y_{it} - x_{it}'\beta_{g,B})^2 - (y_{it} - x_{it}'\beta_{g_i^0(B),B})^2 \right) - \sum_{t=1}^{k^0-1} \left((y_{it} - x_{it}'\beta_{g,B})^2 - (y_{it} - x_{it}'\beta_{g_i^0(B),B})^2 \right).$$

First, considering the case of $k < k^0$, we have that

$$\begin{aligned} |D| =& \left| \sum_{t=k}^{k^0-1} 2u_{it}x_{it}(\beta_{g_i^0(B),B} - \beta_{g,B}) + \sum_{t=k}^{k^0-1} (\beta_{g_i^0(B),B} - \beta_{g,B})' x_{it}x_{it}'(2\beta^0_{g_i^0(B),B} - \beta_{g_i^0(B),B} - \beta_{g,B}) \right| \\ \leq& M_1(k^0 - \underline{k}) \frac{1}{k^0 - \underline{k}} \sum_{t=\underline{k}}^{k^0-1} \|u_{it}x_{it}\| + M_2(k^0 - \underline{k}) \left\| \frac{1}{k^0 - \underline{k}} \sum_{t=\underline{k}}^{k^0-1} x_{it}x_{it}' \right\|, \end{aligned}$$

where M_1 and M_2 are constants independent of (i, g, k, β). Let $M_T = T^f / \log T$. By the Markov inequality, we have

$$\Pr \left(\frac{1}{k^0 - \underline{k}} \left| \sum_{t=\underline{k}}^{k^0-1} (\|u_{it}x_{it}\| - E(\|u_{it}x_{it}\|) \right| > M_T \right) \leq \frac{E\left(\left| \sum_{t=\underline{k}}^{k^0-1} (\|u_{it}x_{it}\| - E(\|u_{it}x_{it}\|) \right|^r \right)}{((k^0 - \underline{k})M_T)^r}.$$

Under Assumption 3(*ix*), we can apply Theorem 1 of Yokoyama (1980) to obtain

$$E\left(\left| \sum_{t=\underline{k}}^{k^0-1} (\|u_{it}x_{it}\| - E(\|u_{it}x_{it}\|) \right|^r \right) \leq C(k^0 - \underline{k})^{r/2}.$$

Noting that $(k^0 - \underline{k})^{-1} \sum_{t=\underline{k}}^{k^0-1} E(\|u_{it}x_{it}\|)$ converges and $M_T \to \infty$, we have that

$$\Pr \left((k^0 - \underline{k})^{-1} \sum_{t=\underline{k}}^{k^0-1} \|u_{it}x_{it}\| > M_T \right) = O\left(T^{-r/4 - qr/2 - rf} (\log T)^{r/2} \right).$$

Similarly, it holds that $\Pr(\|(k^0 - \underline{k})^{-1} \sum_{t=\underline{k}}^{k^0-1} x_{it}x_{it}'\| > M_T) = O(T^{-r/4-qr/2-rf}(\log T)^{r/2})$. These imply that there exists a sequence that satisfies $C_T = O(M_T)$ and $C_T \to \infty$ as $T \to \infty$, such that

$$\Pr \left(\frac{1}{k^0}|D| > \frac{k^0 - \underline{k}}{k^0} C_T \right) = O(T^{-r/4-qr/2-rf}(\log T)^{r/2}).$$

Using a similar argument, we can show that for $k \geq k^0$,

$$\Pr\left(\frac{1}{k^0}|D| > \frac{\bar{k}-k^0}{k^0}C_T\right) = O(T^{-r/4-qr/2-rf}(\log T)^{r/2}).$$

Next, we consider $\sum_{t=1}^{k^0-1}\left((y_{it}-x_{it}'\beta_{g,B})^2-(y_{it}-x_{it}'\beta_{g_i^0(B),B})^2\right)$. As in the proof of Lemma 4, we have

$$\begin{aligned}
&\sum_{t=1}^{k^0-1}\left((y_{it}-x_{it}'\beta_{g,B})^2-(y_{it}-x_{it}'\beta_{g_i^0(B),B})^2\right)\\
&=\sum_{t=1}^{k^0-1}2u_{it}x_{it}(\beta^0_{g_i^0(B),B}-\beta^0_{g,B})+\sum_{t=1}^{k^0-1}(x_{it}'(\beta^0_{g_i^0(B),B}-\beta^0_{g,B}))^2+\Psi,
\end{aligned}$$

where

$$\begin{aligned}
\Psi=&\sum_{t=1}^{k^0-1}2u_{it}x_{it}(\beta_{g_i^0(B),B}-\beta_{g,B}-\beta^0_{g_i^0(B),B}+\beta^0_{g,B})\\
&+\sum_{t=1}^{k^0-1}(\beta_{g_i^0(B),B}-\beta_{g,B}-\beta^0_{g_i^0(B),B}+\beta^0_{g,B})'x_{it}x_{it}'(2\beta^0_{g_i^0(B),B}-\beta_{g_i^0(B),B}-\beta_{g,B})\\
&+\sum_{t=1}^{k^0-1}(\beta^0_{g_i^0(B),B}-\beta^0_{g,B})'x_{it}x_{it}'(\beta^0_{g_i^0(B),B}-\beta_{g_i^0(B),B}-\beta_{g,B}+\beta^0_{g,B}).
\end{aligned}$$

By the Cauchy-Schwarz inequality, Assumption 1(*ii*) and the definition of $\mathcal{N}$ imply that

$$|\Psi| \leq \eta C_1\left\|\sum_{t=1}^{k^0-1}u_{it}x_{it}\right\|+\eta C_2\left\|\sum_{t=1}^{k^0-1}x_{it}x_{it}'\right\|,$$

where C_1 and C_2 are constants independent of η and T. We then have that

$$\begin{aligned}
&\mathbf{1}\left(\sum_{t=1}^{k-1}(y_{it}-x_{it}'\beta_{g,B})^2<\sum_{t=1}^{k-1}(y_{it}-x_{it}'\beta_{g_i^0(B),B})^2\right)\\
\leq&\mathbf{1}\Bigg(\sum_{t=1}^{k^0-1}2u_{it}x_{it}'(\beta^0_{g_i^0(B),B}-\beta^0_{g,B})\\
&\quad-\sum_{t=1}^{k^0-1}(x_{it}'(\beta^0_{g_i^0(B),B}-\beta^0_{g,B}))^2+\eta C_1\left\|\sum_{t=1}^{k^0-1}u_{it}x_{it}\right\|+\eta C_2\left\|\sum_{t=1}^{k^0-1}x_{it}x_{it}'\right\|+|D|\Bigg).
\end{aligned}$$

Note that the right hand side does not depend on β. Thus, as in the proof of Lemma 4, we have

$$\Pr\left(\sup_{\beta\in\mathcal{N}}\max_{k\in K}\mathbf{1}(\hat{g}_i(B)(k,\beta)\neq g_i^0(B))\neq 0\right)$$

$$\leq \sum_{g \in \mathbb{G}^B \setminus \{g_i^0(B)\}} \left(\Pr\left(\frac{1}{k^0} \sum_{t=1}^{k^0-1} (x_{it}'(\beta^0_{g_i^0(B),B} - \beta^0_{g,B}))^2 \leq \frac{c_T''}{2} \right) + \Pr\left(\left\| \frac{1}{k^0} \sum_{t=1}^{k^0-1} u_{it}x_{it} \right\| \geq M \right) \right.$$

$$+ \Pr\left(\left\| \frac{1}{k^0} \sum_{t=1}^{k^0-1} x_{it}x_{it}' \right\| \geq M \right) + \Pr\left(\frac{1}{k^0}|D| > \frac{k^0 - \underline{k}}{k^0} C_T \right)$$

$$\left. + \Pr\left(\frac{1}{k^0} \sum_{t=1}^{k^0-1} 2u_{it}x_{it}'(\beta^0_{g_i^0(B),B} - \beta^0_{g,B}) < -\frac{c_T''}{2} + \eta C_1 M + \eta C_2 M + \frac{k^0 - \underline{k}}{k^0} C_T \right) \right),$$

where we take $c_T'' = c_T \times \rho^*$ for c in Assumption 3(*viii*) and ρ^* in Assumption 1(*iv*).

We observe that

$$\Pr\left(\left\| \frac{1}{k^0} \sum_{t=1}^{k^0-1} x_{it}x_{it}' \right\| \geq M \right) \leq \Pr\left(\frac{1}{k^0} \sum_{t=1}^{k^0-1} \| x_{it}x_{it}' \| \geq M \right) = \Pr\left(\frac{1}{k^0} \sum_{t=1}^{k^0-1} x_{it}'x_{it} \geq M \right).$$

We then apply the Markov inequality and Theorem 1 of Yokoyama (1980) with respect to $x_{it}'x_{it} - E(x_{it}'x_{it})$ so that under Assumption 3(*ix*) it holds that

$$\Pr\left(\left\| (k^0)^{-1} \sum_{t=1}^{k^0-1} x_{it}x_{it}' \right\| \geq M \right) = O\left((k^0)^{-r/2} \right) = O\left(T^{-r/2} \right),$$

where the last equality holds by Assumption 1(*vii*). Similarly, Assumption 1(*ix*) also implies that $\Pr\left(\left\| (k^0)^{-1} \sum_{t=1}^{k^0-1} u_{it}x_{it} \right\| \geq M \right) = O\left(T^{-r/2} \right)$. Moreover, a similar argument shows that under Assumptions 1(*iv*), 3(*viii*) and 3(*ix*), the Markov inequality combined with Theorem 1 of Yokoyama (1980) yields that

$$\Pr\left(\left| \frac{1}{k^0} \sum_{t=1}^{k^0-1} (x_{it}'(\beta^0_{g_i^0(B),B} - \beta^0_{g,B}))^2 - \frac{1}{k^0} \sum_{t=1}^{k^0-1} E((x_{it}'(\beta^0_{g_i^0(B),B} - \beta^0_{g,B}))^2) \right| \geq \frac{c_T''}{2} \right) = O(T^{-er}),$$

which in turn implies that the following equation holds uniformly over g:

$$\Pr\left(\frac{1}{k^0} \sum_{t=1}^{k^0-1} (x_{it}'(\beta^0_{g_i^0(B),B} - \beta^0_{g,B}))^2 \leq \frac{c_T''}{2} \right) = O(T^{-er}).$$

Now we have shown that $\Pr\{(k^0)^{-1}|D| > ((k^0 - \underline{k})/k^0)C_T\} = O(T^{-r/4-qr/2-rf}(\log T)^{r/2})$. Moreover, by similar arguments as above that use the combination of the Markov inequality and Theorem 1 of Yokoyama (1980) we have

$$\Pr\left(\frac{1}{k^0} \sum_{t=1}^{k^0-1} 2u_{it}x_{it}'(\beta^0_{g_i^0(B),B} - \beta^0_{g,B}) < -\frac{c_T''}{2} + \eta C_1 M + \eta C_2 M + \frac{k^0 - \underline{k}}{k^0} C_T \right)$$

$$\leq \Pr\left(\frac{1}{k^0} \sum_{t=1}^{k^0-1} 2u_{it}x_{it}'(\beta^0_{g_i^0(B),B} - \beta^0_{g,B}) < -\frac{c_T''}{4} \right) = O(T^{-er})$$

uniformly over g under Assumption 3(*ix*), where the first inequality holds because $c''_T = O(c_T) = O(T^{-1/2+e})$, $\eta = o(T^{-1/2+e})$, and $((k^0 - \underline{k})/k^0)C_T = O(T^{-1/2+q+f}) = o(c_T)$ under the assumption of $q + f < e$. It thus follows that

$$
\begin{aligned}
&\Pr\left(\max_{1\leq i\leq N}\sup_{\beta\in\mathcal{N}}\max_{k\in K}\mathbf{1}(\hat{g}_i(B)(k,\beta)\neq g_i^0(B))\neq 0\right)\\
\leq&\sum_{i=1}^{N}\Pr\left(\sup_{\beta\in\mathcal{N}}\max_{k\in K}\mathbf{1}(\hat{g}_i(B)(k,\beta)\neq g_i^0(B))\neq 0\right)\\
=&O(N(T^{-er}+T^{-r/2}+T^{-r/4-qr/2-fr}(\log T)^{r/2}))=O(NT^{-er}),
\end{aligned}
$$

where the last equality follows because $1/4 + 2q + f > e$ by the choice of f. Similarly, we can show that $\Pr\left(\max_{1\leq i\leq N}\sup_{\beta\in\mathcal{N}}\max_{k\in K}\mathbf{1}\left[\hat{g}_i(A)(k,\beta)\neq g_i^0(A)\right]\neq 0\right) = O(NT^{-er})$.

□

Proof of Theorem A.1

Proof. We first note that we can take f to satisfy $e - 2q - 1/4 < f < e - q$ under the condition of the theorem so that Lemma 8 can be applied. The proof is almost identical to that of Theorem 1. There are three main differences. First, Lemmas 6, 7 and 8 are used instead of 2, 3 and 4. Second, the lower bound on $\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}(x'_{it}(\beta_{g_i^0(A),A}-\beta_{g_i^0(B),B}))^2/(NT)$ is given by $(k-k^0)\left(\underline{m}-C\eta\right)/T$ because of Assumption 3(*vi*), and it is replaced by $(k-k^0)\underline{m}/(2T)$ because $f < 1/2 - q$ is assumed. Third, we have

$$
\begin{aligned}
&\Pr\left(\sup_{\beta\in\mathcal{N}}\max_{k^0<k\leq\bar{k}}\left(-2\frac{1}{N}\frac{1}{k-k^0}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}x'_{it}(\beta_{g_i^0(A),A}-\beta_{g_i^0(B),B})u_{it}\right)>\frac{\underline{m}}{2}\right)\\
\leq&\Pr\left(\max_{k^0<k\leq\bar{k}}\left(-2\frac{1}{N}\frac{1}{k-k^0}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}x'_{it}(\beta^0_{g_i^0(A),A}-\beta^0_{g_i^0(B),B})u_{it}\right)>\frac{\underline{m}}{4}\right)\\
&+\Pr\left(\eta C\max_{k^0<k\leq\bar{k}}\left\|\frac{1}{N}\frac{1}{k-k^0}\sum_{t=k^0}^{k-1}\sum_{i=1}^{N}x_{it}u_{it}\right\|>\frac{\underline{m}}{4}\right)=O\left(\frac{T^{2q}}{N}\right),
\end{aligned}
$$

where the last equality follows by applying Bai and Perron (1998, Lemma A.6), noting that under Assumptions 3(*ix*) and 1(*x*), $x_{it}u_{it}$ is an L_2 mixingale. Also note that $T^{2q}/N \to 0$ is assumed in the statement of the theorem.

□

Proof of Corollary A.1

Proof. The proof is identical to that of Corollary 1 except that the current proof is based on Theorem A.1 while that of Corollary 1 is based on Theorem 1.

□