Universitat Pompeu Fabra Barcelona
Department of Economics and Business

Economics Working Paper Series

Working Paper No. 1954

Grouped fixed-effects estimation is more robust than you think

Oriol González-Casasús

October 2026

Grouped Fixed-Effects Estimation Is More Robust Than You Think

Oriol González-Casasús*

Universitat Pompeu Fabra and Barcelona School of Economics

September 30, 2026

Abstract

The grouped fixed-effects (GFE) approach of Bonhomme and Manresa (2015) has become an increasingly popular tool in applied economics for modeling latent heterogeneity in panel data, yet its implementation raises two recurring concerns: how many groups to select, and whether groups are sufficiently distinct to be meaningfully recovered. It is shown here that both concerns may be largely immaterial for inference on common parameters. First, the GFE estimator remains asymptotically equivalent to the oracle that knows the true groups even when the number of groups is overspecified, provided the additional groups fit idiosyncratic noise rather than latent structure in the covariates. Second, oracle equivalence continues to hold under weak group separation provided the cross-section does not grow too fast relative to the time dimension, at a rate that depends on the degree of separation. Simulations confirm that the robustness properties emerge quickly in finite samples, and identify the situations in which they fail. The paper concludes with updated guidance for applied researchers, revisiting the income-democracy debate as an illustration. Overall, the findings reframe GFE as a more reliable and user-friendly tool for applied work than previously recognized.

JEL Classification: C23, C38, C13, C52.

Key words: Fixed-Effects, Robust Econometrics, Weak Separation, Misspecification, Information Criteria, Latent Grouped Heterogeneity.

*Department of Economics and Business, Universitat Pompeu Fabra. Email: oriol.gonzalez@upf.edu. This paper is based on materials in the appendix to Chapter 1 of my PhD dissertation at the University of Pennsylvania. I thank Frank Schorfheide for helpful suggestions. I acknowledge financial support from the Spanish Agencia Estatal de Investigación, through the Severo Ochoa Programme for Centres of Excellence in R&D (Barcelona School of Economics CEX2024-001476-S).

1 Introduction

Panel data are invaluable in empirical economics for their ability to account for unobserved heterogeneity across individuals, firms, or other entities observed over time. By exploiting both cross-sectional and temporal variation, panel datasets enable the identification of latent traits that, if ignored, could render estimates biased or inconsistent, hence compromise causal interpretation. Traditional fixed-effects estimators mitigate such biases by allowing for unit-specific intercepts, but this comes at the cost of estimating a large number of parameters and could lead to the well-known incidental parameters problem. The grouped fixed-effects (GFE) estimator introduced by Bonhomme and Manresa (2015) (hereafter BM) offers a compelling alternative by assuming that unobserved heterogeneity can be parsimoniously captured through a finite number of latent types or groups. This approach retains the benefits of fixed-effects while uncovering economically meaningful patterns of heterogeneity, allowing researchers to summarize high-dimensional unobserved variation in a tractable and interpretable way. This paper demonstrates that the GFE estimator of common slope parameters possesses two robustness properties not originally acknowledged that may substantially broaden its relevance and usefulness in applied work, and change the practice of GFE estimation as a byproduct.

The GFE approach simultaneously assigns individuals to groups and estimates (time-varying) group-specific fixed-effects, treating group membership as an unknown to be recovered from the data. As such, GFE estimators provide a powerful framework for empirical work where the nature of heterogeneity itself is of substantive interest. Despite its appeal, the GFE estimator has sparked methodological discussion in the literature. A central concern precisely involves the identification and interpretation of the latent groups: while the grouping structure may capture meaningful economic variation, it is fundamentally data-driven and may reflect artifacts of the estimation procedure or misspecification rather than true latent types. Accurate identification of latent types is crucial, since poor classification may in turn harm the estimates of common slope parameters—the grouped structure can be regarded as a nuisance parameter.

In particular, the implementation of the GFE estimator may raise at least two practical challenges. A first recurring concern among practitioners is the choice of the number of latent groups, which is typically unknown ex ante. Indeed, BM write: *"The choice of [the number of groups] G, and the related issue of how inference on the model's parameters is affected by this choice, are difficult questions that deserve further investigation."* Although applied researchers tend to resort to information criteria to select this unknown quantity, these methods can be sensitive to specification choices and introduce the usual model-selection risks for inference (Leeb and Pötscher, 2005). A second common concern arises when groups in the data happen to look insufficiently different from each other, a problem referred to as weak group separation, challenging correct classification. This is particularly problematic in short panels, e.g., when firms or individuals exhibit similar but not identical time patterns, and might lead to inconsistent parameter estimates (cf. Onatski, 2012). It could also pose problems in long panels if there are convergence clubs (Phillips and Sul, 2009). BM, too, acknowledge this difficulty: *"[The strong separation] condition is instrumental to derive asymptotic [oracle] equivalence. (…) Providing methods to test for group separation, and to achieve valid inference on the model's parameters when it fails, are interesting questions for future work."*

The main contribution of this paper is to derive two new results on the robustness properties of the GFE estimator that address the aforementioned concerns. The performance of the GFE estimator is assessed relative to that of an oracle estimator that uses the true group memberships—a natural benchmark for

evaluating whether the recovery of latent heterogeneity affects estimation. BM show that the GFE estimator is asymptotically equivalent to the oracle under correct specification; i.e., when the true number of groups is known and groups are strongly separated. It is shown here that such an equivalence continues to hold for common slope parameters even when (i) the number of groups used in the estimation is overspecified, under conditions that supplement those in BM with a restriction on the interplay between the covariates and the errors; and (ii) groups are weakly separated, albeit at the cost of stronger conditions on the relative dimensions of the panel. Importantly, the robustness properties concern common parameters only and do not carry over to estimated group-specific parameters, whose distribution crucially depends on the group structure (see Corollary 1 in BM). These properties render the aforementioned concerns largely irrelevant for inference on slope parameters hence substantially alleviate two of the biggest practical obstacles in applying the GFE estimator.

The key to demonstrate property (i) is the derivation of a new ANOVA-type decomposition of the GFE estimator which exploits the grouped fixed-effects structure. Such a decomposition allows to compare both the Gram matrices and scores under any refinement of the true group partition and under the true grouping structure. This is necessary to study the asymptotic behavior of the difference of GFE estimators that use different numbers of groups. The comparison is then carried out at the partition selected by the GFE estimator, exploiting that it minimizes the GFE objective. On the other hand, property (ii) is demonstrated by carefully balancing the rate at which the distance across groups vanishes with that at which consistent classification remains possible.

The novel properties uncovered herein enhance the appeal of the GFE estimator and should encourage practitioners to confidently use it in a broad range of empirical settings. At the same time, they yield important implications for applied practice and redefine how the GFE estimator should be used. For instance, although researchers often face ambiguity about the correct number of unobserved types, property (i) eliminates the need for consistent model-selection procedures or cross-validation, which may be computationally burdensome and statistically delicate (see, e.g., Onatski, 2010; Ahn and Horenstein, 2013; Moon and Weidner, 2015). Moreover, property (ii) guarantees that GFE estimates remain valid even when groups are not strongly distinguishable in finite samples, so long as the cross-section does not grow too fast relative to the time dimension.

Related Literature. Analogous properties have been derived for other estimators. Similarly to property (i), Moon and Weidner (2015) show that overspecifying the number of factors has no effect on the estimation of slope parameters in interactive fixed-effects models. While grouped fixed-effects are a special case of interactive fixed-effects under a constrained factor structure, the estimation strategy is significantly different. Liu, Shang, Zhang and Zhou (2020) study classification consistency under group overspecification in a general latent class framework. The current paper extends those insights to panel models with time-varying heterogeneity, and leverages the particular structure of such models to derive explicit consequences for parameter estimation. It is also worth noting that BM have a discussion on overspecification of groups in their Supplemental Material S.3 and raise the result as a conjecture, which is proved here under a condition on the interplay between the covariates and the errors.

To achieve the analogous of property (i), Moon and Weidner (2015) employ a high-level condition on the eigenvectors of the error covariance matrices (their Assumption EV), which requires the additional factors, estimated from the residuals, to be asymptotically uncorrelated with the regressors, and which they verify

only for i.i.d. normal errors through random matrix theory. In the grouped fixed-effects case, the special structure is that residualization is by cross-sectional averages over (sub)groups. By the novel ANOVA-type decomposition, the extra variation induced by refining the partition—i.e., overspecifying the number of groups—shows up as between-(sub)group means. Those are sample averages across units, and because of mixing across t, averaging across time further shrinks their variance. Thus, their contribution to the score vanishes faster than sampling error. This elementary variance-shrinking mechanism is why one can avoid strong Gaussian assumptions, and simply rely on law-of-large-numbers/mixing bounds for averages. The mechanism operates exactly at the classification that the GFE estimator would select if the slope were known, whose cells depend on the idiosyncratic errors only. That the estimated classification behaves in the same way—rather than picking up a latent structure of the covariates through the estimated slope—is a condition on the interplay between covariates and errors, stated as Assumption 5 in Appendix A and discussed in Section 3.

Few papers consider the consequences of weak group separation on parameter estimates as in property (ii). Dzemski and Okui (2024) and Lumsdaine, Okui and Wang (2023) consider models with group-specific regression coefficients—in the supplemental materials, they extend their frameworks to allow for weak separation and show oracle equivalence. Unlike in their models, time variation of group-specific parameters is unrestricted in this paper, which leads to different rates in the results. Another important difference is that the ultimate focus here is on common rather than group-specific objects.

The rest of the paper is organized as follows. Section 2 describes the grouped panel data model under consideration and defines the GFE estimator. Sections 3 and 4 formally state and discuss each of the robustness properties, while Section 5 provides supporting simulation evidence. Section 6 illustrates the implications for applied practice, and Section 7 concludes by summarizing the recommendations for practitioners. All the proofs are relegated to the Appendix. For ease of comparability, I impose the exact same assumptions as in BM whenever possible.

2 Model and GFE Estimator

Consider a linear panel data model with grouped fixed-effects:

$$y_{it} = x_{it}'\beta^0 + \alpha^0_{g_i^0 t} + v_{it}, \qquad i \in [N],\ t \in [T],\ g \in \left[G^0\right], \tag{1}$$

where x_{it} denotes observed covariates, β^0 is a vector of common parameters, $\alpha^0_{g_i^0 t}$ captures unobserved group-specific time effects, which may be arbitrarily correlated with the observed covariates, and v_{it} is an idiosyncratic error. Each unit i is assigned to one of the G^0 latent groups in an unrestricted way.

The GFE estimator proposed by BM for model (1) seeks to jointly recover the common parameters, group-time effects, and latent group assignments by solving:

$$\left(\hat{\beta}, \hat{\alpha}, \hat{g}\right) = \underset{(b,a,g)\in\mathcal{B}\times\mathcal{A}^{GT}\times\Gamma_G}{\arg\min} \sum_{i=1}^{N}\sum_{t=1}^{T} \left(y_{it} - x_{it}'b - a_{g_i t}\right)^2, \tag{2}$$

where $\mathcal{B}$ and $\mathcal{A}$ are some parameter spaces, and Γ_G is the set of all possible partitions of the N units into G groups.

The minimization problem in (2) is typically solved by an iterative procedure that alternates between updating group assignments and estimating parameters conditional on those assignments; see Algorithm 1 in BM. This structure highlights the importance of accurately recovering the latent grouping: misclassification can impair both the interpretation and precision of the estimated parameters. Indeed, the performance of the GFE estimator is typically compared against that of an oracle estimator that assumes knowledge of the true group memberships:

$$\left(\tilde{\beta}, \tilde{\alpha}\right) = \underset{(b,a)\in\mathcal{B}\times\mathcal{A}^{G^0 T}}{\arg\min} \sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - x_{it}'b - a_{g_i^0 t}\right)^2. \tag{3}$$

By using the true group structure, the oracle serves as a lower bound on estimation error and thus provides a benchmark for parameter estimation that abstracts from the uncertainty of recovering the latent heterogeneity. Under correct model specification, BM demonstrate that the GFE estimator $\hat{\beta}$ is asymptotically equivalent to the oracle $\tilde{\beta}$. The rest of this paper shows that such an equivalence remains true under different plausible forms of model misspecification.

3 Overspecification of Groups

A central operational challenge in the application of the GFE estimator is the selection of the number of groups, G. In practice, the risks and detectability of misspecifying G are highly asymmetric. Underspecifying the number of groups ($G < G^0$) forces heterogeneous units to share time-profiles, leaving unabsorbed heterogeneity that induces an $O(1)$ asymptotic bias in the common slope parameters. Fortunately, this misspecification is straightforward to detect, as it results in a macroscopic penalty to the sum of squared residuals. In contrast, overspecifying the number of groups ($G > G^0$) perfectly absorbs the latent heterogeneity, meaning any further reduction in the objective function is driven solely by the fitting of idiosyncratic noise. Because finite-sample noise easily obscures the true rank of the heterogeneity, data-driven selection procedures frequently struggle to distinguish G^0 from larger models, often leading to overestimation in applied settings.[1] Confronted with this asymmetry, researchers naturally prefer to err on the side of choosing a generously large G to avoid the severe omitted variable bias of underspecification. This empirical reality motivates investigating the behavior of the GFE estimator under an overspecified number of groups.

Let $\hat{\beta}_G$ denote the GFE estimator computed using G groups. BM demonstrate that, under some regularity conditions, the estimator $\hat{\beta}_{G^0}$ is consistent and asymptotically equivalent to the oracle $\tilde{\beta}$. When the number of groups is underspecified, $G < G^0$, $\hat{\beta}_G$ is generally inconsistent for β^0 if the unobserved effects are correlated with the observed covariates, due to omitted variable bias. In contrast, when overspecifying the number of groups, $G > G^0$, $\hat{\beta}_G$ remains consistent for β^0 under the conditions of Theorem 1 in BM, since the proof of that theorem is unaffected in this case.

Importantly, it is demonstrated here that overspecifying the number of groups, i.e., estimating the model with $G \geq G^0$ groups, does not have any consequences on the asymptotic distribution of the GFE estimator either, which remains equivalent to the oracle's.

[1] See Section S.3 in the Supplementary Appendix of BM and, for the analogous problem of selecting the number of factors, Bai and Ng (2002). Appendix C provides numerical evidence in the design of Section 5.

To establish the result, I adopt the same set of assumptions as in BM, which impose certain mixing and moment conditions, complemented by two conditions that are stated and discussed in Appendix A (Assumptions 4 and 5); I refer the reader to BM for a detailed discussion and interpretation of their conditions.

Assumption 1 (Assumption 1 in BM)**.** *There exists a constant $M > 0$ such that:*

(a) $\mathcal{B}$ *and* $\mathcal{A}$ *are compact subsets of* $\mathbb{R}^d$ *and* $\mathbb{R}$*, respectively;*

(b) $\mathbb{E}[\|x_{it}\|^2] \leq M$ *and* $\mathbb{E}[x_{it}v_{it}] = 0$*;*

(c) $\mathbb{E}[v_{it}] = 0$ *and* $\mathbb{E}[v_{it}^4] \leq M$*;*

(d) $\left|(NT)^{-1}\sum_{i=1}^N\sum_{t=1}^T\sum_{s=1}^T\mathbb{E}[v_{it}v_{is}x_{it}'x_{is}]\right| \leq M$*;*

(e) $N^{-1}\sum_{i=1}^N\sum_{j=1}^N\left|T^{-1}\sum_{t=1}^T\mathbb{E}[v_{it}v_{jt}]\right| \leq M$*;*

(f) $\left|N^{-2}T^{-1}\sum_{i=1}^N\sum_{j=1}^N\sum_{t=1}^T\sum_{s=1}^T\mathrm{Cov}(v_{it}v_{jt}, v_{is}v_{js})\right| \leq M$*;*

(g) For all groupings $\gamma = \{g_1, \ldots, g_N\} \in \Gamma_G$*, let* $\hat{\rho}(\gamma)$ *be the minimum eigenvalue of*

$$\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T(x_{it} - \bar{x}_{g_i^0 \wedge g_i, t})(x_{it} - \bar{x}_{g_i^0 \wedge g_i, t})'.$$

Then, $\mathrm{plim}_{N,T\to\infty}\min_{\gamma\in\Gamma_G}\hat{\rho}(\gamma) = \rho > 0$*.*

Assumption 2 (Assumption 2 in BM)**.**

(a) For all $g \in [G^0]$*,* $\mathrm{plim}_{N\to\infty} N^{-1}\sum_{i=1}^N \mathbb{1}\{g_i^0 = g\} = \pi_g > 0$*;*

(b) For all $(g, \tilde{g}) \in [G^0]^2$ *such that* $g \neq \tilde{g}$*,* $\mathrm{plim}_{T\to\infty} T^{-1}\sum_{t=1}^T(\alpha_{gt}^0 - \alpha_{\tilde{g}t}^0)^2 = c_{g,\tilde{g}} > 0$*;*

(c) There exist constants $a > 0$ *and* $d_1 > 0$ *and a sequence* $\alpha[t] \leq \exp(-at^{d_1})$ *such that, for all* $i \in [N]$ *and* $(g, \tilde{g}) \in [G^0]$ *such that* $g \neq \tilde{g}$*,* $\{v_{it}\}_t$*,* $\{\alpha_{gt}^0 - \alpha_{\tilde{g}t}^0\}_t$*, and* $\{(\alpha_{gt}^0 - \alpha_{\tilde{g}t}^0)v_{it}\}_t$ *are strongly mixing processes with mixing coefficients* $\alpha[t]$*. Moreover,* $\mathbb{E}[(\alpha_{gt}^0 - \alpha_{\tilde{g}t}^0)v_{it}] = 0$*;*

(d) There exist constants $b > 0$ *and* $d_2 > 0$ *such that* $\mathbb{P}(|v_{it}| > m) \leq \exp(1 - (m/b)^{d_2})$ *for all* i, t*, and* $m > 0$*;*

(e) There exists a constant $M^* > 0$ *such that, as* N, T *tend to infinity,*

$$\sup_{i\in[N]}\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^T\|x_{it}\| > M^*\right) = o_p(T^{-\delta}) \quad \text{for all } \delta > 0.$$

For $g \in [G]$, define $\hat{\mathcal{C}}_g = \{i \in [N] : \hat{g}_i = g\}$, the collection of units belonging to the gth estimated group. Define also $\mathcal{C}_g^0 = \{i \in [N] : g_i^0 = g\}$ for $g \in [G^0]$, the population analog based on the true group membership variables. We have the following result.

Theorem 1. *Suppose* $G \geq G^0$ *and Assumptions 1, 2 and 4 hold. Then, as* $N, T \to \infty$ *such that, for some* $\nu > 0$*,* $NT^{-\nu} = o(1)$*:*

(a) $\mathbb{P}\left(\forall g \in [G],\ \exists \tilde{g} \in [G^0] : \hat{\mathcal{C}}_g \subseteq \mathcal{C}_{\tilde{g}}^0\right) = 1 - o(1)$*.*

(b) If, in addition, Assumption 5 holds and $\sqrt{NT}(\hat{\beta}_{G^0} - \beta^0) = O_p(1)$, $\hat{\beta}_G = \hat{\beta}_{G^0} + o_p(1/\sqrt{NT})$.

Part (a) of Theorem 1 shows that the classification done by the GFE estimator is valid under overspecification of the number of groups. In particular, with probability approaching one, any grouped units asymptotically belong to one same population group. As a consequence, any population group is either identical to an estimated group or is partitioned into subgroups without any misclassification—which is the best result one can expect under $G \geq G^0$. In the special case where $G = G^0$, Theorem 1 implies classification consistency—Theorem 2 in BM.

Part (b) of Theorem 1 implies that the asymptotic distribution of the GFE estimator with $G \geq G^0$ groups is identical to that of the estimator based on the true number of groups. In turn, $\hat{\beta}_G$ is asymptotically equivalent to the oracle estimator $\tilde{\beta}$ per Corollary 1 in BM. The condition $\sqrt{NT}(\hat{\beta}_{G^0} - \beta^0) = O_p(1)$ is standard, and holds under Assumption 3 in BM by their Theorem 3.

The intuition behind the proof of Theorem 1 is as follows. When $G \geq G^0$ and the estimated groups refine the true ones—as happens with probability approaching one—, grouped fixed-effects cancel out exactly in the first-order conditions for (2), and an ANOVA-type decomposition shows that $\hat{\beta}_G$ and $\hat{\beta}_{G^0}$ can differ for two reasons only: (i) the covariate variation absorbed by the refinement, and (ii) the covariate–error covariance it generates. For a refinement that fits the idiosyncratic errors, both are built from cross-sectional averages of errors, whose variance shrinks with N, and are therefore of smaller order than sampling noise. This is why, unlike irrelevant regressors in a generic regression, the additional subgroup–time dummies do not inflate the asymptotic variance: the only new regressors are contrasts whose true coefficients are zero, and they vanish rather than persist. The difficulty is that the estimated refinement is chosen as a function of the data and can align its cells with covariates or errors, so that (i) and (ii) cannot be controlled uniformly over refinements. The proof exploits instead that the estimated refinement minimizes the GFE objective.

The estimated classification is selected jointly with the slope, and a refinement of the true groups that aligns with a latent structure of the covariates and with the errors lowers the sum of squared residuals by creating a spurious covariate–error covariance, at the cost of fitting the errors less well. Assumption 5 requires this cost to dominate: refinements whose loss in the fit to the errors is compensated by the covariate channel must be asymptotically uninformative about the covariates. It is a high-level condition on the interplay between the classification of the errors and the covariates, not implied by Assumptions 1–2, that buys the absence of variance inflation. Appendix A discusses it, verifies it for the error-driven classification under primitive conditions, and works out an example with a latent binary structure in the covariates. Assumption 4 restricts the cross-sectional dependence of the errors and is needed for part (a). Absent this condition, two units from different groups with nearly identical error paths could be allocated to a spare group of their own.

The main takeaway from Theorem 1 for applied researchers is that, for inference on common regression coefficients, it is not necessary to estimate the number of groups consistently. Despite the true number of groups being unknown in practice, it is safe to use a relatively large number of groups to classify the units and to obtain consistent and asymptotically normal parameter estimates. Otherwise, if the number of groups is underspecified, neither the parameter estimation nor the group classification are consistent. This recommendation is summarized in the following corollary.

Corollary 1. *Let* $\widehat{G}$ *be a selection rule that is asymptotically conservative, i.e.,* $\mathbb{P}(\widehat{G} < G^0) = o(1)$. *Then,*

under the conditions of Theorem 1, the GFE estimator under the selection rule $\hat{\beta}_{\widehat{G}}$ is asymptotically equivalent to the oracle $\tilde{\beta}$.

Examples of a valid selection rule $\widehat{G}$ could be a known upper bound on G^0, or an estimate based on information criteria. For instance, BM propose to use the Bayesian information criterion (BIC) following the work of Bai (2009). Note that the proposed selection rule may be inconsistent yet valid for our purposes. In particular, in contrast with the discussion in BM, the rates of N and T are not required to be equal and we may have $N \gg T$ when using the BIC estimate, since $\operatorname{plim}_{N,T\to\infty} \widehat{G} \geq G^0$ turns out to be sufficient.

4 Weak Group Separation

In the grouped panel data framework (1), units belong to a group to the extent that they share time paths of unobserved heterogeneity. The degree of separation across these group-specific paths plays a critical role in the classification step. Specifically, it implies an interplay between the cost (for parameter estimation) and complexity of classification: if all groups exhibit identical fixed-effect trajectories, then accurate classification becomes both infeasible and irrelevant. Conversely, if the trajectories are highly distinct, correct group assignment is both crucial and straightforward. The empirically relevant and most challenging case lies in between, when group-specific time paths are similar but not identical. This section examines the performance of the GFE estimator under a region of this non-trivial regime of weak group separation.

To formalize this environment, I fix the number of groups at the true value, $G = G^0$, and retain Assumption 1.[2] However, I depart from Assumption 2, which imposes strict group separation, and instead adopt a modified version—Assumption 3—that relaxes the separation and dependence requirements. First, Assumption 3(b) weakens the separation condition by allowing the distance between groups to vanish asymptotically. While in finite samples any pair of groups is required to exhibit distinct time paths, this distinction may shrink with growing sample size. This differs from the standard strict separation assumption—Assumption 2(b)—, under which groups remain distinguishable in the limit. Second, Assumption 3(c) relaxes some of the weak dependence conditions in BM, which is useful in our case to obtain explicit rates on the amount of group separation that can be considered, at the expense of assuming strict stationarity.

Assumption 3.

(a) For all $g \in [G]$, $\operatorname{plim}_{N\to\infty} N^{-1} \sum_{i=1}^{N} \mathbb{1}\{g_i^0 = g\} = \pi_g > 0$;

(b) There exists a nonrandom sequence $c_T > (N \wedge T)^{-1/4+e}$ for some $e > 0$ such that for all $(g, \tilde{g}) \in [G]^2$ with $g \neq \tilde{g}$, $T^{-1} \sum_{t=1}^{T} (\alpha_{gt}^0 - \alpha_{\tilde{g}t}^0)^2 > c_T$;

(c) There exist constants $r > 2$, $b > 0$, and a sequence $a[t]$ with $\sum_{t=0}^{\infty} (t+1)^{r/2-1} a[t]^{b/(r+b)} < \infty$ such that, for all $i \in [N]$, $\{v_{it}^2\}_t$ and, for all $(g, \tilde{g}) \in [G]^2$ with $g \neq \tilde{g}$, $\{(\alpha_{gt}^0 - \alpha_{\tilde{g}t}^0) v_{it}\}_t$ are strictly stationary and mixing processes with mixing coefficients bounded by $a[t]$, and have $(r+b)$-th moments bounded uniformly in i; moreover, $\mathbb{E}[(\alpha_{gt}^0 - \alpha_{\tilde{g}t}^0) v_{it}] = 0$;

(d) For r in part (c) and some constant $\tilde{M} > 0$, $\sup_{i\in[N]} \mathbb{P}\left(T^{-1} \sum_{t=1}^{T} \|x_{it}\| \geq \tilde{M}\right) = O(T^{-r/2})$.

Under these less stringent conditions on the groups, we have the following result.

[2] In this section, the subscript in $\hat{\beta}$ is suppressed since there is no confusion anymore.

Theorem 2. *Let Assumptions 1 and 3 hold. As $N, T \to \infty$ with $NT^{-r/2}c_T^{-r} = o(1)$, it holds that:*

(a) $\mathbb{P}\left(\hat{g}_i = g_i^0 \text{ for all } i \in [N]\right) = 1 - o(1)$.

(b) $\hat{\beta} = \tilde{\beta} + o_p(1/\sqrt{NT})$.

Part (a) of Theorem 2 shows that the GFE estimator consistently recovers all group memberships despite groups in the model being weakly separated, so long as the rate at which separation vanishes is slower than $(N \wedge T)^{-1/4}$. The result must be understood up to group relabeling, as usual. Part (b) of Theorem 2 implies that, in this case, the GFE estimator remains asymptotically equivalent to the oracle, hence $\sqrt{NT}$-consistent and asymptotically normal.

Examining the proof of Theorem 2 reveals that the rate at which the distance across groups is allowed to vanish per Assumption 3 is precisely slow enough so that consistent classification remains possible. In that case, it is therefore straightforward to demonstrate oracle behavior. If groups became increasingly similar faster than such a rate, we could not guarantee part (a) in the theorem. However, due to the special idiosyncrasy of grouped fixed-effects discussed earlier in this section, we could still have the result in part (b). Indeed, Section 3 could be understood as an extreme case of weak group separation where the distance across (the extra) groups is exactly zero. Oracle behavior then follows per the same arguments.

Theorem 2 gives hope to practitioners using GFE estimates when groups are at risk of being weakly separated. However, it also makes it clear that, under weak group separation, the requirements on the relative rates of N and T are stronger. With c_T of the order of $(N \wedge T)^{-1/4+e}$—the slowest decay compatible with Assumption 3(b), so that the requirements below are the most stringent ones—, the rate condition of Theorem 2 reads $N = o(T^{2r/(4+r-4re)})$ when $N \leq T$ and $N = o(T^{r/4+re})$ when $N \geq T$. In contrast with BM, where any polynomial rate $N = o(T^{\nu})$ is admissible, the admissible exponent is pinned down by the number of moments r in Assumption 3(c) and shrinks as separation weakens ($e \to 0$). In particular, a panel with $N \geq T$ is admissible only if $r > 4/(1 + 4e)$ (a condition that holds whenever v_{it} has more than four moments, and that requires more moments the weaker the separation), and otherwise T must grow faster than N. The condition has a transparent finite-sample counterpart: $NT^{-r/2}c_T^{-r}$ is a bound on the expected number of misclassified units (Lemma A.6), so that, for a given T and a given separation, a larger cross-section makes misclassification more likely. Empirical researchers must carefully consider the dimensions of their panel when justifying the validity of GFE estimates in this case.

The results in Sections 3 and 4 rely on slightly different assumptions to deviate as little as possible from the original work by BM, hence guarantee the fairest comparison. However, it seems natural to conjecture that both results are compatible: overspecifying the number of groups in a weakly separated environment may not preclude from oracle behavior, provided the relevant conditions hold. Put differently, having some groups suffering from weak and others from extremely weak separation may not preclude from $\sqrt{NT}$-consistency, standard asymptotic normality, and even no variance inflation.[3]

[3]To be more specific, here "weak" refers to the distance $T^{-1}\sum_{t=1}^{T}(\alpha_{gt}^0 - \alpha_{\bar{g}t}^0)^2$ vanishing slower than $(N \wedge T)^{-1/4}$, and "extremely weak" that vanishing faster than (squared) sampling error; see Proposition 1 in Appendix B. The rates in between ("very weak" separation) are covered by neither result. More generally, it remains an open question what is the slowest rate at which the distance between groups can vanish such that oracle equivalence holds for common parameters without uniform consistent classification, or what is the behavior of the GFE estimator when both consistent group classification and oracle equivalence are unattainable.

5 Simulation Evidence

This section validates the previous asymptotic results via finite-sample simulations. Subsection 5.1 deals with the results regarding Theorem 1, while Subsection 5.2 illustrates those involving Theorem 2. Overall, the findings in this section are optimistic for applied work because the robustness properties quickly emerge in small samples.

5.1 Overspecification of Groups

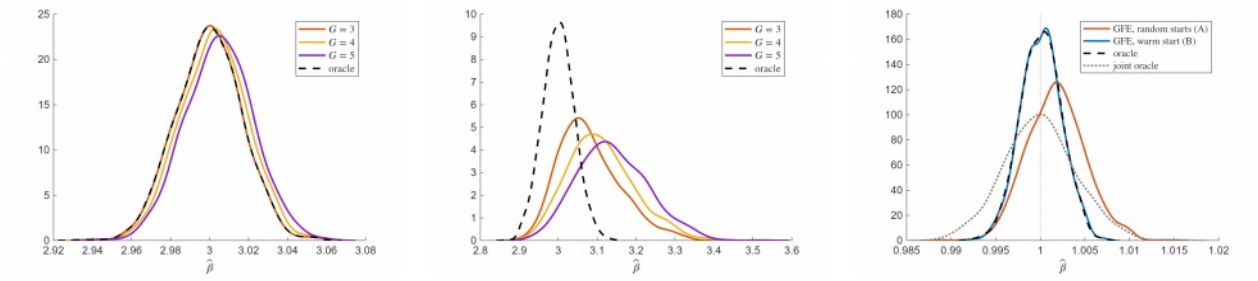
I simulate data according to the following data generating process:

$$y_{it} = x_{it}\beta^0 + \alpha^0_{g_i^0 t} + v_{it}, \qquad x_{it} = \alpha^0_{g_i^0 t} + u_{it}, \qquad i \in [N],\ t \in [T],\ g \in [G^0].$$

I set the dimension of the panel to $(N, T) = (75, 50)$, with $\beta^0 = 3$ and $G^0 = 3$.[4] The variables v_{it}, u_{it} and α^0_{gt} are all drawn from independent standard Normal distributions. True group memberships are drawn uniformly at random on $[G^0]$ independently for each $i \in [N]$.

To verify the implications of Theorem 1, I estimate the model using the GFE estimator with $G \in \{G^0 - 1, \ldots, G^0 + 2\}$ groups, and compute the oracle estimator that uses the true group memberships. I also include the BIC estimate based on the suggestion in BM. I use 1,000 Monte Carlo replications, and the results are presented in Figure I(a); see also Figure C.1 in Appendix C.

We can clearly observe that overspecifying the number of groups does not affect the distribution of the GFE estimator, in line with Theorem 1(b). The results are very encouraging in that not only there is no efficiency loss asymptotically, but also in finite samples. The reason is that, for $G \geq G^0$, the GFE estimator achieves perfect classification, in line with Theorem 1(a); see Table C.1 in Appendix C. The additional results in Figure C.1 provide two further insights about Figure C.1(a). First, underspecifying the number of groups has important consequences due to omitted variable bias, and the distribution of the GFE estimator is no longer centered at the truth. Second, pre-selection based on the BIC has no model-selection consequences either, since it selects $\hat{G} \geq G^0$ in every sample.



(a) $(N, T) = (75, 50)$, $G \geq G^0$ (b) $(N, T) = (75, 9)$, $G \geq G^0$ (c) Latent covariate structure

Figure I: Distribution of the GFE estimator for different number of groups

Appendix C reports two additional experiments, and panels (b) and (c) of Figure I show one case of each. The first varies $T \in \{9, 20, 50\}$ with $N = 75$, where $T = 9$ roughly matches the total sample size of the application in Section 6. With $T = 9$ (panel (b)) classification is poor—13% of the units are misclassified

[4] The value of β^0 is immaterial. Since the objective in (2) satisfies $Q(y + xc; b, a, \gamma) = Q(y; b - c, a, \gamma)$, the GFE estimator is exactly equivariant and the distribution of $\hat{\beta}_G - \beta^0$ does not depend on β^0, for any G.

at $G = G^0-$, all estimators with $G \geq G^0$ are biased, and the bias increases mildly with G; the picture of panel (a) emerges from $T = 20$ onwards, where misclassification is about 1%. The robustness property is a large-T statement and cannot be expected to neutralize the non-classical misclassification biases of very short panels. The second experiment lets the covariates carry a latent binary structure independent of the groups of the outcome equation, $x_{it} = \alpha^0_{g_i^0 t} + \sigma_h \kappa_i h_t + u_{it}$ with $\kappa_i \in \{0, 1\}$, so that the joint structure of (y_{it}, x_{it}) has $2G^0$ cells, and adds as a benchmark the "joint oracle" based on those cells—the least-squares estimator for the fixed informative refinement of Remark 1. In the design of that remark, the GFE estimator with $G > G^0 = 1$ behaves like the oracle at every sample size considered (i.e., the estimated partitions ignore the covariate structure, in line with Assumption 5), whereas the joint oracle has a standard deviation up to 90% larger than the oracle's. With $G^0 = 2$ and a strong covariate structure ($\sigma_h = 3$), however, the outcome depends on how the estimator is computed. Panel (c) shows the case $(N, T) = (300, 200)$, $G = 4$, where Algorithm 1 in BM started at random assignments converges to partitions partly aligned with the covariate structure, whose standard deviation is 1.4 times the oracle's (the variance inflation of Remark 3), whereas the same algorithm initialized at the k-means classification of the residuals of the estimate with $G - 1$ groups converges to error-driven partitions that behave like the oracle. The latter attain the smaller objective in about nine replications out of ten, by a margin of the order of 10^{-3} of the objective. Appendix C documents the mechanism—a path dependence created by the omitted-variable bias of the slope at a random initial partition—, shows that the estimator remains $\sqrt{NT}$-consistent, and describes a diagnostic based on the within-group covariate variation across G. In such designs, the finite-sample behavior of the estimator is governed by the computational strategy, and Assumption 5 cannot be taken for granted.

5.2 Weak Group Separation

I simulate data according to the following data generating process:

$$y_{it} = x_{it}\beta^0 + \alpha^0_{g_i^0 t} + v_{it}, \qquad x_{it} = \alpha^0_{g_i^0 t} + u_{it}, \qquad i \in [N],\ t \in [T],\ g \in [G^0].$$

To model the weak separation problem, I consider the following specification:

$$\alpha^0_{g_i^0 t} = \kappa_{g_i^0} \delta_{g_i^0 t} + (1 - \kappa_{g_i^0}) \mu_t.$$

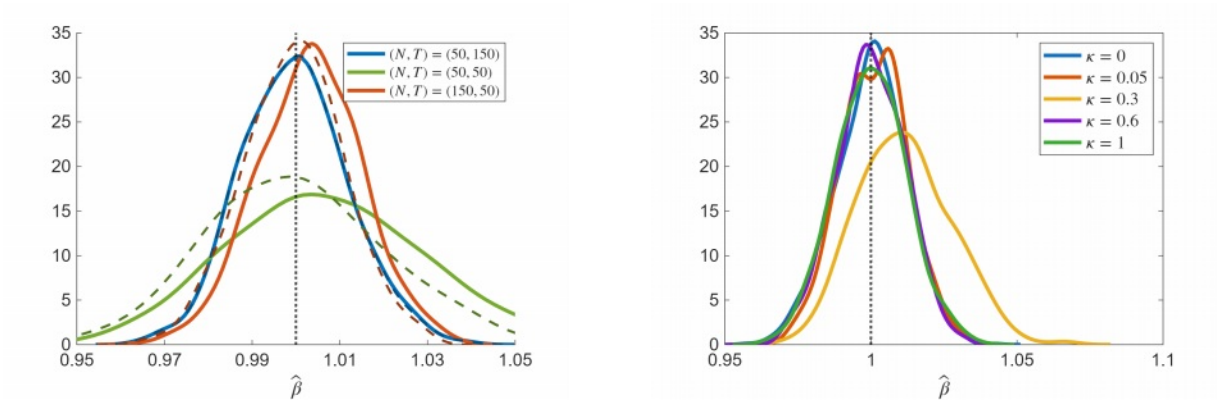
The parameters $\kappa_g \in [0, 1]$ control the strength of separation across groups. The strong separation case happens when $\kappa_g = 1$ for all g, since δ_{gt} are normalized to be strongly separated. Analogously, the extreme case of no separation happens when $\kappa_g = 0$ for all g. I consider two groups, $G^0 = 2$, hence only one value of κ is sufficient to control the degree of separation, since for $\kappa := \kappa_1 = \kappa_2$, group separation is given by

$$\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{1t} - \alpha^0_{2t}\right)^2 = \kappa^2 \left(\frac{1}{T}\sum_{t=1}^{T}(\delta_{1t} - \delta_{2t})^2\right).$$

The slope parameter is fixed at $\beta^0 = 1$, and the variables v_{it}, u_{it}, δ_{gt} and μ_t are all drawn from independent standard Normal distributions. True group memberships are drawn uniformly at random on $[G^0]$ independently for each $i \in [N]$.

To verify the implications of Theorem 2, I conduct two analyses, reported in Figure II. First, to explore the conditions on the relative dimensions of the panel, I estimate the model using the GFE estimator with

three different sample sizes. The sample dimensions considered are $(N,T) = (50,150)$, $(N,T) = (50,50)$ and $(N,T) = (150,50)$. I set $\kappa = 0.45$ to represent weak group separation. I use 1,000 Monte Carlo replications, and the results are presented in Figure II(a). The blue lines display the case where the time dimension is significantly larger than the cross-sectional dimension, the green lines the intermediate case, while the orange lines indicate the opposite. For each design, solid lines trace out the distribution of the GFE estimator, and dashed lines are used for the distribution of the oracle. The evidence from the plot is clear: under weak separation, the GFE estimator behaves like the oracle only when the time dimension is significantly larger than that of the cross-section. The intermediate case separates the roles of the two dimensions: at fixed N, a longer panel removes the distortion, whereas at fixed T a larger cross-section reduces the bias in absolute terms—the group effects are estimated more precisely and fewer units are misclassified—but only mildly relative to the oracle's sampling error, which shrinks faster. A larger cross-section thus does not substitute for a longer panel, in line with the rate condition of Theorem 2, in which N may only grow polynomially in T. Here, the biases induced from a larger cross-section are only moderate, but this may not be the case in general. Appendix C.5 repeats the three designs with heavy-tailed errors and finds that, with fewer than three moments, the GFE estimator departs from the oracle at every panel size considered, in line with the role of the moment condition in the rate condition of Theorem 2.



(a) Dimension of the panel

(b) Degree of separation

Figure II: Distribution of the GFE estimator under weak group separation

The second experiment consists on a sensitivity analysis of the GFE estimator's distribution as a function of the strength of group separation. In Figure II(b), we see the distribution of the GFE estimator under different degrees of separation when $(N,T) = (50,150)$. The results make it evident that the GFE estimator is correctly centered under strong separation ($\kappa = 1$), weak separation ($\kappa = 0.45, 0.6$; see panel (a) for the former) and extremely weak separation ($\kappa = 0, 0.05$). In contrast, there exist large biases when groups are very weakly separated ($\kappa = 0.3$). This illustrates the potential non-standard behavior of the GFE estimator in some regions of the weak separation regime argued in footnote 3.

6 Consequences for Applied Practice

While the novel robustness results yield sharp and clear theoretical consequences, the scope of the implications for applied practice are naturally sample-dependent. However, the updated knowledge about the behavior of the GFE estimator allows us to revisit older results, and interpret the conclusions therein more thoroughly. By looking at the empirical application in BM in hindsight, this section aims to put the benefits

of the novel results into perspective and guide applied economists on how to fully exploit them.

Consider the model for income and democracy in Acemoglu, Johnson, Robinson and Yared (2008), adapted to account for grouped fixed-effects:

$$democracy_{it} = \beta_1 democracy_{i,t-1} + \beta_2 income_{i,t-1} + \alpha_{g_i t} + v_{it}. \tag{4}$$

The model includes two regressors: lagged democracy, measured by the Freedom House indicator, and lagged income, measured by logged GDP per capita. The grouped fixed-effects capture factors that simultaneously affect economic and political development which may be shared by groups of countries. I follow the literature and estimate the model on the 1970–2000 balanced subsample of Acemoglu et al. (2008). This yields a sample of size $(N, T) = (90, 7)$. Two caveats apply. The panel is very short, and Section 5 shows that, with T of this order, misclassification is the first-order concern. Moreover, the model is dynamic, so that the lagged dependent variable is not strictly exogenous. The verification of Assumption 5 in Lemma B.4 does not cover this case, and the exercise below should be read as an illustration of how the results can inform applied practice rather than as an application in which their conditions are known to hold.

6.1 Selecting the number of groups

The first question we must address is how many groups G we shall use to estimate model (4). Suppose we have a selection rule $\widehat{G}$ satisfying Corollary 1. While theoretically valid in large samples, in practice the finite-sample properties of the estimators $(\hat{\beta}_{1,\widehat{G}}, \hat{\beta}_{2,\widehat{G}})$ will crucially depend on the finite-sample properties of $\widehat{G}$. Moreover, being extremely conservative will have adverse consequences on the estimates of group-specific parameters—their distribution does depend on the group specifications.

One possible solution to effectively answer the question is to estimate the model over a fine grid of G values. Per Theorem 1(b), parameter estimates of (β_1, β_2), as well as their standard errors, should stabilize after $G \geq G^0$. Such a threshold would then hint at the true number of groups. What this rule identifies depends on Assumption 5. Under it, the extra groups fit the idiosyncratic errors and estimates stabilize at G^0. If instead the classification picks up a latent structure of the covariates (Remark 3), estimates stabilize once G exceeds the number of cells of the joint structure of (y_{it}, x_{it}), at an estimator which is oracle-equivalent for that joint structure but has a larger variance than the oracle for the outcome equation. In both cases, the threshold is the smallest G beyond which further overspecification is harmless for inference on the common parameters. Whether it coincides with G^0 cannot be settled from the stabilization pattern alone, but the within-group covariate variation $\widehat{W}(\hat{\gamma}_G)$ helps: under Assumption 5 it stabilizes once $G \geq G^0$, whereas a continued sharp decline signals that the additional groups absorb covariate structure; see Appendix C.2. Here, the within-group variation of lagged income drops from 1.12 at $G = 1$ to 0.61 at $G = 4$ and is flat afterwards (0.54–0.61 for $G \geq 4$; Table C.4), so that the movements of the estimates between $G = 4$ and $G = 8$ do not appear to be driven by covariate structure absorbed by the additional groups.

Figure III, which replicates Figure 1 in BM, shows the parameter estimates and standard errors on lagged democracy and income for different values of the number of groups G. I have highlighted in red the value of G after which estimation results appear to stabilize.[5] Thus, given the novel findings derived in this paper, we have some support for the hypothesis that the true number of groups is no less than 8. On the

[5] The red line is merely an eyeballed indicator rather than a formal statistical test.

other hand, the BIC provides an upper bound on the true number of groups if T grows at a slower rate than N. The choice $G = 10$ is optimal according to BIC, which suggests that the true number of groups should be no larger than 10. All in all, we now have empirical evidence that the true number of groups G^0 may be either 8, 9 or 10. The new findings on G^0 contrast with the choice in BM, who set $G = 4$, and a

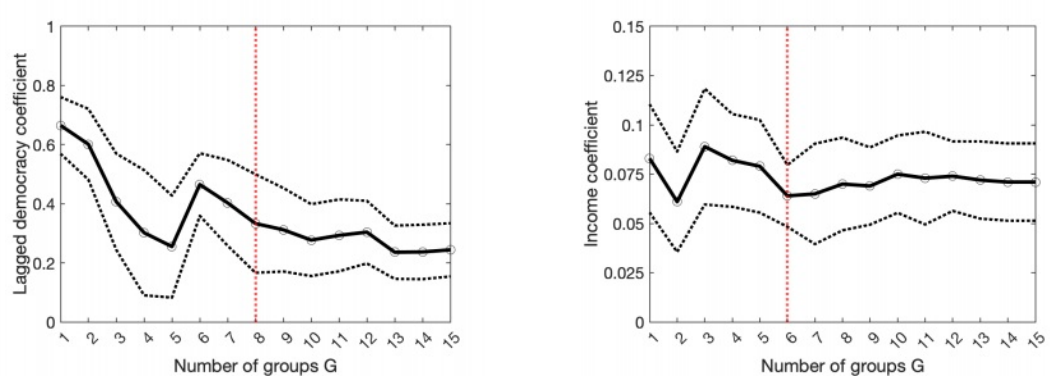


Figure III: Sensitivity analysis in G

Notes: The x-axis shows the number of groups G used in estimation. The y-axis reports parameter values. 95%-confidence intervals clustered at the country level are shown in dashed lines. Confidence intervals are based on the fixed-T formulas in BM. The red line shows the value of G for which results stabilize.

long literature who follow their strategy (e.g., Aguilar Loyo and Boot, 2025; Lewis, Melcangi, Pilossoph and Toner-Rodgers, 2023).[6] Naturally, the difference in the estimation results directly translates to the economic predictions derived from the model. For instance, the estimated cumulative income effect $\hat{\beta}_{2,G}/(1-\hat{\beta}_{1,G})$ decreases from 0.15 for $G = 3$, to 0.1 for $G = 9$.

In order to better understand some of the forces underlying Figure III, it is interesting to explore the estimated group composition across G values. According to Theorem 1(a), groups should increasingly split, but not mix, for large enough G. Figure IV displays the group composition for $G = 2$ against $G = 3$, as well as for $G = 8$ against $G = 9$.

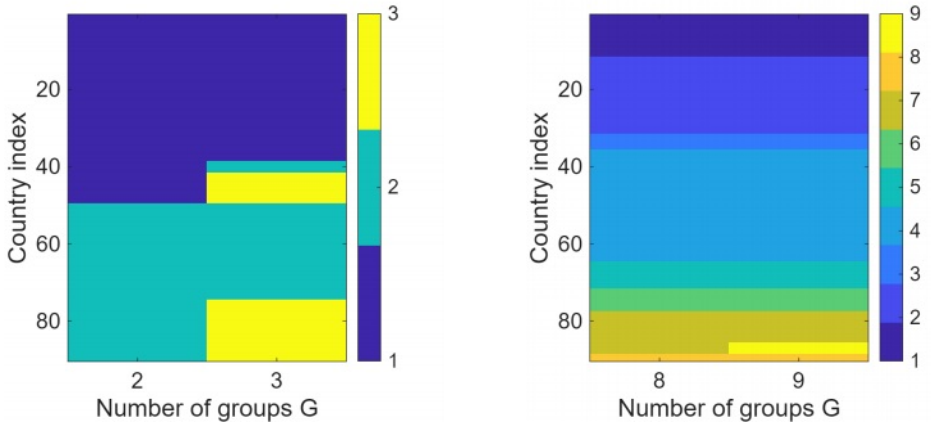


Figure IV: Changes in group composition

Notes: The x-axis shows the number of groups G used in estimation. The y-axis are country indicators. Each color represents one group; labels are arbitrary yet aligned across each comparison.

[6]To be fair, the choice of $G = 4$ in BM seems mainly motivated for a clean display of the results. In Section S.7.2. they extend the main analysis and consider other choices of G.

Figure IV shows that, when moving from two to three groups, both groups in the original estimated structure (cian and blue) contribute units to the third group (yellow), whereas, when moving from eight to nine groups, the new group (yellow) is carved out of a single group (olive) and no other unit changes group. If Theorem 1(a) applied exactly with $G \geq G^0$, no pair of units separated with G groups would be joined with $G+1$, and no unit would move beyond a pure split. For $2 \to 3$, 15% of the pairs grouped together with three groups were separated with two, 12% of the units move beyond a pure split, and the adjusted Rand index between the two partitions is 0.48. For $8 \to 9$, no pair is joined and no unit moves, with adjusted Rand index of 0.98, hence is a pure split. The numerical measures of similarity in Table C.4 in Appendix C confirm that the intermediate transitions lie in between while transitions beyond $G = 8$ remain pure splits, which is consistent with the evidence from Figure III that the classification settles at about eight groups. Two caveats apply. With $T = 7$, partitions that reshuffle a few units attain values of the GFE objective very close to that of the reported one, so the group composition is only as reliable as the numerical minimization. Moreover, the misclassification rates BM report in simulations with $T = 7$ are of the order of one unit in ten, so that the evidence from group composition is suggestive rather than conclusive.

6.2 Assessing the consequences of group separation

Understanding the behavior of the GFE estimator under different possible strengths of group separation in (4) is equally important. To assess the presence of separation issues, one may be tempted to employ a naive statistic that directly targets the strength of separation, such as

$$\min_{g,\tilde{g}} \frac{1}{T} \sum_{t=1}^{T} (\hat{\alpha}_{g,t} - \hat{\alpha}_{\tilde{g},t})^2.$$

However, this is not a valid statistic in general since it endogenously depends on the strength of separation—when groups are too weakly separated such that Theorem 2 does not hold, model parameters may not be accurately estimated. Moreover, weak separation is only a problem to the extent that parameter estimates are biased—any indicator of weak separation should then focus on the consequences and not on the cause.[7]

To illustrate the effects of group separation in model (4), I employ data perturbation to enforce weak separation. In particular, I take the estimated grouped fixed-effects, and then gradually shrink the differences between them toward their grand mean. This creates a sequence of datasets with continuously varying separation, from original separation down to zero separation. I then re-compute the GFE estimator for each shrinkage level and track its behavior. This procedure directly manipulates separation while leaving covariates and noise structure intact.

The approach works as follows:

1. Fit the GFE estimator at (a proxy for) G^0, and record firm group assignments and grouped fixed-effects $\hat{\alpha}_{gt}$. Given the previous discussion, I assume $G^0 = 8$.

2. For a grid $\lambda \in [0, 1]$, create perturbed grouped fixed-effects $\alpha_{gt}^{(\lambda)} = (1-\lambda)\hat{\alpha}_{gt} + \lambda\bar{\alpha}_t$, where the

[7] This contrasts with other similar regimes in the literature. E.g., the consequences of weak instruments vary monotonically with the strength of the instruments (Staiger and Stock, 1997), which is why the strength of the instruments can be assessed from the first stage alone, without reference to the parameter of interest. Here, instead, the consequences are not monotone in the strength of separation: for extremely weak separation we recover $\sqrt{NT}$-consistency as under strong separation, whereas for very weak separation the results are not guaranteed; see footnote 3 for a formal definition of those regimes. Figuring out what the exact thresholds of weak, very weak and extremely weak are in finite samples seems an arduous task.

time-specific mean is $\bar{\alpha}_t = \frac{1}{G^0}\sum_{g=1}^{G^0}\hat{\alpha}_{gt}$. Higher λ represents weaker separation.

3. Generate a new dependent variable $y_{it}^{(\lambda)} = x_{it}'\hat{\beta} + \alpha_{g_i t}^{(\lambda)} + \hat{v}_{it}$, where $\hat{v}_{it}$ are residuals from the original fit.[8]

4. Re-run the GFE estimator on each λ-dataset.

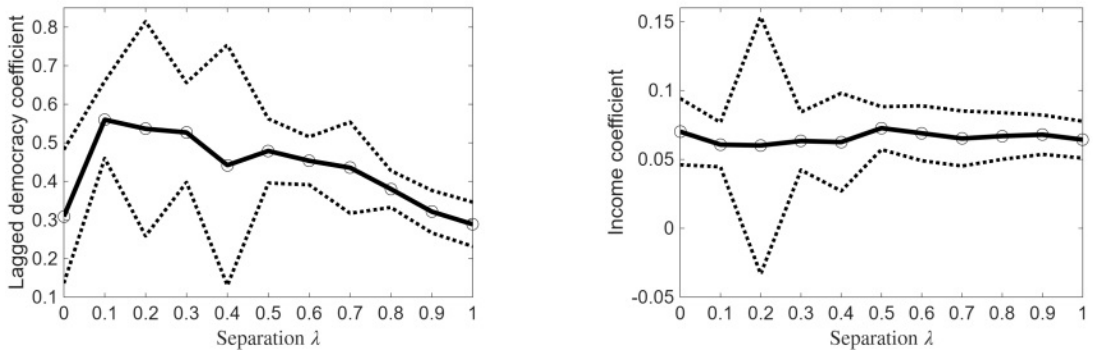


Figure V: Sensitivity analysis in λ

Notes: The x-axis shows the perturbation parameter λ. The y-axis reports estimated parameter values. 95%-confidence intervals clustered at the country level are shown in dashed lines. Confidence intervals are based on the fixed-T formulas in BM. The figure reports, for each λ, the best of 100 starting values together with a start at the baseline classification.

Figure V plots the GFE estimator $(\hat{\beta}_{1,G^0}, \hat{\beta}_{2,G^0})$ against the separation parameter λ. Since the oracle estimator equals the baseline estimate for every λ (footnote 8), any movement in the figure is due to classification error. The income coefficient is essentially flat, with increased standard errors for $\lambda = 0.2$. The lagged democracy coefficient moves away from the baseline as soon as separation is reduced and remains well above it up to $\lambda = 0.7$, with confidence bands that widen irregularly for intermediate values of λ, and it then returns towards the baseline as λ approaches 1, where the bands are narrowest. This is what the results of Section 4 lead one to expect in a panel with $T = 7$. The classification is fragile already at the original separation (Section 5 and Figure IV), so that reducing separation quickly moves the estimator into the very weak regime, where the results are not guaranteed, whereas for λ close to 1 the groups are so close that the estimator behaves as the pooled oracle—which here coincides with the baseline estimate—in line with Proposition 1. The magnitude of the distortions in between, up to 0.25 or about 1.5 baseline standard errors, should be read with the caveat that, with $T = 7$ and $G = 8$, the minimization of the GFE objective is difficult and the computed path depends on the search (Appendix C.2).

The extent of the distortions also hinges on the original separation in the model. Whether groups in model (4) are (very) weakly separated, or unobserved heterogeneity has a more complex structure, are interesting questions that deserve further attention, yet go beyond the purpose of this section.

[8] In a dynamic panel as the one considered here, perturbing the dependent variable may also perturb the regressors. However, the perturbation considered here sidesteps this complication since it implies $y_{it}^{(\lambda)} = y_{it} + \lambda(\bar{\alpha}_t - \hat{\alpha}_{\hat{g}_i t})$. Moreover, for every λ, the oracle estimator that uses the baseline classification returns exactly $\hat{\beta}$: the residuals $\hat{v}_{it}$ have zero group–time means and are orthogonal to the within-group deviations of x_{it}, so that the perturbed data satisfy the model's first-order conditions at $(\hat{\beta}, \alpha^{(\lambda)})$. The exercise thus isolates the effect of classification error.

7 Conclusion and Message for Practitioners

Models that incorporate time-varying, group-specific fixed-effects have become central in applied research for capturing rich forms of unobserved heterogeneity. This paper demonstrates that the widely used grouped fixed-effects estimator possesses two key robustness properties that enhance its applicability across diverse empirical settings. These results yield several practical takeaways for researchers doing estimation and inference on common regression coefficients:

i. Overselecting the number of groups may escape notice yet is costless, provided the additional groups fit idiosyncratic noise rather than latent structure in the covariates; underselecting is extremely costly yet easily detectable.

ii. There is no need to consistently estimate the number of groups; selecting any large enough number of groups yields standard asymptotically normal inference and circumvents model-selection concerns.

iii. Weak group separation is not a problem provided the cross-section does not grow too fast relative to the time dimension; how fast depends on the degree of separation and on the tails of the errors.

iv. If group separation is extremely weak, accurate classification is impossible but the consequences are negligible unless the covariates carry the structure of the groups that the estimator pools.

v. When the covariates may carry latent group structure of their own, computation matters: initialize the classification at the residuals of a preliminary estimate rather than at random assignments, and monitor the within-group covariate variation across G; a sharp decline signals that the additional groups absorb covariate structure, in which case standard errors may be inflated.

This paper formally bridges the gap between theory and practice by showing that the GFE estimator mimics the behavior of an oracle estimator even under realistic deviations from ideal assumptions. Thus, the findings not only clarify the theoretical properties of the GFE estimator but also simplify modeling decisions for practitioners. This lends further credibility to the increasing use of latent group methods in labor, education, industrial organization, and development economics, among many others, and renders the GFE approach as a more robust and accessible method for applied work than previously thought.

References

Acemoglu, D., Johnson, S., Robinson, J. A. and Yared, P. (2008), 'Income and democracy', *American Economic Review* **98**(3), 808–842.

Aguilar Loyo, J. and Boot, T. (2025), 'Grouped heterogeneity in linear panel data models with heterogeneous error variances', *Journal of Business & Economic Statistics* **43**(1), 68–80.

Ahn, S. C. and Horenstein, A. R. (2013), 'Eigenvalue ratio test for the number of factors', *Econometrica* **81**(3), 1203–1227.

Bai, J. (2009), 'Panel data models with interactive fixed effects', *Econometrica* **77**(4), 1229–1279.

Bai, J. and Ng, S. (2002), 'Determining the number of factors in approximate factor models', *Econometrica* **70**(1), 191–221.

Bonhomme, S. and Manresa, E. (2015), 'Grouped patterns of heterogeneity in panel data', *Econometrica* **83**(3), 1147–1184.

Bradley, R. C. (2005), 'Basic properties of strong mixing conditions. A survey and some open questions', *Probability Surveys* **2**, 107–144.

Dzemski, A. and Okui, R. (2024), 'Confidence set for group membership', *Quantitative Economics* **15**(2), 245–277.

Leeb, H. and Pötscher, B. M. (2005), 'Model selection and inference: Facts and fiction', *Econometric Theory* **21**(1), 21–59.

Lewis, D. J., Melcangi, D., Pilossoph, L. and Toner-Rodgers, A. (2023), 'Approximating grouped fixed effects estimation via fuzzy clustering regression', *Journal of Applied Econometrics* **38**(7), 1077–1084.

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

Lumsdaine, R. L., Okui, R. and Wang, W. (2023), 'Estimation of panel group structure models with structural breaks in group memberships and coefficients', *Journal of Econometrics* **233**(1), 45–65.

Merlevède, F., Peligrad, M. and Rio, E. (2011), 'A Bernstein type inequality and moderate deviations for weakly dependent sequences', *Probability Theory and Related Fields* **151**(3), 435–474.

Moon, H. R. and Weidner, M. (2015), 'Linear regression for panel with unknown number of factors as interactive fixed effects', *Econometrica* **83**(4), 1543–1579.

Onatski, A. (2010), 'Determining the number of factors from empirical distribution of eigenvalues', *The Review of Economics and Statistics* **92**(4), 1004–1016.

Onatski, A. (2012), 'Asymptotics of the principal components estimator of large factor models with weakly influential factors', *Journal of Econometrics* **168**(2), 244–258.

Phillips, P. C. and Sul, D. (2009), 'Economic transition and growth', *Journal of Applied Econometrics* **24**(7), 1153–1185.

Staiger, D. and Stock, J. H. (1997), 'Instrumental variables regression with weak instruments', *Econometrica* **65**(3), 557–586.

Yokoyama, R. (1980), 'Moment bounds for stationary mixing sequences', *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **52**(1), 45–57.

Appendix

The Appendix collects in Section A the proofs of Theorems 1 and 2. Proofs of auxiliary results used in Section A are gathered in Section B. Throughout, C is an arbitrary positive constant that can change from line to line. Section C contains additional simulation results.

A Proofs of Main Results

A.1 Proof of Theorem 1

The proof of Theorem 1 relies on two conditions that strengthen those of BM, Assumptions 4 and 5 below, and on three auxiliary lemmas, Lemmas A.1–A.3, which are proved in Appendix B. Part (a) rests on Lemmas A.1–A.2 and Assumption 4; part (b) on Lemma A.3 and Assumption 5. I first state the assumptions and the lemmas, together with the notation they require and a discussion of their content, and then give the proof of the theorem, which is short conditional on all these previous results. Throughout the Appendix, for $a, b \in \mathbb{R}^T$ I write $\|a\|_T^2 := T^{-1}\sum_{t=1}^T a_t^2$ and $\langle a, b\rangle_T := T^{-1}\sum_{t=1}^T a_t b_t$.

The first lemma is the half of Lemma B.3 in BM that survives overspecification: every true group effect is approximated by some estimated group effect.

Lemma A.1. *Under Assumptions 1 and 2(a), and $G \geq G^0$, $\max_{g\in[G^0]} \min_{\tilde{g}\in[G]:\hat{\mathcal{C}}_{\tilde{g}}\neq\emptyset} \|\hat{\alpha}_{\tilde{g}} - \alpha_g^0\|_T^2 = o_p(1)$ as $N, T \to \infty$, where the minimum runs over the nonempty estimated groups.*

Lemma A.1 is one-sided on purpose. The converse statement—that every estimated group effect is close to some true effect, so that $\hat{\alpha}$ would be consistent for α^0 in the Hausdorff distance—is false when $G > G^0$: the GFE estimator may allocate a spare group to a single unit, or to a few units, whose effect then fits the idiosyncratic errors of those units and stays at a positive distance from every α_g^0. Such small groups are harmless for part (a) of Theorem 1, which is a statement about the composition of the estimated groups, and part (a) is proved directly, without Hausdorff consistency, by exploiting the optimality of the estimated assignment unit by unit and by comparing the estimated partition with alternative partitions. This requires the following strengthening of the conditions of BM.

Assumption 4.

(a) The processes $\{v_{it}\}_t$, $i \in [N]$, are independent across i;

(b) There exists a constant $M_2 > 0$ such that $\sup_{i\in[N]} \mathbb{P}\left(T^{-1}\sum_{t=1}^T \|x_{it}\|^2 > M_2\right) = o(T^{-\delta})$ for all $\delta > 0$.

Part (a) restricts the cross-sectional dependence of the errors, which Assumptions 1(e)–(f) only restrict on average; BM note that their conditions hold in this case. It is used through a single consequence, $\max_{i\neq j} |\langle v_i, v_j\rangle_T| = o_p(1)$ (Lemma B.1(b)): the error paths of any two units are asymptotically orthogonal. Some restriction of this kind is necessary for part (a) of Theorem 1, and not only for its proof. Suppose two units i and j in different true groups had identical error paths, $v_{it} = v_{jt}$ for all t, which Assumptions 1–2 allow. With $G = G^0 + 1$, assigning i and j to a spare group of their own reduces the sum of squared residuals by approximately $T[2\|v_i\|_T^2 - \frac{1}{2}\|\alpha_{g_i^0}^0 - \alpha_{g_j^0}^0\|_T^2]$, whereas any use of the spare group that respects

the true groups reduces it by at most $T \max_{k\in[N]} \|v_k\|_T^2$ up to terms of smaller order when $T/N \to \infty$. Hence, heuristically, with homoskedastic errors whose variance exceeds half the separation between the two groups, the GFE estimator forms a mixed group with probability approaching one. Part (b) of Assumption 4 is the second-moment analogue of Assumption 2(e) and holds under the conditions discussed by BM for the latter (bounded covariates, or tail and dependence conditions on x_{it} similar to those on v_{it}). These two requirements have the same form as the conditions under which Liu et al. (2020) prove the analogue of part (a) of the theorem for latent-class models with group-specific slopes. Their Assumption A1(a) imposes cross-sectional independence of the observations conditional on the group structure, and their Assumptions A1(b) and A1(e) impose exponential mixing and exponential tails, which play the role of Assumption 2(c)–(d) and Assumption 4(b) here. In this sense, Assumption 4 brings the conditions for classification with an overspecified number of groups in line with theirs, rather than adding a genuinely new restriction.

Lemma A.2. *Under Assumptions 1, 2 and 4, when $G \geq G^0$ and $N, T \to \infty$ such that $NT^{-\nu} \to 0$ for some $\nu > 0$,*

$$\mathbb{P}\left(\forall g \in [G],\ \exists \tilde{g} \in [G^0] :\ \hat{\mathcal{C}}_g \subseteq \mathcal{C}^0_{\tilde{g}}\right) = 1 - o(1).$$

One can immediately see that part (a) of Theorem 1 is Lemma A.2.

Let $\hat{\gamma}^{(G)}$ denote the partition estimated with G groups, so that $\hat{\beta}_G = \hat{\beta}(\hat{\gamma}^{(G)})$ with $\hat{\beta}(\cdot)$ the least-squares estimator conditional on a partition (see step (a) of the proof of Lemma A.3), and define the event

$$E = \left\{\text{for each } g \in [G],\, G \geq G^0, \text{ there exists } \tilde{g} \in [G^0] \text{ such that } \hat{\mathcal{C}}_g \subseteq \mathcal{C}^0_{\tilde{g}}\right\}, \tag{A.1}$$

whose probability tends to one by part (a) of Theorem 1. Part (b) of Theorem 1 follows from Lemma A.3 below, which requires a condition on the joint behavior of the covariates, the errors and the true groups, and which I now introduce. The same notation is used throughout the proof of Lemma A.3 in Appendix B. A partition $\gamma \in \Gamma_G$ is said to refine γ^0 if every cell of γ is contained in a single true group; on the event E, $\hat{\gamma}^{(G)}$ refines γ^0. For a refinement γ, denote its cells by S, write $\bar{x}_{St}$, $\bar{v}_{St}$ for the cell means at time t, and $\bar{x}_{g^0t}$, $\bar{v}_{g^0t}$ for the means over the true group $\mathcal{C}^0_g \supseteq S$. Define the between-cell (within-true-group) second moments

$$\Omega(\gamma) := \frac{1}{NT}\sum_{t=1}^{T}\sum_{S} |S| \left(\bar{x}_{St} - \bar{x}_{g^0t}\right)\left(\bar{x}_{St} - \bar{x}_{g^0t}\right)', \qquad D(\gamma) := \frac{1}{NT}\sum_{t=1}^{T}\sum_{S} |S| \left(\bar{v}_{St} - \bar{v}_{g^0t}\right)^2,$$
$$\delta(\gamma) := \frac{1}{NT}\sum_{t=1}^{T}\sum_{S} |S| \left(\bar{x}_{St} - \bar{x}_{g^0t}\right)\left(\bar{v}_{St} - \bar{v}_{g^0t}\right),$$

where $\sum_S$ runs over all cells of γ. By the ANOVA-type identities of Appendix B, $\Omega(\gamma) = \widehat{W}(\gamma^0) - \widehat{W}(\gamma) \succeq 0$ is the share of the (within-true-group) covariate variation absorbed by the refinement, $\delta(\gamma) = \Delta(\gamma^0) - \Delta(\gamma)$ is the covariate–error covariance the refinement creates, and $D(\gamma)$ is the improvement in the fit of the grouped effects to the idiosyncratic errors achieved by the refinement. The objects $\widehat{W}(\cdot)$ and $\Delta(\cdot)$ denote the Gram matrix and the score of the residualized regression, see (B.2) and (B.5). Let $\tilde{\gamma}$ be a (measurable selection of a) maximizer of $D(\gamma)$ over the refinements of γ^0 in Γ_G. As shown in step (e) of the proof of Lemma A.3, $\tilde{\gamma}$ is the classification that the GFE estimator would select if β^0 were known, so that $\tilde{\gamma}$ depends on the errors and the true groups only.

For a refinement γ, let $L(\gamma) := NT[D(\tilde{\gamma}) - D(\gamma)] \geq 0$ be its loss in the fit to the errors relative to $\tilde{\gamma}$, and for $c, K > 0$ define the class of refinements[9]

$$\mathcal{N}_{c,K} := \left\{ \gamma \in \Gamma_G \text{ refining } \gamma^0 : \; L(\gamma) \leq K + c\, NT\, \delta(\gamma)' \widehat{W}(\gamma)^{-1} \delta(\gamma) \right\},$$

which contains $\tilde{\gamma}$ and all refinements whose loss in the fit to the errors is at most a constant plus c times the gain they obtain through the covariate channel, $NT\delta(\gamma)'\widehat{W}(\gamma)^{-1}\delta(\gamma)$, which is the reduction in the sum of squared residuals that the spurious covariate–error covariance $\delta(\gamma)$ generates when the slope is re-estimated at γ; see (B.11) below.

Assumption 5. *There exists a constant $c > 1$ such that, for every $K > 0$,*

$$\sup_{\gamma \in \mathcal{N}_{c,K}} \left\{ \|\Omega(\gamma)\| + \sqrt{NT}\, \|\delta(\gamma)\| \right\} = o_p(1).$$

Assumption 5 involves no estimator of β^0 or of the partition: $\mathcal{N}_{c,K}$ is a random class of refinements determined by the covariates, the errors and the true groups, and the supremum is over that class. Its content is best understood from step (e) of the proof of Lemma A.3, which shows that the GFE estimator can only select a refinement whose loss in the fit to the errors is compensated by the reduction in the sum of squared residuals obtained through the covariate channel, and this reduction exceeds the covariate-channel gain $NT\delta(\gamma)'\widehat{W}(\gamma)^{-1}\delta(\gamma)$ only by a term driven by the oracle score, which is $O_p(1)$. Hence, for every $c > 1$ and with probability arbitrarily close to one, $\hat{\gamma}^{(G)}$ belongs to $\mathcal{N}_{c,K}$ for some large K, and the assumption states that all refinements in this class are asymptotically uninformative about the covariates ($\Omega(\gamma) = o_p(1)$) and generate no covariate–error covariance of the order of the sampling error ($\sqrt{NT}\delta(\gamma) = o_p(1)$).

Three types of refinements are worth distinguishing. First, $\tilde{\gamma}$ itself, for which $L(\tilde{\gamma}) = 0$. Lemma B.4 in Appendix B shows that $\Omega(\tilde{\gamma}) = O_p(N^{-1})$ and $\sqrt{NT}\delta(\tilde{\gamma}) = O_p(N^{-1/4} + T^{-1/4})$ under primitive conditions—strict exogeneity of the covariates and weak dependence of their within-group deviations—through the elementary variance-shrinking mechanism described in the Introduction: the cells of $\tilde{\gamma}$ are averages of errors that are independent of the covariates. Second, refinements with $L(\gamma) = O(1)$, which fit the errors almost as well as $\tilde{\gamma}$. Heuristically, they differ from $\tilde{\gamma}$ in the assignment of a vanishing fraction of units located near the boundaries of its cells, and reassigning $o(N)$ units changes $\Omega(\gamma)$ and $\sqrt{NT}\delta(\gamma)$ by $o_p(1)$ only. Third, refinements that align with a latent structure of the covariates and with the errors, which can generate $\sqrt{NT}\|\delta(\gamma)\|$ of order $\sqrt{N}$. Assumption 5 requires that their loss $L(\gamma)$ exceed $c\, NT\delta(\gamma)'\widehat{W}(\gamma)^{-1}\delta(\gamma)$, i.e., that the fit to the errors they forgo dominate, by a factor larger than one, the gain they obtain through the covariate channel. Whether this holds depends on the design. Remark 1 discusses an example with a latent binary structure in the covariates, in which the loss exceeds the covariate-channel gain by a factor bounded away from one for all values of the variances.

Assumption 5 is not implied by Assumptions 1–2 alone, and its verification beyond $\gamma = \tilde{\gamma}$ is left for future research. I conjecture that it can be derived from two ingredients. The first is a strengthening of condition (b) of Lemma B.4 from second moments to sub-Gaussian tails. Conditional on the errors and the true groups, this would make the covariate–error covariance $\sqrt{NT}\delta(\gamma)$ of any refinement γ that does not depend on the covariates concentrate at the scale $D(\tilde{\gamma})^{1/2}$, i.e., $(N \wedge T)^{-1/4}$, as in the proof of that lemma. The second

[9] $\widehat{W}(\gamma)$ is invertible for every $\gamma \in \Gamma_G$ with probability approaching one (Lemma B.5). On the complement of this event, $\mathcal{N}_{c,K}$ may be defined arbitrarily, which is immaterial for a statement in probability.

is a bound on the number of refinements whose loss $L(\gamma)$ in the fit to the errors is at most ℓ, of the form $\exp(a_{NT} + b_{NT}\ell)$ with a_{NT} and b_{NT} of smaller order than $(N \wedge T)^{1/2}$. It is then possible to show, by a union bound over the refinements with a given loss and the observation that a refinement in $\mathcal{N}_{c,K}$ with a large covariate–error covariance must have a loss proportional to the square of that covariance, that no refinement in $\mathcal{N}_{c,K}$ is asymptotically informative about the covariates. Assumption 5 would thus reduce to a counting condition on the near-optimal classifications of weakly dependent noise, a problem that involves neither the covariates nor the slope and that, to my knowledge, has not been studied. A heuristic count indicates when such a condition can be expected to hold: reassigning a unit from its cell in $\tilde{\gamma}$ costs an amount of order $\sqrt{T/N}$ in the fit to the errors (Remark 2), so that about $\ell N^{3/2}T^{-1/2}$ units can be reassigned at a loss of at most ℓ, and if the costs of such reassignments were additive the counting condition would hold if and only if $N^2/T = o(1)$. Such a regime is between the stability regime $T/N^3 \to \infty$ of Remark 2 and the regime $N \asymp T$, in which the simulations of Appendix C.2 find covariate-aligned solutions in designs with a strong latent structure in the covariates.

The comparison with the interactive fixed-effects case is instructive. The robustness result of Moon and Weidner (2015) rests on a high-level condition (their Assumption EV) requiring the eigenvalues of the error covariance matrices to be sufficiently separated and their eigenvectors—which determine the additional factors—to be sufficiently uncorrelated with the regressors. They verify it for i.i.d. normal errors using random matrix theory and conjecture that it holds more generally. Assumption 5 plays the same role here: the near-optimal classifications of the errors must be uninformative about the covariates. It is verified for the error-driven classification $\tilde{\gamma}$ (Lemma B.4) under weaker distributional assumptions than theirs, but the classification is discrete and selected jointly with the slope, so that, unlike the estimated factors, the additional groups may align with the covariates through the covariate channel, which is what Assumption 5 excludes. Remark 2 discusses a simpler but stronger sufficient condition, and Remark 3 what remains true when the supremum in Assumption 5 is bounded but does not vanish. Appendix C provides numerical evidence, showing that, in the example of Remark 1, the estimated partitions ignore the covariate structure at every sample size considered, whereas in a design with two groups in the outcome equation and a strong covariate structure, local minima of the GFE objective aligned with the covariates coexist with error-driven ones at nearly the same objective value, and the behavior of the computed estimator depends on the initialization of the algorithm.

Lemma A.3. *Under Assumptions 1, 2 and 5, if* $\sqrt{NT}(\hat{\beta}_{G^0} - \beta^0) = O_p(1)$, *when* $G \geq G^0$ *and* $N, T \to \infty$ *such that* $NT^{-\nu} \to 0$ *for some* $\nu > 0$, *for every* $C > 0$,

$$\mathbb{P}\left(\left\|\hat{\beta}_G - \hat{\beta}_{G^0}\right\| > C/\sqrt{NT}, E\right) = o(1).$$

This rather long presentation of the additional assumptions and notation already sketched the key steps of the proof of Theorem 1, which is formalized as follows.

Proof of Theorem 1. Part (a) is Lemma A.2. For part (b), note that

$$\mathbb{P}\left(\left\|\hat{\beta}_G - \hat{\beta}_{G^0}\right\| > C/\sqrt{NT}\right) \leq \mathbb{P}\left(\left\|\hat{\beta}_G - \hat{\beta}_{G^0}\right\| > C/\sqrt{NT}, E\right) + \mathbb{P}\left(E^c\right). \tag{A.2}$$

The first term on the RHS in (A.2) is $o(1)$ by Lemma A.3 for every $C > 0$, while the second term is $o(1)$ by part (a) of the theorem. This concludes the proof. □

A.2 Proof of Theorem 2

To prove Theorem 2, as in the previous subsection, I start by outlining a sequence of auxiliary results that will make the proof of the actual result short. The following sequence of lemmas borrow some arguments from Appendices A and B in BM, Lumsdaine et al. (2023, Supplemental Material C) and Dzemski and Okui (2024, Supplemental Material D).

Lemma A.4. *Suppose Assumptions 1 and 3(a), (c) hold, and $NT^{-r/2} \to 0$. Then,*

$$\max_{g\in[G]} \min_{\tilde{g}\in[G]} \frac{1}{T}\sum_{t=1}^{T} \left(\hat{\alpha}_{\tilde{g}t} - \alpha^0_{gt}\right)^2 = O_p\left((N\wedge T)^{-1/2}\right) \quad \text{and} \quad \|\hat{\beta} - \beta^0\|^2 = O_p\left((N\wedge T)^{-1/2}\right).$$

Lemma A.5. *Suppose Assumptions 1 and 3(a)-(c) hold, and $NT^{-r/2} \to 0$. Then there exists a permutation $\sigma : [G] \to [G]$ such that, for any $g \in [G]$,*

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \left(\hat{\alpha}_{\sigma(g)t} - \alpha^0_{gt}\right)^2 = O_p\left((N\wedge T)^{-1/2}\right).$$

Lemmas A.4 and A.5 imply that $d_H(\hat{\alpha}, \alpha^0) = O_p\left((N\wedge T)^{-1/2}\right)$, where, because the objective function is invariant to a relabeling of the groups, distances between vectors of group effects are measured by the Hausdorff distance

$$d_H(a,b)^2 := \max\left\{\max_{g\in[G]}\min_{\tilde{g}\in[G]} \|a_{\tilde{g}} - b_g\|_T^2,\ \max_{\tilde{g}\in[G]}\min_{g\in[G]} \|a_{\tilde{g}} - b_g\|_T^2\right\}, \qquad a, b \in \mathcal{A}^{GT},$$

with $a_g := (a_{gt})_{t\in[T]}$ and $\|\cdot\|_T$ as defined at the beginning of this Appendix. For the rest of the proof, I take $\sigma(g) = g$ without loss of generality. The rate $(N\wedge T)^{-1/2}$ for these preliminary estimates is what dictates the separation rate in Assumption 3(b): the neighborhood in which classification can be controlled uniformly (Lemma A.6) must contain the preliminary estimates, and its radius must be of smaller order than $\sqrt{c_T}$.

To establish that the group membership structure is correct asymptotically in a local neighborhood of the true values, for any $\eta > 0$ let

$$\mathcal{N}_\eta := \left\{(\beta,\alpha) \in \mathcal{B}\times\mathcal{A}^{GT} : \|\beta - \beta^0\|^2 < \eta,\ \max_{g\in[G]} \frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt} - \alpha^0_{gt}\right)^2 < \eta\right\}.$$

Lemma A.6. *Suppose that Assumptions 1 and 3 hold, and let $\eta = \eta_{NT}$ be defined by $\sqrt{\eta_{NT}} = (N\wedge T)^{-1/4+f}$ for some $0 < f < e$, with e from Assumption 3(b). As $N, T \to \infty$ such that $NT^{-r/2}c_T^{-r} = o(1)$, it holds that*

$$\mathbb{P}\left(\hat{g}(\beta,\alpha) \neq g^0 \text{ for some } (\beta,\alpha) \in \mathcal{N}_\eta\right) = O\left(NT^{-r/2}\left(1 + c_T^{-r}\right)\right) = o(1).$$

Let us now put the previous results together and show the two parts of the theorem.

Proof of Theorem 2. For part (a), note that

$$\mathbb{P}\left(\hat{g} \neq g^0\right) \leq \mathbb{P}\left(\hat{g} \neq g^0, (\hat{\beta}, \hat{\alpha}) \in \mathcal{N}_\eta\right) + \mathbb{P}\left((\hat{\beta}, \hat{\alpha}) \notin \mathcal{N}_\eta\right).$$

The first term on the RHS is $o(1)$ by Lemma A.6, with $\eta = \eta_{NT}$ as defined there, and the second is $o(1)$ because, by Lemmas A.4 and A.5, $\|\hat{\beta} - \beta^0\|^2$ and $\max_g T^{-1}\sum_t(\hat{\alpha}_{gt} - \alpha^0_{gt})^2$ are $O_p((N \wedge T)^{-1/2}) = o_p(\eta_{NT})$, so that $(\hat{\beta}, \hat{\alpha}) \in \mathcal{N}_{\eta_{NT}}$ with probability approaching one; note that $NT^{-r/2}c_T^{-r} = o(1)$ implies $NT^{-r/2} \to 0$ because $c_T \leq C$ by Assumption 1(a). Hence the result follows.

For part (b), observe that

$$\mathbb{P}\left(\left\|\hat{\beta} - \tilde{\beta}\right\| > C/\sqrt{NT}\right) \leq \mathbb{P}\left(\left\|\hat{\beta} - \tilde{\beta}\right\| > C/\sqrt{NT}, \hat{g} = g^0\right) + \mathbb{P}\left(\hat{g} \neq g^0\right).$$

The first term on the RHS is zero because $\hat{\beta} = \tilde{\beta}$ under $\hat{g} = g^0$, and the second term is $o(1)$ by part (a) of the theorem. This concludes the proof. □

B Proofs of Auxiliary Results

Proof of Lemma A.1. The argument is the first half of the proof of Lemma B.3 in BM.

For all $g \in [G^0]$, define

$$\sigma(g) = \underset{\tilde{g} \in [G]: \hat{\mathcal{C}}_{\tilde{g}} \neq \emptyset}{\arg\min} \frac{1}{T}\sum_{t=1}^{T}\left(\hat{\alpha}_{\tilde{g}t} - \alpha^0_{gt}\right)^2.$$

If there are multiple minimizers, just pick one of them. The minimum is over the nonempty estimated groups, which is the set over which the lemma is stated. Since $\hat{\mathcal{C}}_{\hat{g}_i} \ni i$ is nonempty for every i, the inequality below remains valid.

Let $N_g = \sum_{i=1}^N \mathbb{1}\{g_i^0 = g\}$ denote the size of group g. By the definition of σ, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\left(\hat{\alpha}_{\sigma(g)t} - \alpha^0_{gt}\right)^2 &= \frac{1}{TN_g}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{1}\{g_i^0 = g\}\left(\hat{\alpha}_{\sigma(g_i^0)t} - \alpha^0_{g_i^0 t}\right)^2 \\ &\leq \frac{1}{TN_g}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{1}\{g_i^0 = g\}\left(\hat{\alpha}_{\hat{g}_i t} - \alpha^0_{g_i^0 t}\right)^2\end{aligned}$$

for any $g \in [G^0]$. Therefore,

$$\max_{g \in [G^0]}\frac{1}{T}\sum_{t=1}^{T}\left(\hat{\alpha}_{\sigma(g)t} - \alpha^0_{gt}\right)^2 \leq \max_{g \in [G^0]}\frac{1}{TN_g}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{1}\{g_i^0 = g\}\left(\hat{\alpha}_{\hat{g}_i t} - \alpha^0_{g_i^0 t}\right)^2,$$

which is $o_p(1)$ by Theorem 1 in BM. In particular, note that Theorem 1 in BM is unaffected by $G \geq G^0$. This is sufficient for the result by Assumption 2(a). □

Proof of Lemma A.2. Write $e := \hat{\beta}_G - \beta^0$, $u_i := v_i - x_i e \in \mathbb{R}^T$, with x_i the $T \times d$ matrix of covariates of unit i, and $r_i := y_i - x_i\hat{\beta}_G = \alpha^0_{g_i^0} + u_i$, where $\alpha^0_g := (\alpha^0_{gt})_{t \in [T]}$. For a nonempty $S \subseteq [N]$ let $\bar{r}_S$, $\bar{u}_S$ be the averages of r_i, u_i over $i \in S$, $w_g(S) := |S \cap \mathcal{C}^0_g|/|S|$ the composition of S, $\bar{\alpha}(S) := \sum_{g \in [G^0]} w_g(S)\alpha^0_g$,

and, for $i \in S$,

$$D_i(S) := \alpha^0_{g_i^0} - \bar{\alpha}(S) = \sum_{g \in [G^0]} w_g(S) \left(\alpha^0_{g_i^0} - \alpha^0_g \right),$$

so that $r_i - \bar{r}_S = D_i(S) + u_i - \bar{u}_S$. Note that $\sum_{i \in S} D_i(S) = 0$ and that $D_i(S)$ is a convex combination of the fixed vectors $\alpha^0_{g_i^0} - \alpha^0_g$. Given $\hat{\beta}_G$, the pair $(\hat{\alpha}, \hat{\gamma})$ minimizes $\sum_i \|r_i - a_{g_i}\|^2$ over (a, γ), so that $\hat{\alpha}_g = \bar{r}_{\hat{\mathcal{C}}_g}$ for every nonempty estimated group and $\mathrm{SSR}(\hat{\gamma}) := \sum_g \sum_{i \in \hat{\mathcal{C}}_g} \|r_i - \bar{r}_{\hat{\mathcal{C}}_g}\|^2 \leq \mathrm{SSR}(\gamma)$ for every $\gamma \in \Gamma_G$, where $\mathrm{SSR}(\gamma)$ is defined with the cell means of γ. Empty estimated groups, if any, play no role in what follows (the statement of the lemma concerns nonempty groups only).

Let $c := \min_{g \neq \tilde{g}} c_{g,\tilde{g}} > 0$ from Assumption 2(b), and fix $\varepsilon > 0$ small (to be chosen). By Lemmas A.1 and B.1, Assumptions 2(a)–(b) and the consistency of $\hat{\beta}_G$ (Theorem 1 in BM, which holds for $G \geq G^0$), there is an event A_{NT} with $\mathbb{P}(A_{NT}) \to 1$ on which:

(i) $\|\alpha^0_g - \alpha^0_{\tilde{g}}\|^2_T \geq c/2$ for all $g \neq \tilde{g}$ in $[G^0]$, and $|\mathcal{C}^0_g| \geq \pi_g N/2$ for all g;

(ii) $\max_i \|u_i\|^2_T \leq C_u$ for a constant C_u; $\zeta := \max_{i \neq j} |\langle u_i, u_j \rangle_T| \leq \varepsilon$; $\xi := \max_i \max_{g \neq \tilde{g}} |\langle \alpha^0_g - \alpha^0_{\tilde{g}}, u_i \rangle_T| \leq \varepsilon$;

(iii) $\eta_1 := \max_{g \in [G^0]} \min_{\tilde{g} \in [G]: \hat{\mathcal{C}}_{\tilde{g}} \neq \emptyset} \|\hat{\alpha}_{\tilde{g}} - \alpha^0_g\|_T \leq \varepsilon$.

Indeed, (i) follows from Assumption 2(a)–(b); for (ii), write $u_i = v_i - x_i e$ and use Lemma B.1: $\|x_i e\|_T \leq \sqrt{M_2}\|e\| = o_p(1)$ uniformly in i, $\|u_i\|_T \leq \|v_i\|_T + \|x_i e\|_T$ (so that $C_u = 2C_v$ works, with C_v from Lemma B.1(a)),

$$|\langle u_i, u_j \rangle_T| \leq |\langle v_i, v_j \rangle_T| + 2 \max_k \|v_k\|_T \max_k \|x_k e\|_T + \max_k \|x_k e\|^2_T$$

and $|\langle \alpha^0_g - \alpha^0_{\tilde{g}}, u_i \rangle_T| \leq |\langle \alpha^0_g - \alpha^0_{\tilde{g}}, v_i \rangle_T| + \|\alpha^0_g - \alpha^0_{\tilde{g}}\|_T \|x_i e\|_T$, where $\|\alpha^0_g - \alpha^0_{\tilde{g}}\|_T$ is bounded by Assumption 1(a); (iii) is Lemma A.1. For every $g \in [G^0]$, let $\sigma(g) \in [G]$ attain the minimum in (iii); $\hat{\mathcal{C}}_{\sigma(g)}$ is nonempty by construction.

I work on A_{NT} throughout and show that, for ε small enough and N, T large enough, every nonempty estimated group is contained in a true group, which proves the lemma. I proceed in five steps.

(a) *A mixed group contains a unit with a large D_i.*

If S contains units i, j with $g^0_i \neq g^0_j$, then $D_i(S) - D_j(S) = \alpha^0_{g_i^0} - \alpha^0_{g_j^0}$ and, by the triangle inequality and (i),

$$\max\{\|D_i(S)\|_T, \|D_j(S)\|_T\} \geq \tfrac{1}{2}\|\alpha^0_{g_i^0} - \alpha^0_{g_j^0}\|_T \geq \tfrac{1}{2}\sqrt{c/2},$$

i.e., $\max\{\|D_i(S)\|^2_T, \|D_j(S)\|^2_T\} \geq c/8$.

(b) *Unit-level optimality bounds D_i by the inverse group size.*

Let $S = \hat{\mathcal{C}}_g$ be an estimated group with $m := |S| \geq 2$, let $i \in S$ and $S' := \hat{\mathcal{C}}_{\sigma(g_i^0)}$. Suppose first $S' \neq S$. By Lemma B.3(a), moving i from S to S' cannot decrease the objective, which gives

$$\|r_i - \bar{r}_S\|^2_T \leq \|r_i - \bar{r}_{S'}\|^2_T.$$

Since $\bar{r}_{S'} = \hat{\alpha}_{\sigma(g_i^0)}$, (iii) gives $\|r_i - \bar{r}_{S'}\|_T \leq \|\alpha^0_{g_i^0} - \hat{\alpha}_{\sigma(g_i^0)}\|_T + \|u_i\|_T \leq \eta_1 + \|u_i\|_T$, so the right-hand

side is at most $\|u_i\|_T^2 + 2\sqrt{C_u}\eta_1 + \eta_1^2$. For the left-hand side, $r_i - \bar{r}_S = D_i(S) + u_i - \bar{u}_S$ and

$$\|r_i - \bar{r}_S\|_T^2 = \|D_i(S)\|_T^2 + \|u_i - \bar{u}_S\|_T^2 + 2\langle D_i(S), u_i - \bar{u}_S\rangle_T \geq \|D_i(S)\|_T^2 + \|u_i - \bar{u}_S\|_T^2 - 4\xi,$$

because $|\langle D_i(S), u_j\rangle_T| \leq \sum_g w_g(S)|\langle \alpha^0_{g_i^0} - \alpha^0_g, u_j\rangle_T| \leq \xi$ for every j, hence also $|\langle D_i(S), \bar{u}_S\rangle_T| \leq \xi$. Moreover, writing $u_i - \bar{u}_S = (1-m^{-1})u_i - m^{-1}\sum_{j\in S\setminus\{i\}} u_j$,

$$\|u_i - \bar{u}_S\|_T^2 \geq (1-m^{-1})^2\|u_i\|_T^2 - 2(1-m^{-1})m^{-1}\sum_{j\in S\setminus\{i\}}\langle u_i, u_j\rangle_T \geq (1-m^{-1})^2\|u_i\|_T^2 - 2\zeta.$$

Collecting terms and using $1-(1-m^{-1})^2 \leq 2/m$,

$$\|D_i(S)\|_T^2 \leq \frac{2}{m}\|u_i\|_T^2 + 2\zeta + 4\xi + 2\sqrt{C_u}\eta_1 + \eta_1^2 \leq \frac{2C_u}{m} + \kappa, \qquad \kappa := 2\zeta + 4\xi + 2\sqrt{C_u}\eta_1 + \eta_1^2. \tag{B.1}$$

If instead $S' = S$, then (iii) gives $\|\bar{\alpha}(S) + \bar{u}_S - \alpha^0_{g_i^0}\|_T \leq \eta_1$, so $\|D_i(S)\|_T \leq \eta_1 + \|\bar{u}_S\|_T$, and $\|\bar{u}_S\|_T^2 = m^{-2}\sum_{j,k\in S}\langle u_j, u_k\rangle_T \leq m^{-1}C_u + \zeta$, hence (B.1) holds in this case as well (with $2\eta_1^2 + 2\zeta$ in place of κ, which is smaller as soon as $\eta_1 \leq 2\sqrt{C_u}$, i.e., for $\varepsilon \leq 2\sqrt{C_u}$). On A_{NT}, $\kappa \leq \kappa(\varepsilon) := 6\varepsilon + 2\sqrt{C_u}\varepsilon + \varepsilon^2$.

(c) *Large estimated groups are pure.*

Choose ε so small that $\kappa(\varepsilon) < c/16$, and let $m_0 := \lceil 32C_u/c\rceil$. If an estimated group S with $|S| > m_0$ were mixed, step (a) would give a unit $i \in S$ with $\|D_i(S)\|_T^2 \geq c/8$, while (B.1) gives $\|D_i(S)\|_T^2 < c/16 + c/16 = c/8$, a contradiction. Hence every estimated group of size larger than m_0 is contained in a true group.

(d) *Every true group has a large pure estimated group.*

By (i), the $|\mathcal{C}^0_g| \geq \pi_g N/2$ units of true group g are spread over at most G estimated groups, so one of them, call it $R(g)$, contains at least $\pi_g N/(2G) > m_0$ units of group g for N large. By step (c), $R(g)$ is pure, hence $R(g) \subseteq \mathcal{C}^0_g$ and $|R(g)| \geq \pi_g N/(2G)$. Its effect satisfies $\bar{r}_{R(g)} = \alpha^0_g + \bar{u}_{R(g)}$ with $\|\bar{u}_{R(g)}\|_T^2 \leq C_u/|R(g)| + \zeta \leq \eta_2^2$, $\eta_2^2 := 2GC_u/(\pi N) + \varepsilon$, where $\pi := \min_g \pi_g$.

(e) *Small estimated groups are pure.*

Suppose an estimated group S with $2 \leq m = |S| \leq m_0$ is mixed, and let $i^* \in S$ maximize $\|u_j\|_T$ over $j \in S$. Consider the partition γ' obtained from $\hat{\gamma}$ by keeping i^* alone in the group of S and moving every other $j \in S$ to $R(g_j^0)$, which is a different group since $|R(g_j^0)| > m_0 \geq m$. By Lemma B.3(b), the contribution of the enlarged group $R(g_j^0)\cup\{j\}$ to $\mathrm{SSR}(\gamma')$ is at most its sum of squared deviations from the old mean $\bar{r}_{R(g_j^0)}$, so that

$$\mathrm{SSR}(\gamma') - \mathrm{SSR}(\hat{\gamma}) \leq T\sum_{j\in S\setminus\{i^*\}}\|r_j - \bar{r}_{R(g_j^0)}\|_T^2 - T\sum_{j\in S}\|r_j - \bar{r}_S\|_T^2.$$

By step (d), $\|r_j - \bar{r}_{R(g_j^0)}\|_T = \|u_j - \bar{u}_{R(g_j^0)}\|_T \leq \|u_j\|_T + \eta_2$, so the first sum is at most $\sum_{j\neq i^*}\|u_j\|_T^2 + (m-1)(2\sqrt{C_u}\eta_2 + \eta_2^2)$. For the second sum, $r_j - \bar{r}_S = D_j(S) + u_j - \bar{u}_S$ and, using $\sum_{j\in S} D_j(S) = 0$,

$|\langle D_j(S), u_j\rangle_T| \le \xi$ and Lemma B.3(c),

$$\begin{aligned}\sum_{j\in S}\|r_j-\bar r_S\|_T^2 &= \sum_{j\in S}\|D_j(S)\|_T^2+\sum_{j\in S}\|u_j\|_T^2 - m\|\bar u_S\|_T^2 + 2\sum_{j\in S}\langle D_j(S),u_j\rangle_T\\ &\ge \sum_{j\in S}\|D_j(S)\|_T^2+\sum_{j\in S}\|u_j\|_T^2 - \|u_{i^*}\|_T^2 - (m-1)\zeta - 2m\xi,\end{aligned}$$

where $m\|\bar u_S\|_T^2 = m^{-1}\sum_{j,k\in S}\langle u_j,u_k\rangle_T \le \max_{j\in S}\|u_j\|_T^2 + (m-1)\zeta$ was used. Combining, and using step (a),

$$\begin{aligned}T^{-1}\left[\mathrm{SSR}(\gamma') - \mathrm{SSR}(\hat\gamma)\right] &\le -\sum_{j\in S}\|D_j(S)\|_T^2 + (m-1)\left(2\sqrt{C_u}\eta_2+\eta_2^2+\zeta\right) + 2m\xi\\ &\le -\frac{c}{8} + m_0\left(2\sqrt{C_u}\eta_2+\eta_2^2+3\varepsilon\right),\end{aligned}$$

which is negative for ε small and N large, because $m_0 = \lceil 32C_u/c\rceil$ depends neither on N,T nor on ε, while $\eta_2\to 0$ as $\varepsilon\to 0$ and $N\to\infty$ (shrink ε further if necessary, this does not affect m_0). This contradicts the optimality of $\hat\gamma$. Hence no estimated group of size between 2 and m_0 is mixed, groups of size one are trivially pure, and step (c) covers the remaining ones. The lemma follows.

□

Proof of Lemma A.3. I proceed in five steps.

(a) *Profiling out grouped fixed-effects.*

Fix any partition $\gamma = \{g_1,\dots,g_N\} \in \Gamma_G$, and let $\mathcal{C}_{\tilde g}(\gamma) = \{i\in[N] : g_i = \tilde g\}$ for $\tilde g\in[G]$. For each group g and time t, define

$$\bar y_{gt}(\gamma) = \frac{1}{|\mathcal{C}_g(\gamma)|}\sum_{i\in\mathcal{C}_g(\gamma)} y_{it}, \qquad \bar x_{gt}(\gamma) = \frac{1}{|\mathcal{C}_g(\gamma)|}\sum_{i\in\mathcal{C}_g(\gamma)} x_{it},$$

and residualize

$$\widetilde y_{it}(\gamma) := y_{it} - \bar y_{g_it}(\gamma), \qquad \widetilde x_{it}(\gamma) := x_{it} - \bar x_{g_it}(\gamma).$$

Profiling the GFE objective function in (2) over $\{a_{gt}\}$ yields the conditional (on γ) GFE estimator

$$\widehat\beta(\gamma) = \widehat W(\gamma)^{-1}\widehat S(\gamma),$$

where

$$\widehat W(\gamma) := \frac{1}{NT}\sum_{i,t}\widetilde x_{it}(\gamma)\widetilde x_{it}(\gamma)', \qquad \widehat S(\gamma) := \frac{1}{NT}\sum_{i,t}\widetilde x_{it}(\gamma)\widetilde y_{it}(\gamma). \tag{B.2}$$

Note that, by Lemma B.5, with probability approaching one $\widehat W(\gamma)^{-1}$ exists and $\|\widehat W(\gamma)^{-1}\| \le 2/\rho$ for every $\gamma\in\Gamma_G$.

Under the grouped panel model (1) we obtain, for any partition γ,

$$\widehat W(\gamma)(\widehat\beta(\gamma)-\beta^0) = \underbrace{\frac{1}{NT}\sum_{i,t}\widetilde x_{it}(\gamma)(\alpha^0_{g_i^0 t} - \bar\alpha^0_{g_it}(\gamma))}_{\text{approximation term}} + \underbrace{\frac{1}{NT}\sum_{i,t}\widetilde x_{it}(\gamma)(v_{it}-\bar v_{g_it}(\gamma))}_{\text{stochastic term}}, \tag{B.3}$$

where $\bar{\alpha}^0_{gt}(\gamma) = |\mathcal{C}_g(\gamma)|^{-1} \sum_{j \in \mathcal{C}_g(\gamma)} \alpha^0_{g^0_j t}$ and $\bar{v}_{gt}(\gamma) = |\mathcal{C}_g(\gamma)|^{-1} \sum_{j \in \mathcal{C}_g(\gamma)} v_{jt}$.

Let $\widehat{\gamma}^{(G)}$ (respectively $\widehat{\gamma}^{(G^0)}$) denote the partition minimizing the objective with exactly G (resp. G^0) groups. Thus, $\widehat{\beta}_G = \widehat{\beta}(\widehat{\gamma}^{(G)})$ and $\widehat{\beta}_{G^0} = \widehat{\beta}(\widehat{\gamma}^{(G^0)})$.

(b) *Approximation terms vanish on E.*

By definition of E in (A.1), on E each estimated group in $\widehat{\gamma}^{(G)}$ is a subset of a single true group, hence for every i, t,

$$\alpha^0_{g^0_i t} = \bar{\alpha}^0_{g_i t}(\widehat{\gamma}^{(G)}),$$

and the approximation term in (B.3) is identically zero when $\gamma = \widehat{\gamma}^{(G)}$, for any $G \geq G^0$. Trivially, the same holds for $\gamma = \hat{\gamma}^{(G^0)}$.

(c) *ANOVA-type identity for cross-products under refinements.*

Fix a true group $\mathcal{C}^0_g$ and time t. Suppose $\mathcal{C}^0_g$ is partitioned into disjoint subgroups S, i.e., the cells induced by a refinement γ. For each subgroup S, define subgroup means $\bar{x}_{St}$, $\bar{v}_{St}$ and the coarse group means $\bar{x}_{g^0 t}, \bar{v}_{g^0 t}$.

Start from the sum of within-subgroup cross-terms

$$\sum_{i \in S} (x_{it} - \bar{x}_{St})(v_{it} - \bar{v}_{St}).$$

Write $x_{it} - \bar{x}_{St} = (x_{it} - \bar{x}_{g^0 t}) + (\bar{x}_{g^0 t} - \bar{x}_{St})$ and similarly for v. Expanding yields four terms:

$$\begin{aligned}\sum_{i \in S} (x_{it} - \bar{x}_{g^0 t})(v_{it} - \bar{v}_{g^0 t}) + (\bar{x}_{g^0 t} - \bar{x}_{St}) \sum_{i \in S} (v_{it} - \bar{v}_{g^0 t}) \\ + (\bar{v}_{g^0 t} - \bar{v}_{St}) \sum_{i \in S} (x_{it} - \bar{x}_{g^0 t}) + |S|(\bar{x}_{g^0 t} - \bar{x}_{St})(\bar{v}_{g^0 t} - \bar{v}_{St}).\end{aligned}$$

But $\sum_{i \in S}(v_{it} - \bar{v}_{g^0 t}) = |S|(\bar{v}_{St} - \bar{v}_{g^0 t})$ and similarly for x. Substituting and canceling gives that, for each subgroup S, we have the exact algebraic identity

$$\sum_{i \in S} (x_{it} - \bar{x}_{St})(v_{it} - \bar{v}_{St}) = \sum_{i \in S} (x_{it} - \bar{x}_{g^0 t})(v_{it} - \bar{v}_{g^0 t}) - |S|(\bar{x}_{St} - \bar{x}_{g^0 t})(\bar{v}_{St} - \bar{v}_{g^0 t}).$$

Summing this identity over all subgroups S inside the group $\mathcal{C}^0_g$ and then summing over all groups g and times t yields the global within–between identity

$$\begin{aligned}\sum_{g=1}^{G^0} \sum_{S \subset \mathcal{C}^0_g} \sum_{i \in S} (x_{it} - \bar{x}_{St})(v_{it} - \bar{v}_{St}) - \sum_{g=1}^{G^0} \sum_{i \in \mathcal{C}^0_g} (x_{it} - \bar{x}_{g^0 t})(v_{it} - \bar{v}_{g^0 t}) \\ = - \sum_{g=1}^{G^0} \sum_{S \subset \mathcal{C}^0_g} |S|(\bar{x}_{St} - \bar{x}_{g^0 t})(\bar{v}_{St} - \bar{v}_{g^0 t}).\end{aligned} \tag{B.4}$$

Define the score coming from the stochastic term in (B.3) as

$$\Delta(\gamma) := \frac{1}{NT} \sum_{i,t} \widetilde{x}_{it}(\gamma)\, \widetilde{v}_{it}(\gamma). \tag{B.5}$$

The ANOVA-type identity in (B.4) is purely algebraic, holds for each t and, when divided by NT and interpreted component-wise, gives the exact difference between the score evaluated at a refinement and the score evaluated at the coarse (true) partition:[10]

$$\Delta(\gamma) - \Delta(\gamma^0) = -\frac{1}{NT} \sum_{t=1}^{T} \sum_{g=1}^{G^0} \sum_{S \subset \mathcal{C}_g^0} |S| (\bar{x}_{St} - \bar{x}_{g^0 t})(\bar{v}_{St} - \bar{v}_{g^0 t}). \tag{B.6}$$

(d) *An exact expression for the difference of estimators.*

Fix a refinement γ of γ^0 and use the notation $\Omega(\gamma)$, $\delta(\gamma)$, $D(\gamma)$ introduced before Assumption 5. Identity (B.6) reads $\Delta(\gamma) = \Delta(\gamma^0) - \delta(\gamma)$, and the same ANOVA-type identity applied to the cross-products $x_{it}x_{it}'$ instead of $x_{it}v_{it}$ gives

$$\widehat{W}(\gamma) = \widehat{W}(\gamma^0) - \Omega(\gamma), \qquad \Omega(\gamma) \succeq 0, \tag{B.7}$$

so that $\widehat{W}(\gamma) \preceq \widehat{W}(\gamma^0)$: refining the partition can only remove covariate variation. Since the approximation terms in (B.3) vanish for refinements, $\widehat{\beta}(\gamma) - \beta^0 = \widehat{W}(\gamma)^{-1}\Delta(\gamma)$ and $\widehat{\beta}(\gamma^0) - \beta^0 = \widehat{W}(\gamma^0)^{-1}\Delta(\gamma^0)$. Writing $\Delta(\gamma^0) = \widehat{W}(\gamma^0)(\widehat{\beta}(\gamma^0) - \beta^0) = [\widehat{W}(\gamma) + \Omega(\gamma)](\widehat{\beta}(\gamma^0) - \beta^0)$ and substituting yields the exact decomposition

$$\widehat{\beta}(\gamma) - \widehat{\beta}(\gamma^0) = \widehat{W}(\gamma)^{-1}\Big[\Omega(\gamma)(\widehat{\beta}(\gamma^0) - \beta^0) - \delta(\gamma)\Big], \tag{B.8}$$

which is valid for every refinement γ of γ^0 and isolates the only two channels through which the two estimators can differ: the covariate variation absorbed by the refinement, $\Omega(\gamma)$, which acts on the sampling error of the coarse estimator, and the covariate–error covariance generated by the refinement, $\delta(\gamma)$.

Let E_0 denote the event that $\hat{\gamma}^{(G^0)}$ coincides with γ^0 up to a relabeling of the groups, and note that $\mathbb{P}(E_0) = 1 - o(1)$ by Theorem 2 in BM (classification consistency), whose conditions are those of Theorem 1 with $G = G^0$. On $E \cap E_0$, $\hat{\beta}_{G^0} = \widehat{\beta}(\gamma^0)$ because the residualization operator is invariant to a relabeling of the cells, and $\hat{\beta}_G = \widehat{\beta}(\hat{\gamma})$ with $\hat{\gamma} := \hat{\gamma}^{(G)}$ a refinement of γ^0. Hence, on $E \cap E_0$, (B.8) gives

$$\sqrt{NT}\left(\hat{\beta}_G - \hat{\beta}_{G^0}\right) = \widehat{W}(\hat{\gamma})^{-1}\Big[\Omega(\hat{\gamma})\sqrt{NT}(\hat{\beta}_{G^0} - \beta^0) - \sqrt{NT}\,\delta(\hat{\gamma})\Big]. \tag{B.9}$$

Two of the three ingredients in (B.9) are controlled directly. By Lemma B.5, with probability approaching one $\|\widehat{W}(\hat{\gamma})^{-1}\| \le 2/\rho$. Moreover, on E_0, $\sqrt{NT}\Delta(\gamma^0) = \widehat{W}(\gamma^0)\sqrt{NT}(\hat{\beta}_{G^0} - \beta^0)$, which is $O_p(1)$ by the hypothesis of the lemma, since $\|\widehat{W}(\gamma^0)\| = O_p(1)$ by Assumption 1(b) and Markov's

[10] For each true group $\mathcal{C}_g^0$, the notation $\{S \subset \mathcal{C}_g^0\}$ denotes the (nonempty) blocks produced by intersecting $\mathcal{C}_g^0$ with the cells of the partition γ, $\{S : S = \mathcal{C}_g^0 \cap \mathcal{C}_{g'}(\gamma) \text{ for some } g'\}$. Thus, formally, the set of blocks S inside each true group $\mathcal{C}^0$ change with γ. This is left implicit in the notation.

inequality. It therefore remains to show that

$$\Omega(\hat{\gamma}) = o_p(1) \qquad \text{and} \qquad \sqrt{NT}\,\delta(\hat{\gamma}) = o_p(1). \tag{B.10}$$

Neither statement can be established uniformly over the refinements of γ^0, and for fixed refinements both may fail (Remark 1). The argument must therefore exploit that $\hat{\gamma}$ is not an arbitrary refinement but the minimizer of the GFE objective. This is done in the next step.

(e) *Localization at the estimated partition and conclusion.*

For any $b \in \mathcal{B}$ and partition γ, let $r_{it}(b) := y_{it} - x_{it}'b$ and $\widetilde{r}_{it}(\gamma; b) := r_{it}(b) - \bar{r}_{g_i t}(\gamma; b)$ denote the residualized outcome. Because $(\hat{\beta}_G, \hat{\alpha}, \hat{\gamma})$ jointly minimize the objective in (2), profiling out the grouped effects shows that $(\hat{\beta}_G, \hat{\gamma})$ minimizes $\sum_{i,t} \widetilde{r}_{it}(\gamma; b)^2$ over $(b, \gamma) \in \mathcal{B} \times \Gamma_G$. For a refinement γ of γ^0, $r_{it}(b) = \alpha^0_{g_i^0 t} + v_{it} - x_{it}'(b - \beta^0)$ and the cell means absorb $\alpha^0_{g_i^0 t}$ exactly, so that $\widetilde{r}_{it}(\gamma; b) = \widetilde{v}_{it}(\gamma) - \widetilde{x}_{it}(\gamma)'(b - \beta^0)$. Since $\hat{\beta}_G = \widehat{\beta}(\hat{\gamma})$ by step (a), the value of the objective at the GFE solution is, on E,

$$\begin{aligned} \sum_{i,t} \widetilde{r}_{it}(\hat{\gamma}; \hat{\beta}_G)^2 &= \sum_{i,t} \widetilde{v}_{it}(\hat{\gamma})^2 - NT\,\Delta(\hat{\gamma})'\widehat{W}(\hat{\gamma})^{-1}\Delta(\hat{\gamma}) \\ &= \sum_{i,t} \widetilde{v}_{it}(\gamma^0)^2 - NT\,D(\hat{\gamma}) - NT\,\Delta(\hat{\gamma})'\widehat{W}(\hat{\gamma})^{-1}\Delta(\hat{\gamma}), \end{aligned} \tag{B.11}$$

where the second equality applies the ANOVA-type identity of step (c) to the squares v_{it}^2. The same identity gives, for the value of the objective at $(b, \gamma) = (\beta^0, \tilde{\gamma})$,

$$\sum_{i,t} \widetilde{r}_{it}(\tilde{\gamma}; \beta^0)^2 = \sum_{i,t} \widetilde{v}_{it}(\tilde{\gamma})^2 = \sum_{i,t} \widetilde{v}_{it}(\gamma^0)^2 - NT\,D(\tilde{\gamma}). \tag{B.12}$$

Two remarks are in order. First, (B.11) shows that, on E, $\hat{\gamma}$ maximizes the sum

$$NTD(\gamma) + NT\,\Delta(\gamma)'\widehat{W}(\gamma)^{-1}\Delta(\gamma)$$

over the refinements of γ^0: if β^0 were known, the second term would be absent and the GFE classification would maximize $D(\gamma)$, i.e., it would be the partition $\tilde{\gamma}$ of Assumption 5—with the true groups already accounted for, the extra groups are chosen to fit the idiosyncratic errors. Second, the covariates enter the classification only through the second term, the reduction in the sum of squared residuals obtained by re-estimating the slope at γ. By (B.6), it equals $NT\,[\Delta(\gamma^0) - \delta(\gamma)]'\,\widehat{W}(\gamma)^{-1}\,[\Delta(\gamma^0) - \delta(\gamma)]$, so it rewards refinements that generate a large covariate–error covariance $\delta(\gamma)$. This is exactly the mechanism that invalidates any uniform-in-γ argument, and Assumption 5 states that it cannot overturn the fit to the errors.

Let $z := \sqrt{NT}\|\Delta(\gamma^0)\|$ and $u := \sqrt{NT}\|\delta(\hat{\gamma})\|$. Since the GFE solution minimizes the objective, (B.11) is at most (B.12). Thus, rearranging and using $D(\tilde{\gamma}) \geq D(\hat{\gamma})$ yields, on the event on which $\|\widehat{W}(\hat{\gamma})^{-1}\| \leq 2/\rho$ (which has probability approaching one by Lemma B.5), the key inequality

$$0 \leq L(\hat{\gamma}) = NT\,[D(\tilde{\gamma}) - D(\hat{\gamma})] \leq NT\,\Delta(\hat{\gamma})'\widehat{W}(\hat{\gamma})^{-1}\Delta(\hat{\gamma}) \leq 2\rho^{-1}\,(z + u)^2\,. \tag{B.13}$$

The loss in the fit to the errors incurred by the estimated partition is bounded by what the covariate

channel can return. Now let $c > 1$ be the constant of Assumption 5 and put $\eta := c - 1$. Writing $\Delta(\hat\gamma) = \Delta(\gamma^0) - \delta(\hat\gamma)$ and using $\|a - d\|_A^2 \le (1+\eta)\|d\|_A^2 + (1+\eta^{-1})\|a\|_A^2$ for the norm induced by the positive definite matrix $A = \widehat{W}(\hat\gamma)^{-1}$, the middle term in (B.13) is bounded by

$$L(\hat\gamma) \le 2\rho^{-1}(1+\eta^{-1})z^2 + c\,NT\,\delta(\hat\gamma)'\widehat{W}(\hat\gamma)^{-1}\delta(\hat\gamma), \tag{B.14}$$

so that $\hat\gamma \in \mathcal{N}_{c,K}$ whenever $2\rho^{-1}(1+\eta^{-1})z^2 \le K$. As $z = O_p(1)$ by the oracle rate, for every $\varepsilon > 0$ there is K such that $\mathbb{P}(\hat\gamma \in \mathcal{N}_{c,K}) \ge 1 - \varepsilon - o(1)$ on E, and on this event Assumption 5 gives $\|\Omega(\hat\gamma)\| + u \le \sup_{\gamma\in\mathcal{N}_{c,K}}\{\|\Omega(\gamma)\| + \sqrt{NT}\|\delta(\gamma)\|\} = o_p(1)$. Since ε is arbitrary, this establishes (B.10): the partition selected by the GFE estimator is, like $\tilde\gamma$, asymptotically uninformative about the covariates and generates no covariate–error covariance of the order of the sampling error, because selecting a partition with either property would worsen the fit to the errors by more than the covariate channel can return.

Finally, on the intersection of $E \cap E_0$ with the event of probability $1-o(1)$ on which $\|\widehat{W}(\hat\gamma)^{-1}\| \le 2/\rho$ (Lemma B.5), (B.9) gives

$$\sqrt{NT}\left\|\hat\beta_G - \hat\beta_{G^0}\right\| \le 2\rho^{-1}\Big[\|\Omega(\hat\gamma)\|\sqrt{NT}\|\hat\beta_{G^0} - \beta^0\| + u\Big] = o_p(1)\cdot O_p(1) + o_p(1) = o_p(1).$$

Hence, for every $C > 0$,

$$\mathbb{P}(\|\hat\beta_G - \hat\beta_{G^0}\| > C/\sqrt{NT}, E) \le \mathbb{P}(E_0^c) + o(1) + \mathbb{P}(\sqrt{NT}\|\hat\beta_G - \hat\beta_{G^0}\|\mathbb{1}\{E\cap E_0\} > C) = o(1).$$

Combining steps (a)–(e) proves the lemma. □

Remark 1. *One may wonder whether* $\|\Delta(\gamma)\| = O_p((NT)^{-1/2})$, $\Delta(\gamma) - \Delta(\gamma^0) = O_p(N^{-1}T^{-1/2})$ *and* $\widehat{W}(\gamma) - \widehat{W}(\gamma^0) = O_p(N^{-1})$ *uniformly over the refinements* γ *of* γ^0*, on the grounds that the between-cell mean differences* $\bar x_{St} - \bar x_{g^0 t}$ *and* $\bar v_{St} - \bar v_{g^0 t}$ *are* $O_p(N^{-1/2})$ *and that mixing over* t *shrinks the average of their products. This is not true for two distinct reasons, and the exact decomposition* (B.8) *shows precisely where. First, a refinement may be informative about the covariates. Suppose* $G^0 = 1$*, so that* γ^0 *is the trivial partition, and let* $x_{it} = f_t + \kappa_i h_t + \varepsilon_{it}$ *with* $\kappa_i \in \{0,1\}$ *for half of the units each,* $\{h_t\}$ *a stationary process. For the refinement* γ_κ *that splits the units according to* κ_i*,* $\bar x_{St} - \bar x_t = \pm h_t/2 + O_p(N^{-1/2})$*, hence* $\Omega(\gamma_\kappa) = \frac{1}{4T}\sum_{t=1}^T h_t^2 + o_p(1)$ *does not vanish and* $\sqrt{NT}\delta(\gamma_\kappa) = \frac{\sqrt{NT}}{4T}\sum_{t=1}^T h_t(\bar v_{S_1 t} - \bar v_{S_2 t}) + o_p(1) = O_p(1)$*. By* (B.8)*,* $\sqrt{NT}(\widehat\beta(\gamma_\kappa) - \widehat\beta(\gamma^0)) = O_p(1)$ *but not* $o_p(1)$*: the conditional estimator based on this fixed refinement is* $\sqrt{NT}$*-consistent, yet not oracle-equivalent—its asymptotic variance is inflated. Assumptions 1–2 do not exclude latent structure in the covariates beyond* γ^0*, so this possibility cannot be dismissed. Second, even when the covariates carry no such structure, so that* $\Omega(\gamma) = o_p(1)$ *uniformly over refinements, the rate* $\delta(\gamma) = O_p(N^{-1}T^{-1/2})$ *that an elementary variance calculation delivers at any fixed refinement cannot be made uniform: the cells of a refinement may be chosen as a function of the data so as to align covariates and errors. For instance, with* $G^0 = 1$ *and* x_{it}, v_{it} *i.i.d. standard normal, the refinement with cell* $S = \{i : \widetilde{x}_{i1} > 0,\ \widetilde{v}_{i1} > 0\}$ *has* $NT\delta(\gamma) = \frac{2}{3\pi}N + O_p(\sqrt{T})$*, which is not* $o(\sqrt{NT})$ *unless* $N/T \to 0$*. More generally, a union bound over the* G^N *refinements inflates the fixed-*γ *rate* $N^{-1}T^{-1/2}$ *by the factor* $\sqrt{N}$*, i.e., to the order* $(NT)^{-1/2}$ *of the sampling error itself, hence no argument that holds uniformly over refinements*

can deliver Lemma A.3. Instead, the data-dependence of $\hat{\gamma}$ must be used, and step (e) does so through the maximization property of $\hat{\gamma}$ at the cost of Assumption 5. In the example above, Assumption 5 is not violated by γ_κ: since κ_i is unrelated to the errors, γ_κ fits the errors no better than an arbitrary split, $NTD(\gamma_\kappa) = O_p(T)$, whereas $NTD(\tilde{\gamma})$ is the maximal between-cell variation of N weakly dependent noise vectors in $\mathbb{R}^T$ and exceeds that of an arbitrary split by an amount that diverges with N and T (heuristically, at least of order $\sqrt{NT}$), hence $L(\gamma_\kappa) \to \infty$ while $NT\|\delta(\gamma_\kappa)\|^2 = O_p(1)$, and $\gamma_\kappa \notin \mathcal{N}_{c,K}$ with probability approaching one. The refinements that matter for Assumption 5 in this example are those aligned with both the latent structure and the errors, such as $S = \{i : \kappa_i = 1,\ \sum_t h_t \widetilde{v}_{it} > 0\}$, for which a direct calculation (with κ_i independent of the errors, all shocks Gaussian and independent, and variances $\sigma_h^2, \sigma_\varepsilon^2, \sigma_v^2$) gives $NT\|\delta(\gamma)\|^2 \approx 0.018\, N\sigma_h^2\sigma_v^2$, $\widehat{W}(\gamma) \approx \sigma_\varepsilon^2 + \sigma_h^2/6$ and, when N/T is large, $L(\gamma) \gtrsim 0.42\, N\sigma_v^2$, so that the loss exceeds the covariate-channel gain $NT\delta(\gamma)'\widehat{W}(\gamma)^{-1}\delta(\gamma)$ by a factor of at least $24\sigma_\varepsilon^2/\sigma_h^2 + 4 \geq 4$ whatever the variances. Allowing the cells to be defined by thresholds on $\sum_t h_t \widetilde{v}_{it}$ that differ across the two values of κ_i lowers this factor, but the same calculation shows that it remains above 1.5 for all thresholds and variances, and the conclusion holds a fortiori when N/T is small, where the loss is of order $\sqrt{NT}$ rather than N. Such refinements therefore lie outside $\mathcal{N}_{c,K}$ for any $c < 1.5$, and the GFE estimator does not select them: Theorem 1(b) is consistent with the example, and what fails is the uniform argument. The calculation also makes clear that Assumption 5 is a condition on constants of the design, not on rates: it fails in designs where a latent structure of the covariates is so strongly aligned with the errors that splitting along it costs less in the fit to the errors than it returns through the covariate channel.

Remark 2. *A simpler, but stronger, sufficient condition is classification stability: $\mathbb{P}(\hat{\gamma}^{(G)} = \tilde{\gamma}) \to 1$, i.e., the estimation error in the slope does not alter the classification of the units into the extra groups. Under stability,* (B.10) *follows from Lemma B.4 alone, so Theorem 1(b) holds under Assumptions 1–2 and the primitive conditions (a)–(b) of Lemma B.4, without Assumption 5. Inequality* (B.13) *shows when stability can be expected. If $\hat{\gamma} \neq \tilde{\gamma}$, then $L(\hat{\gamma}) \leq 2\rho^{-1}(z+u)^2$ with $z = O_p(1)$, so stability holds whenever the classification problem for the errors has a margin,*

$$\inf_{\gamma \neq \tilde{\gamma}} \frac{L(\gamma)}{\left(1 + \sqrt{NT}\|\delta(\gamma)\|\right)^2} \to \infty \quad \text{in probability,}$$

the infimum being over the refinements of γ^0 in Γ_G other than $\tilde{\gamma}$. The margin condition requires every deviation from $\tilde{\gamma}$ to be costly in the fit to the errors. It is demanding: reassigning a single unit located near the boundary between two cells of $\tilde{\gamma}$ changes $NTD(\gamma)$ by an amount which, for independent errors with variance σ_v^2, is of order $\sigma_v^2\sqrt{T/N}$ for a typical unit and of order $\sigma_v^2\sqrt{T}N^{-3/2}$ for the unit closest to the boundary, so that the margin condition—and stability—can only be expected when $T/N^3 \to \infty$. When N is not small relative to T, a vanishing fraction of units near the cell boundaries does switch cells, and Assumption 5 is the appropriate condition: it tolerates such switches, requiring only that they do not make the classification informative about the covariates. The discussion following Assumption 5 in Appendix A explains why such switches can be expected to be harmless as long as the near-optimal classifications of the errors are not too numerous, which a heuristic count places in the regime $N^2/T = o(1)$.

Remark 3. *Assumption 5 is precisely what delivers the absence of variance inflation. Suppose instead that the supremum in Assumption 5 is $O_p(1)$ rather than $o_p(1)$. The argument of step (e) then gives $\Omega(\hat{\gamma}) = O_p(1)$ and $\sqrt{NT}\delta(\hat{\gamma}) = O_p(1)$, and* (B.9) *yields $\sqrt{NT}(\hat{\beta}_G - \hat{\beta}_{G^0}) = O_p(1)$: the GFE estimator with $G > G^0$ groups remains $\sqrt{NT}$-consistent, but it is no longer asymptotically equivalent to the oracle, and its limit distribution—*

which depends on the joint behavior of the estimated partition and the covariates—is not characterized by the present argument. The case of a fixed informative refinement in Remark 1 shows that the inflation is real: there, $\sqrt{NT}(\widehat{\beta}(\gamma_\kappa) - \widehat{\beta}(\gamma^0))$ has a non-degenerate limit.

Proposition 1. *Suppose the true groups $[G^0]$ are partitioned into clusters $K_1, \dots, K_{G^*}$, $1 \le G^* \le G^0$, let $k_i \in [G^*]$ denote the cluster of unit i, γ^* the partition of $[N]$ into the cluster cells $\mathcal{C}_k^* := \{i : k_i = k\}$, and*

$$a_{NT} := \max_{k \in [G^*]} \max_{g, \tilde{g} \in K_k} \|\alpha_g^0 - \alpha_{\tilde{g}}^0\|_T$$

the largest within-cluster distance. Suppose Assumptions 1, 2(a),(c)–(e) and 4 hold, that Assumption 2(b) holds for every pair of groups in different clusters, and that $N, T \to \infty$ with $NT^{-\nu} \to 0$ for some $\nu > 0$.

(a) If $a_{NT} = o(1)$, then for every $G \ge G^*$, $\mathbb{P}\left(\forall g \in [G],\ \exists k \in [G^*] :\ \hat{\mathcal{C}}_g \subseteq \mathcal{C}_k^*\right) = 1 - o(1)$.

(b) Let $\tilde{\beta}^* := \hat{\beta}(\gamma^*)$ be the least-squares estimator that uses the cluster partition. If, in addition, $a_{NT} = o(1/\sqrt{NT})$, $\sqrt{NT}(\tilde{\beta}^* - \beta^0) = O_p(1)$, and Assumption 5 holds with (γ^*, u_{it}) in place of (γ^0, v_{it}), where $u_{it} := v_{it} + \alpha^0_{g_i^0 t} - \alpha^*_{k_i t}$ and $\alpha^*_{kt} := |\mathcal{C}_k^*|^{-1} \sum_{i \in \mathcal{C}_k^*} \alpha^0_{g_i^0 t}$, then $\hat{\beta}_G = \tilde{\beta}^* + o_p(1/\sqrt{NT})$ for every $G \ge G^*$.

(c) If, in addition to the conditions of (b), the true partition γ^0 is asymptotically uninformative about the covariates within clusters, in the sense that $\Omega^* := \widehat{W}(\gamma^*) - \widehat{W}(\gamma^0) = o_p(1)$ and $\sqrt{NT}[\Delta(\gamma^*) - \Delta(\gamma^0)] = o_p(1)$, where $\Delta(\gamma) := \widehat{W}(\gamma)(\hat{\beta}(\gamma) - \beta^0)$ denotes the full score of the residualized regression (approximation and stochastic terms in (B.3)*), then also $\hat{\beta}_G = \tilde{\beta} + o_p(1/\sqrt{NT})$, where $\tilde{\beta} = \hat{\beta}(\gamma^0)$ is the oracle estimator of* (3)*.*

Proposition 1, which formalizes the "extremely weak" regime of footnote 3, shows that groups whose effects differ by less than the sampling error, $a_{NT}^2 = o((NT)^{-1})$, cannot be told apart by the GFE estimator—part (a) says nothing about the composition of the estimated groups within clusters—, but they need not be. The estimator is asymptotically equivalent to the least-squares estimator that pools them, which is $\sqrt{NT}$-consistent because the unabsorbed heterogeneity within clusters is of smaller order than the sampling error. Part (a) requires only that the within-cluster distances vanish, at any rate. Part (c) makes precise when pooling is innocuous relative to the oracle that knows the true groups: the two oracles differ at the order $(NT)^{-1/2}$ whenever the covariates carry structure along the true groups within a cluster ($\Omega^* \not\to 0$), in which case $\tilde{\beta}^*$ has the smaller asymptotic variance, since it uses that covariate variation to estimate β^0 without incurring a bias of the order of the sampling error.

Proof. Write $a_i := (\alpha^0_{g_i^0 t} - \alpha^*_{k_i t})_{t \in [T]}$, so that $u_i = v_i + a_i$ and the model reads $y_{it} = x_{it}'\beta^0 + \alpha^*_{k_i t} + u_{it}$ with the cluster partition as group structure. Since α_k^* is a convex combination of $\{\alpha_g^0\}_{g \in K_k}$, $\|a_i\|_T \le a_{NT}$ for every i, and $\|\alpha_k^* - \alpha_{\tilde{k}}^*\|_T \ge \min_{g \in K_k, \tilde{g} \in K_{\tilde{k}}} \|\alpha_g^0 - \alpha_{\tilde{g}}^0\|_T - 2a_{NT}$, so that the clusters satisfy Assumption 2(a)–(b) eventually.

(a) The GFE objective is the same function of (b, a, γ) under both representations, only its decomposition into signal and error changes. The proof of Lemma A.2 applied to the cluster representation uses the errors only through the consistency of $\hat{\beta}_G$, Lemma A.1, and the uniform bounds (ii) in that proof. The first two are statements of the form $o_p(1)$ obtained in BM (Theorem 1 and Lemma B.3) from uni-

form approximations of the objective by its deterministic part. With u_i in place of v_i, every additional term in those approximations is a sample cross-product of a_{it} with a quantity whose root-mean-square over (i,t) is bounded (by Assumptions 1(a),(b) and 4(b)), hence is bounded uniformly over the parameter space by $Ca_{NT} = o(1)$ by the Cauchy–Schwarz inequality, and the approximations remain valid. For the uniform bounds, $\|u_i\|_T \le \|v_i\|_T + a_{NT}$, $|\langle u_i, u_j\rangle_T| \le |\langle v_i, v_j\rangle_T| + 2\max_k \|v_k\|_T a_{NT} + a_{NT}^2$, and $|\langle \alpha_k^* - \alpha_{\tilde{k}}^*, u_i\rangle_T| \le \max_{g\neq\tilde{g}} |\langle \alpha_g^0 - \alpha_{\tilde{g}}^0, v_i\rangle_T| + Ca_{NT}$, because $\alpha_k^* - \alpha_{\tilde{k}}^*$ is a convex combination of differences $\alpha_g^0 - \alpha_{\tilde{g}}^0$ with $\|\alpha_g^0 - \alpha_{\tilde{g}}^0\|_T \le C$. Thus, by Lemma B.1 and $a_{NT} = o(1)$, the bounds (ii) hold with u in place of v. The proof of Lemma A.2 then goes through verbatim and gives (a).

(b) Let E^* be the event in (a), which by definition has probability tending to one. Note that on E^*, $\hat{\gamma}^{(G^*)} = \gamma^*$ up to relabeling, so that $\hat{\beta}_{G^*} = \tilde{\beta}^*$, and $\hat{\gamma}^{(G)}$ refines γ^*. In the cluster representation, the approximation term in (B.3) vanishes on E^* exactly as in step (b) of the proof of Lemma A.3, and steps (a) and (c)–(e) of that proof are algebraic identities and inequalities in which the errors enter only through u_{it}, together with three probabilistic inputs: the uniform invertibility of $\widehat{W}(\gamma)$, which follows from Assumption 1(g) as in Lemma B.5, since $\widehat{W}(\gamma^* \wedge \gamma) \succeq \widehat{W}(\gamma^0 \wedge \gamma)$ for every γ (γ^0 refines γ^*, so residualizing on $\gamma^0 \wedge \gamma$ removes at least as much variation); the rate $\sqrt{NT}(\tilde{\beta}^* - \beta^0) = O_p(1)$; and Assumption 5 for (γ^*, u). All three hold by assumption, hence $\hat{\beta}_G = \tilde{\beta}^* + o_p(1/\sqrt{NT})$. The condition $a_{NT} = o(1/\sqrt{NT})$ is what makes the rate condition on $\tilde{\beta}^*$ attainable: by (B.3) at $\gamma = \gamma^*$ in the original representation, $\widehat{W}(\gamma^*)(\tilde{\beta}^* - \beta^0)$ is the sum of the oracle-type score $(NT)^{-1}\sum_{i,t} \widetilde{x}_{it}(\gamma^*)\widetilde{v}_{it}(\gamma^*)$, which is $O_p((NT)^{-1/2})$ under the conditions of Theorem 3 in BM, and of the approximation term $(NT)^{-1}\sum_{i,t} \widetilde{x}_{it}(\gamma^*)a_{it}$, whose norm is at most $\mathrm{tr}(\widehat{W}(\gamma^*))^{1/2} a_{NT} = O_p(a_{NT}) = o_p((NT)^{-1/2})$ by the Cauchy–Schwarz inequality.

(c) The identity (B.8), applied in the original representation to the fixed refinement γ^0 of γ^*, gives $\tilde{\beta} - \tilde{\beta}^* = \widehat{W}(\gamma^0)^{-1}[\Omega^*(\tilde{\beta}^* - \beta^0) - \delta^*]$ with $\delta^* := \Delta(\gamma^*) - \Delta(\gamma^0)$. The identity is algebraic and holds with the full scores, $\Delta(\gamma^*)$ including the approximation term of the previous paragraph and $\Delta(\gamma^0)$ reducing to the stochastic term. Under the conditions of (c), $\sqrt{NT}(\tilde{\beta} - \tilde{\beta}^*) = o_p(1)$, and the claim follows from (b). □

Proof of Lemma A.4. Let us start by defining

$$\hat{Q}(\beta, \alpha, \gamma) := \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - x_{it}'\beta - \alpha_{g_i t})^2,$$
$$\tilde{Q}(\beta, \alpha, \gamma) := \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}'(\beta^0 - \beta) + \alpha_{g_i^0 t}^0 - \alpha_{g_i t})^2 + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} v_{it}^2.$$

Moreover, write $e := \beta^0 - \hat{\beta}$, $a_{it} := \alpha_{g_i^0 t}^0 - \hat{\alpha}_{\hat{g}_i t}$, $q^2 := (NT)^{-1}\sum_{i,t}(x_{it}'e + a_{it})^2 = \tilde{Q}(\hat{\beta}, \hat{\alpha}, \hat{\gamma}) - \tilde{Q}(\beta^0, \alpha^0, \gamma^0)$ and $\varepsilon_a^2 := (NT)^{-1}\sum_{i,t} a_{it}^2$. Expanding the square in $\hat{Q}$ gives, for every (β, α, γ), the identity

$$\tilde{Q}(\beta, \alpha, \gamma) - \hat{Q}(\beta, \alpha, \gamma) = -\frac{2}{NT}\sum_{i,t}[x_{it}'(\beta^0 - \beta) + (\alpha_{g_i^0 t}^0 - \alpha_{g_i t})]v_{it},$$

so that $\tilde{Q} - \hat{Q}$ vanishes at $(\beta^0, \alpha^0, \gamma^0)$ and the definition of the GFE estimator gives

$$q^2 = \tilde{Q}(\hat{\beta}, \hat{\alpha}, \hat{\gamma}) - \tilde{Q}(\beta^0, \alpha^0, \gamma^0) \le [\tilde{Q} - \hat{Q}](\hat{\beta}, \hat{\alpha}, \hat{\gamma}) = -\frac{2}{NT}\sum_{i,t} v_{it}\,(x_{it}'e + a_{it})$$
$$\le 2\|e\|\Big\|\frac{1}{NT}\sum_{i,t} v_{it}x_{it}\Big\| + \frac{2}{NT}\Big|\sum_{i,t} v_{it}a_{it}\Big|.$$

By the Cauchy–Schwarz inequality,

$$\Big\|\frac{1}{NT}\sum_{i,t} v_{it}x_{it}\Big\|^2 \le \frac{1}{N}\sum_{i=1}^{N}\Big\|\frac{1}{T}\sum_{t=1}^{T} v_{it}x_{it}\Big\|^2,$$

whose expectation is at most M/T by Assumption 1(d), and Markov's inequality implies that the first term is $\|e\|O_p(T^{-1/2})$. For the second, note that a_{it} depends on i only through $(g_i^0, \hat{g}_i)$, so that, writing $B_{g\tilde{g}} := \mathcal{C}_g^0 \cap \hat{\mathcal{C}}_{\tilde{g}}$ for the (at most G^2) nonempty cells of $\gamma^0 \wedge \hat{\gamma}$ and $a_{g\tilde{g}t}$ for the common value of a_{it} on $B_{g\tilde{g}}$, the Cauchy–Schwarz inequality and Lemma B.2 give

$$\Big|\sum_{i,t} v_{it}a_{it}\Big| = \Big|\sum_{g,\tilde{g}}\sum_{t} a_{g\tilde{g}t}\sum_{i\in B_{g\tilde{g}}} v_{it}\Big| \le \left(\sum_{g,\tilde{g}}|B_{g\tilde{g}}|\sum_{t} a_{g\tilde{g}t}^2\right)^{1/2}\left(\sum_{g,\tilde{g}}\frac{\|\sum_{i\in B_{g\tilde{g}}} v_i\|^2}{|B_{g\tilde{g}}|}\right)^{1/2}$$
$$\le \sqrt{NT}\,\varepsilon_a\, G\|V\|_{op}.$$

Hence $q^2 \le 2\|e\|O_p(T^{-1/2}) + 2G\varepsilon_a(\|V\|_{op}^2/NT)^{1/2}$.

On the other hand, since a_{it} is constant on the cells of $\gamma^0 \wedge \hat{\gamma}$ at each t, the argument of Lemma A.2 in BM (projecting the covariates on the cell-time indicators) and Assumption 1(g) give $q^2 \ge \hat{\rho}(\hat{\gamma})\|e\|^2 \ge (\rho/2)\|e\|^2$ with probability approaching one (the last inequality holds uniformly over Γ_G, as in the proof of Lemma B.5), so $\|e\| \le q\sqrt{2/\rho}$. Moreover, by the triangle inequality in $\mathbb{R}^{NT}$ and Markov's inequality applied to $(NT)^{-1}\sum_{i,t}\|x_{it}\|^2$ (Assumption 1(b)),

$$\varepsilon_a = (NT)^{-1/2}\|a\| \le (NT)^{-1/2}\,\|x'e + a\| + (NT)^{-1/2}\,\|x'e\|$$
$$\le q + \|e\|\left((NT)^{-1}\sum_{i,t}\|x_{it}\|^2\right)^{1/2} \le C_1 q$$

with probability approaching one. Substituting, $q^2 \le C_2\, q\,[O_p(T^{-1/2}) + (\|V\|_{op}^2/NT)^{1/2}]$, i.e.,

$$q^2 = O_p\left(T^{-1} + \frac{\|V\|_{op}^2}{NT}\right) = O_p\left(N^{-1/2} + T^{-1/2} + \frac{\max_i\sum_t v_{it}^2}{NT}\right) = O_p\left((N\wedge T)^{-1/2}\right),$$

by Lemma B.2 and because $\max_i T^{-1}\sum_t v_{it}^2 = O_p(1)$, since by Assumption 3(c), Markov's inequality and Theorem 1 in Yokoyama (1980), $\mathbb{P}(T^{-1}\sum_t(v_{it}^2 - \mathbb{E}[v_{it}^2]) > 1) \le CT^{-r/2}$ uniformly in i, and $NT^{-r/2} \to 0$.

Therefore $\|\hat{\beta} - \beta^0\|^2 \le (2/\rho)q^2 = O_p((N\wedge T)^{-1/2})$ and $\varepsilon_a^2 = (NT)^{-1}\sum_{i,t}(\hat{\alpha}_{\hat{g}_i t} - \alpha^0_{g_i^0 t})^2 = O_p((N\wedge T)^{-1/2})$. Finally, for any $g \in [G]$,

$$\frac{1}{NT}\sum_{i=1}^{N}\mathbb{1}\{g_i^0 = g\}\left(\min_{\tilde{g}\in[G]}\sum_{t=1}^{T}\left(\hat{\alpha}_{\tilde{g}t} - \alpha_{gt}^0\right)^2\right) \le \frac{1}{NT}\sum_{i=1}^{N}\mathbb{1}\{g_i^0 = g\}\left(\sum_{t=1}^{T}\left(\hat{\alpha}_{\hat{g}_i t} - \alpha_{gt}^0\right)^2\right) \le \varepsilon_a^2,$$

and the result follows from Assumption 3(a). □

Proof of Lemma A.5. Let $\sigma(g) = \arg\min_{\tilde{g}\in[G]} \left\|\hat{\alpha}_{\tilde{g}} - \alpha_g^0\right\|^2$. By Lemma A.4,

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\hat{\alpha}_{\sigma(g)t} - \alpha_{gt}^0\right)^2 = O_p\left((N\wedge T)^{-1/2}\right).$$

It remains to establish that σ is a permutation. For $g \neq \tilde{g}$, the triangular inequality gives

$$\left\|\hat{\alpha}_{\sigma(g)} - \hat{\alpha}_{\sigma(\tilde{g})}\right\| \geq \left\|\alpha_g^0 - \alpha_{\tilde{g}}^0\right\| - \left\|\alpha_g^0 - \hat{\alpha}_{\sigma(g)}\right\| - \left\|\alpha_{\tilde{g}}^0 - \hat{\alpha}_{\sigma(\tilde{g})}\right\|.$$

Since $\|\alpha_g^0 - \alpha_{\tilde{g}}^0\| \geq \sqrt{Tc_T}$ by Assumption 3(b) and the two remaining terms are $O_p(T^{1/2}(N\wedge T)^{-1/4})$ by Lemma A.4, while $\sqrt{c_T} \geq (N\wedge T)^{-1/8+e/2} \gg (N\wedge T)^{-1/4}$, the RHS is strictly positive with probability approaching one. It thus follows that, with probability approaching 1, $\sigma(g) \neq \sigma(\tilde{g})$ for all $g \neq \tilde{g}$. Thus σ possesses a well-defined inverse and is bijective; i.e., it is a permutation. □

Proof of Lemma A.6. I prove the equivalent statement

$$\max_{i\in[N]}\sup_{(\beta,\alpha)\in\mathcal{N}_\eta} \mathbb{1}\left\{\hat{g}_i(\beta,\alpha) \neq g_i^0\right\} = o_p(1).$$

Note that, from the definition of $\hat{g}_i(\cdot)$,

$$\begin{aligned}\mathbb{1}\left\{\hat{g}_i(\beta,\alpha) \neq g_i^0\right\} &= \max_{g\neq g_i^0}\mathbb{1}\left\{\sum_{t=1}^{T}\left(y_{it} - x_{it}'\beta - \alpha_{gt}\right)^2 < \sum_{t=1}^{T}\left(y_{it} - x_{it}'\beta - \alpha_{g_i^0 t}\right)^2\right\}\\ &= \max_{g\neq g_i^0}\mathbb{1}\left\{\sum_{t=1}^{T}\left(\alpha_{g_i^0 t} - \alpha_{gt}\right)\left(v_{it} + x_{it}'(\beta^0 - \beta) + \alpha_{g_i^0 t}^0 - \frac{\alpha_{g_i^0 t} + \alpha_{gt}}{2}\right) \leq 0\right\}.\end{aligned}$$

Define

$$\begin{aligned}A_T = \Bigg|&\sum_{t=1}^{T}\left(\alpha_{g_i^0 t} - \alpha_{gt}\right)\left(v_{it} + x_{it}'(\beta^0 - \beta) + \alpha_{g_i^0 t}^0 - \frac{\alpha_{g_i^0 t} + \alpha_{gt}}{2}\right)\\ &- \sum_{t=1}^{T}\left(\alpha_{g_i^0 t}^0 - \alpha_{gt}^0\right)\left(v_{it} + \alpha_{g_i^0 t}^0 - \frac{\alpha_{g_i^0 t}^0 + \alpha_{gt}^0}{2}\right)\Bigg|.\end{aligned}$$

By the same arguments as in the proof of Lemma B.4 in BM,

$$A_T \leq TC\sqrt{\eta}\left[\left(\frac{1}{T}\sum_{t=1}^{T}v_{it}^2\right)^{1/2} + \frac{1}{T}\sum_{t=1}^{T}\|x_{it}\| + 1\right].$$

Thus,

$$\begin{aligned}\sup_{(\beta,\alpha)\in\mathcal{N}_\eta}\mathbb{1}\left\{\hat{g}_i(\beta,\alpha) \neq g_i^0\right\} \leq \max_{g\neq g_i^0}\mathbb{1}\Bigg\{&\sum_{t=1}^{T}\left(\alpha_{g_i^0 t}^0 - \alpha_{gt}^0\right)v_{it} \leq -\frac{1}{2}\sum_{t=1}^{T}\left(\alpha_{g_i^0 t}^0 - \alpha_{gt}^0\right)^2\\ &+ TC\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}v_{it}^2\right)^{1/2} + TC\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|\right) + TC\sqrt{\eta}\Bigg\}.\end{aligned}$$

As a result, for c_T in Assumption 3(b),

$$\begin{aligned}\mathbb{P}\left(\sup_{(\beta,\alpha)\in\mathcal{N}_\eta}\mathbb{1}\left\{\hat{g}_i(\beta,\alpha)\neq g_i^0\right\}=1\right)&\leq\sum_{g\neq g_i^0}\mathbb{P}\Bigg(\sum_{t=1}^T\left(\alpha_{g_i^0t}^0-\alpha_{gt}^0\right)v_{it}\leq-\frac{1}{2}\sum_{t=1}^T\left(\alpha_{g_i^0t}^0-\alpha_{gt}^0\right)^2\\&\qquad+TC\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^Tv_{it}^2\right)^{1/2}\\&\qquad+TC\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^T\|x_{it}\|\right)+TC\sqrt{\eta}\Bigg)\\&\leq\sum_{g\neq g_i^0}\Bigg\{\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^T\|x_{it}\|\geq\tilde{M}\right)\\&\qquad+\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^Tv_{it}^2\geq\tilde{M}\right)+\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^T\left(\alpha_{g_i^0t}^0-\alpha_{gt}^0\right)^2\leq\frac{c_T}{2}\right)\\&\qquad+\mathbb{P}\Bigg(\frac{1}{T}\sum_{t=1}^T\left(\alpha_{g_i^0t}^0-\alpha_{gt}^0\right)v_{it}\leq-\frac{c_T}{4}+C\sqrt{\eta\tilde{M}}\\&\qquad+C\sqrt{\eta}\tilde{M}+C\sqrt{\eta}\Bigg)\Bigg\}.\end{aligned}\tag{B.15}$$

Recall that $\sqrt{\eta}=\sqrt{\eta_{NT}}=(N\wedge T)^{-1/4+f}$ with $0<f<e$, so that $\sqrt{\eta}/c_T\leq(N\wedge T)^{f-e}=o(1)$, and suppose $\tilde{M}$ is large enough. The first term on the RHS of (B.15) is $O\left(T^{-r/2}\right)$ by Assumption 3(d). By Markov's inequality,

$$\mathbb{P}\left(\frac{1}{T}\left|\sum_{t=1}^T\left(v_{it}^2-\mathbb{E}\left[v_{it}^2\right]\right)\right|>C\right)\leq\frac{\mathbb{E}\left[\left|\sum_{t=1}^T(v_{it}^2-\mathbb{E}[v_{it}^2])\right|^r\right]}{(CT)^r}.$$

Under Assumption 3(c), we can apply Theorem 1 in Yokoyama (1980) to obtain

$$\mathbb{E}\left[\left|\sum_{t=1}^T(v_{it}^2-\mathbb{E}[v_{it}^2])\right|^r\right]\leq CT^{r/2}.$$

Noting that $T^{-1}\sum_{t=1}^T\mathbb{E}[v_{it}^2]$ is bounded by Assumption 1(c), the second term on the RHS of (B.15) is $O(T^{-r/2})$, uniformly in i. The third term on the RHS of (B.15) is zero by Assumption 3(b) for all g. For the last term, since $\sqrt{\eta}=o(c_T)$, for N,T large enough $C\sqrt{\eta\tilde{M}}+C\sqrt{\eta}\tilde{M}+C\sqrt{\eta}\leq c_T/8$, so that, uniformly over g,

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^T\left(\alpha_{g_i^0t}^0-\alpha_{gt}^0\right)v_{it}\leq-\frac{c_T}{4}+C\sqrt{\eta\tilde{M}}+C\sqrt{\eta}\tilde{M}+C\sqrt{\eta}\right)\\&\quad\leq\mathbb{P}\left(\frac{1}{T}\left|\sum_{t=1}^T\left(\alpha_{g_i^0t}^0-\alpha_{gt}^0\right)v_{it}\right|\geq\frac{c_T}{8}\right)\\&\quad\leq\frac{8^r\,\mathbb{E}[|\sum_{t=1}^T(\alpha_{g_i^0t}^0-\alpha_{gt}^0)v_{it}|^r]}{(Tc_T)^r}\leq CT^{-r/2}c_T^{-r},\end{aligned}$$

by Markov's inequality and Theorem 1 in Yokoyama (1980) applied to the strictly stationary, mixing, mean-zero process $\{(\alpha^0_{g_i^0 t} - \alpha^0_{gt})v_{it}\}_t$ of Assumption 3(c). It then follows that

$$\mathbb{P}\left(\max_{i\in[N]} \sup_{(\beta,\alpha)\in\mathcal{N}_\eta} \mathbb{1}\left\{\hat{g}_i(\beta,\alpha) \neq g_i^0\right\} = 1\right) \leq \sum_{i=1}^{N} \mathbb{P}\left(\sup_{(\beta,\alpha)\in\mathcal{N}_\eta} \mathbb{1}\left\{\hat{g}_i(\beta,\alpha) \neq g_i^0\right\} = 1\right)$$
$$= O\left(NT^{-r/2}\left(1 + c_T^{-r}\right)\right),$$

which proves the result because $NT^{-r/2}c_T^{-r} = o(1)$ and c_T is bounded. Note that Assumption 3(b) gives $c_T^{-r} \leq (N \wedge T)^{r/4-re}$, so the condition $NT^{-r/2}(N \wedge T)^{r/4-re} = o(1)$ is sufficient. □

The following three lemmas collect the auxiliary results used in the proofs of Lemmas A.2, A.4 and B.4. Recall that $\|a\|_T^2 = T^{-1}\sum_t a_t^2$ and $\langle a, b\rangle_T = T^{-1}\sum_t a_t b_t$ for $a, b \in \mathbb{R}^T$, and let $v_i := (v_{it})_{t\in[T]}$.

Lemma B.1. *Suppose Assumptions 1 and 2 hold and $N, T \to \infty$ with $NT^{-\nu} \to 0$ for some $\nu > 0$. Then:*

(a) There is a constant C_v such that $\mathbb{P}(\max_{i\in[N]} \|v_i\|_T^2 \leq C_v) \to 1$;

(b) If, in addition, Assumption 4(a) holds, $\max_{i\neq j} |\langle v_i, v_j\rangle_T| = o_p(1)$;

(c) $\max_{i\in[N]} \max_{g\neq\tilde{g}} |\langle \alpha_g^0 - \alpha_{\tilde{g}}^0, v_i\rangle_T| = o_p(1)$;

(d) If, in addition, Assumption 4(b) holds, $\mathbb{P}(\max_{i\in[N]} T^{-1}\sum_{t=1}^T \|x_{it}\|^2 \leq M_2) \to 1$.

Proof of Lemma B.1. Parts (a) and (c) are established in the proof of Lemma B.4 in BM: under Assumptions 2(c)–(d), the processes $\{v_{it}^2 - \mathbb{E}[v_{it}^2]\}_t$ and $\{(\alpha_{gt}^0 - \alpha_{\tilde{g}t}^0)v_{it}\}_t$ are strongly mixing with coefficients decaying faster than any polynomial, have mean zero, and have tails decaying faster than any polynomial uniformly in i (for the first, because $\mathbb{P}(v_{it}^2 > m) \leq \exp(1 - (\sqrt{m}/b)^{d_2})$; for the second, because $|\alpha_{gt}^0 - \alpha_{\tilde{g}t}^0|$ is bounded by Assumption 1(a)). The Bernstein-type inequality of Merlevède, Peligrad and Rio (2011, Theorem 2) (applied after replacing d_1 and d_2 by $\min\{d_1, 1\}$ and $\min\{d_2, 1\}$, which weakens Assumption 2(c)–(d) and ensures the condition $1/\gamma_1 + 1/\gamma_2 > 1$ of that theorem, and after absorbing the exponent $d_2/2$ of the tails of v_{it}^2 and $v_{it}v_{jt}$ into the constants) then gives, for every $\varepsilon > 0$, constants $c_1, c_2, d_3 > 0$ not depending on i such that $\mathbb{P}(|T^{-1}\sum_t(v_{it}^2 - \mathbb{E}[v_{it}^2])| > \varepsilon) \leq c_1\exp(-c_2T^{d_3})$ and $\mathbb{P}(|\langle \alpha_g^0 - \alpha_{\tilde{g}}^0, v_i\rangle_T| > \varepsilon) \leq c_1\exp(-c_2T^{d_3})$. A union bound over $i \in [N]$ (and over the finitely many pairs $(g, \tilde{g})$) yields probabilities of order $N\exp(-c_2T^{d_3}) = o(1)$, since $N \leq T^\nu$ eventually. Part (a) follows with $C_v := M^{1/2} + 1$ because $T^{-1}\sum_t \mathbb{E}[v_{it}^2] \leq M^{1/2}$ by Assumption 1(c). For (b), fix $i \neq j$. Under Assumption 4(a) the processes $\{v_{it}\}_t$ and $\{v_{jt}\}_t$ are independent, so $\mathbb{E}[v_{it}v_{jt}] = 0$ and the process $\{v_{it}v_{jt}\}_t$ is strongly mixing with coefficients bounded by $2\alpha[t]$ (Bradley, 2005, Theorem 5.1), whose tails satisfy $\mathbb{P}(|v_{it}v_{jt}| > m) \leq 2\exp(1-(\sqrt{m}/b)^{d_2})$. The same inequality therefore gives $\mathbb{P}(|\langle v_i, v_j\rangle_T| > \varepsilon) \leq c_1\exp(-c_2T^{d_3})$ uniformly over pairs, and a union bound over the $N^2 \leq T^{2\nu}$ pairs yields (b). Part (d) follows from Assumption 4(b) and a union bound, since $N\,o(T^{-\delta}) = o(1)$ for $\delta > \nu$. □

Lemma B.2. *Let V denote the $N \times T$ matrix (v_{it}) and $\|\cdot\|_{op}$ the operator norm. Under Assumption 1(c), (e) and (f),*

$$\|VV' - \mathrm{diag}(VV')\|_F = O_p\left(N^{1/2}T + NT^{1/2}\right),$$

hence

$$\|V\|_{op}^2 \le O_p\left(N^{1/2}T + NT^{1/2}\right) + \max_{i\in[N]} \sum_{t=1}^{T} v_{it}^2.$$

Moreover, for every partition of $[N]$ into cells S and every S, $\|\sum_{i\in S} v_i\|^2/|S| \le \|V\|_{op}^2$, and the same holds with v_i replaced by $v_i - \bar{v}_{C^0_{g_i^0}}$, the within-true-group deviation.

Proof of Lemma B.2. By the triangle inequality, $\|VV'\|_{op} \le \|VV' - \mathrm{diag}(VV')\|_{op} + \|\mathrm{diag}(VV')\|_{op} \le \|VV' - \mathrm{diag}(VV')\|_F + \max_i \sum_t v_{it}^2$, and $\|V\|_{op}^2 = \|VV'\|_{op}$. Next,

$$\mathbb{E}\,\|VV' - \mathrm{diag}(VV')\|_F^2 = \sum_{i\neq j} \mathbb{E}\left[\left(\sum_{t=1}^{T} v_{it}v_{jt}\right)^2\right] = \sum_{i\neq j}\left(\sum_{t=1}^{T}\mathbb{E}[v_{it}v_{jt}]\right)^2 + \sum_{i\neq j}\mathrm{Var}\left(\sum_{t=1}^{T} v_{it}v_{jt}\right).$$

By Assumption 1(c) and the Cauchy–Schwarz inequality, $|\mathbb{E}[v_{it}v_{jt}]| \le M^{1/2}$, so

$$\left(\sum_t \mathbb{E}[v_{it}v_{jt}]\right)^2 \le M^{1/2}T\left|\sum_t \mathbb{E}[v_{it}v_{jt}]\right|,$$

and the first sum is at most $M^{1/2}T\sum_{i,j}|\sum_t \mathbb{E}[v_{it}v_{jt}]|$, which is bounded by $M^{3/2}NT^2$ by Assumption 1(e). The second sum consists of nonnegative terms and is bounded by

$$\sum_{i,j}\sum_{t,s}\mathrm{Cov}(v_{it}v_{jt}, v_{is}v_{js}) \le MN^2T$$

by Assumption 1(f). Markov's inequality then gives the first claim. For the last claim,

$$\left\|\sum_{i\in S} v_i\right\|^2 |S|^{-1} = \|\mathbb{1}_S'V\|^2\|\mathbb{1}_S\|^{-2} \le \|V\|_{op}^2;$$

and the matrix of within-true-group deviations is $(\mathrm{I}_N - \Pi_{\gamma^0})V$ for the orthogonal projection Π_{γ^0} onto the indicators of the true groups, whose operator norm is at most $\|V\|_{op}$. □

Lemma B.3. *Let $r_1, \dots, r_N \in \mathbb{R}^T$ and, for a nonempty $S \subseteq [N]$, let $\bar{r}_S$ be the average of r_i over S and $\mathrm{SS}(S) := \sum_{i\in S}\|r_i - \bar{r}_S\|^2$.*

(a) (Exchange.) If S and S' are disjoint and nonempty, $|S| = m \ge 2$, $|S'| = m'$, and $i \in S$, then

$$\mathrm{SS}(S\setminus\{i\}) + \mathrm{SS}(S'\cup\{i\}) - \mathrm{SS}(S) - \mathrm{SS}(S') = \frac{m'}{m'+1}\|r_i - \bar{r}_{S'}\|^2 - \frac{m}{m-1}\|r_i - \bar{r}_S\|^2.$$

(b) (Optimality of the mean.) For every $a \in \mathbb{R}^T$, $\mathrm{SS}(S) \le \sum_{i\in S}\|r_i - a\|^2$.

(c) (ANOVA.) $\mathrm{SS}(S) = \sum_{i\in S}\|r_i\|^2 - |S|\|\bar{r}_S\|^2$.

Proof of Lemma B.3. (b) and (c) follow from $\sum_{i\in S}\|r_i - a\|^2 = \mathrm{SS}(S) + |S|\|\bar{r}_S - a\|^2$, which is obtained by expanding the square, with $a = 0$ for (c). For (a), applying this identity to the set $A\cup\{i\}$ with $|A| = k \ge 1$ and $a = \bar{r}_A$, and noting that $\bar{r}_{A\cup\{i\}} - \bar{r}_A = (r_i - \bar{r}_A)/(k+1)$, gives $\mathrm{SS}(A\cup\{i\}) = \mathrm{SS}(A) + \frac{k}{k+1}\|r_i - \bar{r}_A\|^2$.

Applying it with $A = S'$ gives the first term. Applying it with $A = S \setminus \{i\}$, $k = m-1$, and using $r_i - \bar{r}_S = \frac{m-1}{m}(r_i - \bar{r}_{S\setminus\{i\}})$, gives $\mathrm{SS}(S) = \mathrm{SS}(S\setminus\{i\}) + \frac{m}{m-1}\|r_i - \bar{r}_S\|^2$, which is the second term. □

The next lemma verifies the conclusion of Assumption 5 at $\gamma = \tilde{\gamma}$ under primitive conditions. It formalizes the variance-shrinking mechanism invoked in the Introduction, and shows that it is valid precisely because $\tilde{\gamma}$ is a function of the errors and the true groups only.

Lemma B.4. *Suppose Assumptions 1 and 2 hold, $G \geq G^0$, and in addition:*

(a) $\{v_{it}\}_{i,t}$ is independent of $(\{x_{it}\}_{i,t}, \{g_i^0\}_i)$;

(b) conditional on $\{g_i^0\}_i$, the within-group deviations $\widetilde{x}_{it} := x_{it} - \bar{x}_{g_i^0 t}$ satisfy, for some constant C, for every $g \in [G^0]$, every nonempty $S \subseteq \mathcal{C}_g^0$, every coordinate $k \in [d]$ and every $u \in \mathbb{R}^T$ with $\|u\| = 1$,

$$\mathbb{E}\left[\left(\sum_{t=1}^{T} u_t \sum_{i\in S} \widetilde{x}_{it}^{(k)}\right)^2 \,\middle|\, \{g_i^0\}_i\right] \leq C|S|.$$

Then $\|\Omega(\tilde{\gamma})\| = O_p(N^{-1})$ and $\sqrt{NT}\|\delta(\tilde{\gamma})\| = O_p\left(N^{-1/4} + T^{-1/4}\right)$. In particular, the conclusion of Assumption 5 holds for the element $\tilde{\gamma}$ of $\mathcal{N}_{c,K}$.

Condition (a) is strict exogeneity in its strong form, and it excludes predetermined regressors such as lagged outcomes. Condition (b) requires the within-group deviations of the covariates to have zero mean and to be weakly dependent across units and over time once the group structure is conditioned upon: the variance of any linear combination of the deviations of $|S|$ units grows at most linearly in $|S|$, uniformly over directions in $\mathbb{R}^T$. It holds, for instance, when within each true group the covariate paths are exchangeable across units—as when units are sampled at random from a group population—and their idiosyncratic components have summable autocovariances. It accommodates common (group-level) components, which are removed by the demeaning, and latent structure with random loadings, e.g., $x_{it} = f_t + \kappa_i h_t + \varepsilon_{it}$ with κ_i i.i.d. across i and $\{h_t\}$ a stationary process with summable autocovariances independent of $\{\kappa_i\}$. It rules out deterministic unit-specific heterogeneity in the covariates within a true group, since then $\sum_{i\in S}\mathbb{E}[\widetilde{x}_{it}]$ can be of order $|S|$ for suitably chosen S.

Proof of Lemma B.4. Write $\widetilde{x}_i^{(k)} := (\widetilde{x}_{it}^{(k)})_{t\in[T]} \in \mathbb{R}^T$ and $\widetilde{v}_i := (v_{it} - \bar{v}_{g_i^0 t})_{t\in[T]} \in \mathbb{R}^T$, and for a cell S of a refinement let $a_k(S) := \sum_{i\in S}\widetilde{x}_i^{(k)}$ and $b(S) := \sum_{i\in S}\widetilde{v}_i$, so that $b_t(S) = |S|(\bar{v}_{St} - \bar{v}_{g^0 t})$ and $a_{k,t}(S) = |S|(\bar{x}_{St}^{(k)} - \bar{x}_{g^0 t}^{(k)})$. Then, for any refinement γ,

$$\Omega_{kk}(\gamma) = \frac{1}{NT}\sum_S \frac{\|a_k(S)\|^2}{|S|}, \quad NT\,\delta_k(\gamma) = \sum_S \frac{\langle a_k(S), b(S)\rangle}{|S|}, \quad NT\,D(\gamma) = \sum_S \frac{\|b(S)\|^2}{|S|}. \tag{B.16}$$

The partition $\tilde{\gamma}$ is a function of $(\{v_{it}\}, \{g_i^0\})$ only. Throughout, condition on $\mathcal{F} := \sigma(\{v_{it}\}, \{g_i^0\})$. By condition (a), the conditional distribution of $\{x_{it}\}$ given $\mathcal{F}$ is its conditional distribution given $\{g_i^0\}$, and the cells S of $\tilde{\gamma}$ as well as the vectors $b(S)$ are fixed given $\mathcal{F}$. Hence condition (b) applies to each cell of $\tilde{\gamma}$.

Gram difference. By (B.16) and condition (b) applied with u equal to each canonical basis vector of $\mathbb{R}^T$, $\mathbb{E}[\|a_k(S)\|^2 \mid \mathcal{F}] \leq CT|S|$, so that $\mathbb{E}[\Omega_{kk}(\tilde{\gamma}) \mid \mathcal{F}] \leq CGT/(NT) = CG/N$, because $\tilde{\gamma}$ has at most G cells.

Since $\Omega(\tilde\gamma) \succeq 0$, $\|\Omega(\tilde\gamma)\| \le \operatorname{tr}\Omega(\tilde\gamma)$, and Markov's inequality gives $\|\Omega(\tilde\gamma)\| = O_p(N^{-1})$.

Score difference. Fix k and a cell S of $\tilde\gamma$ with $b(S) \neq 0$, and put $u = b(S)/\|b(S)\|$, which is $\mathcal{F}$-measurable. By condition (b), $\mathbb{E}[\langle a_k(S), b(S)\rangle^2 \mid \mathcal{F}] = \|b(S)\|^2\,\mathbb{E}[(u'a_k(S))^2 \mid \mathcal{F}] \le C|S|\,\|b(S)\|^2$. By the triangle inequality in $L^2(\cdot \mid \mathcal{F})$, (B.16) and the Cauchy–Schwarz inequality over the at most G cells,

$$\begin{aligned}\mathbb{E}\left[(NT\delta_k(\tilde\gamma))^2 \mid \mathcal{F}\right]^{1/2} &\le \sum_S \frac{\sqrt{C|S|}\,\|b(S)\|}{|S|} = \sqrt{C}\sum_S \frac{\|b(S)\|}{\sqrt{|S|}} \\ &\le \sqrt{CG}\left(\sum_S \frac{\|b(S)\|^2}{|S|}\right)^{1/2} = \sqrt{CG\,NT\,D(\tilde\gamma)}.\end{aligned}$$

Summing over $k \in [d]$ and applying Markov's inequality conditionally on $\mathcal{F}$ gives $\sqrt{NT}\|\delta(\tilde\gamma)\| = O_p(D(\tilde\gamma)^{1/2})$. This is the announced mechanism: given the errors, the between-cell means of the covariates are averages of $|S|$ weakly dependent, mean-zero terms, which is what a fixed-partition calculation uses. The calculation is legitimate at $\tilde\gamma$ because $\tilde\gamma$ does not depend on the covariates.

The fit to the errors vanishes. It remains to show $D(\tilde\gamma) = o_p(1)$, with a rate. By (B.16) and Lemma B.2, $NT\,D(\gamma) \le G\|V\|_{op}^2$ for every refinement γ, and $\|V\|_{op}^2 \le O_p(N^{1/2}T + NT^{1/2}) + \max_i \sum_t v_{it}^2$. Since $\max_i \sum_t v_{it}^2 \le \sum_t \max_i v_{it}^2$ and, by the exponential tail bound in Assumption 2(d), $\mathbb{E}[\max_{i\in[N]} v_{it}^2] \le C(1+\log N)^{2/d_2}$ for every t, the last term is $O_p(T(\log N)^{2/d_2})$ by Markov's inequality. Hence

$$D(\tilde\gamma) \le \frac{G\|V\|_{op}^2}{NT} = O_p\left(N^{-1/2} + T^{-1/2} + \frac{(\log N)^{2/d_2}}{N}\right) = O_p\left(N^{-1/2} + T^{-1/2}\right),$$

where the last step uses that $(\log N)^{2/d_2}N^{-1/2} = o(1)$. Combining with the previous paragraph, we obtain $\sqrt{NT}\|\delta(\tilde\gamma)\| = O_p(N^{-1/4} + T^{-1/4})$, and the last claim of the lemma follows from the two rates. □

The last result of this appendix records the uniform invertibility of the profiled Gram matrix $\widehat{W}(\gamma)$ defined in (B.2), which is used repeatedly in Appendices A and B. It is where Assumption 1(g) enters the arguments, and it bridges the residualization on the cells of $\gamma^0 \wedge \gamma$ in that assumption and the residualization on the cells of γ in $\widehat{W}(\gamma)$.

Lemma B.5. *Suppose Assumption 1(g) holds. Then, with probability approaching one, for every $\gamma \in \Gamma_G$, $\lambda_{\min}(\widehat{W}(\gamma)) \ge \hat\rho(\gamma) \ge \rho/2$, so that $\widehat{W}(\gamma)^{-1}$ exists and satisfies $\|\widehat{W}(\gamma)^{-1}\| \le 2/\rho$.*

Proof of Lemma B.5. Let x denote the $NT \times d$ matrix with rows x_{it}', and for a partition γ let $\mathcal{P}_\gamma$ denote the orthogonal projection of $\mathbb{R}^{NT}$ onto the span of the GT cell–time indicators $e_{g,t} \in \mathbb{R}^{NT}$, $g \in [G]$, $t \in [T]$, with entries $(e_{g,t})_{(i,s)} = \mathbb{1}\{g_i = g\}\mathbb{1}\{s = t\}$, so that residualizing on the cells of γ is the map $w \mapsto (\mathrm{I} - \mathcal{P}_\gamma)w$. Then $NT\,\widehat{W}(\gamma) = x'(\mathrm{I} - \mathcal{P}_\gamma)x$, and the matrix in Assumption 1(g) is $(NT)^{-1}x'(\mathrm{I} - \mathcal{P}_{\gamma^0\wedge\gamma})x$. Since every cell of $\gamma^0 \wedge \gamma$ is contained in a cell of γ, the range of $\mathcal{P}_\gamma$ is a subspace of that of $\mathcal{P}_{\gamma^0\wedge\gamma}$, so $\mathcal{P}_{\gamma^0\wedge\gamma} - \mathcal{P}_\gamma$ is itself an orthogonal projection and hence positive semi-definite. Therefore, for every $u \in \mathbb{R}^d$ with $\|u\| = 1$,

$$u'\widehat{W}(\gamma)u = \frac{\|(\mathrm{I} - \mathcal{P}_\gamma)xu\|^2}{NT} \ge \frac{\|(\mathrm{I} - \mathcal{P}_{\gamma^0\wedge\gamma})xu\|^2}{NT} \ge \hat\rho(\gamma),$$

i.e., $\lambda_{\min}(\widehat{W}(\gamma)) \ge \hat\rho(\gamma)$ for every $\gamma \in \Gamma_G$ and every realization of the data. By Assumption 1(g), $\min_{\gamma\in\Gamma_G}\hat\rho(\gamma) \ge \rho/2$ with probability approaching one, which gives the first claim. The second follows

from $\|\widehat{W}(\gamma)^{-1}\| = \lambda_{\min}(\widehat{W}(\gamma))^{-1}$. □

C Additional Simulation Results

This appendix reports the two additional experiments announced in Section 5.1, numerical evidence for the claim of Section 3 on the detectability of under- and overspecification, the similarity measures used in Section 6, and the experiment with heavy-tailed errors announced in Section 5.2. Unless otherwise indicated, the GFE estimator is computed with Algorithm 1 in BM from 10 random starting values (random assignments of the units to groups), and the solution with the smallest objective is retained, as in Section 5. Standard deviations are reported relative to that of the oracle estimator $\tilde{\beta}$ based on the true groups ("sd ratio"), while "mixing" denotes the share of units whose true group is not the majority true group of their estimated group, which Theorem 1(a) predicts to vanish for every $G \geq G^0$ and which, at $G = G^0$, is the usual misclassification rate.

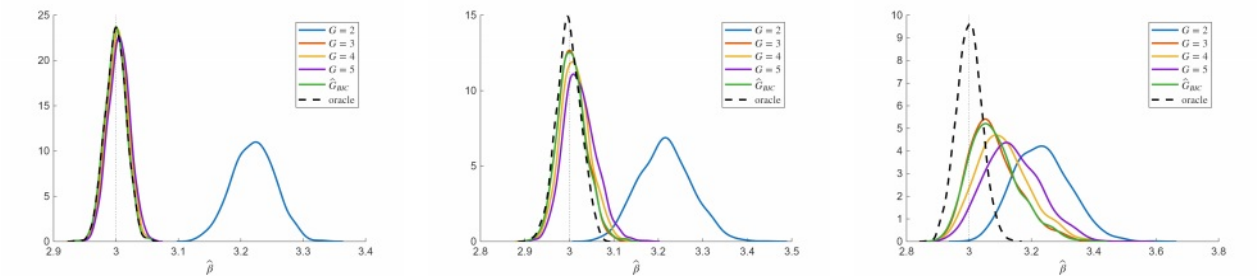
C.1 Short Panels

The BIC is computed according to the recommendation in BM:

$$BIC(G) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - x_{it}'\hat{\beta}_G - \hat{\alpha}_{\hat{g}_i t,G})^2 + \hat{\sigma}^2 \frac{GT+N+K}{NT}\log(NT),$$

where

$$\hat{\sigma}^2 = \frac{1}{NT - G_{max}T - N - K}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - x_{it}'\hat{\beta}_{G_{max}} - \hat{\alpha}_{\hat{g}_i t,G_{max}})^2, \quad G_{max} = G^0 + 2.$$



(a) $(N,T) = (75,50)$ (b) $(N,T) = (75,20)$ (c) $(N,T) = (75,9)$

Figure C.1: Distribution of the GFE estimator for different number of groups, including $G < G^0$ and BIC

Table C.1 reports the design of Section 5.1 for $T \in \{9, 20, 50\}$ with $N = 75$. With $T = 9$, classification is poor: 13% of the units are misclassified at $G = G^0$, and the mixing rate increases to 16–18% for $G = G^0+1, G^0+2$. All estimators with $G \geq G^0$ are then biased—by two oracle standard deviations at $G = G^0$, and by three to three and a half at $G = G^0+1, G^0+2$—and their standard deviations are about twice the oracle's. Overspecification adds a moderate further distortion to the misclassification biases that dominate in such short panels. With $T = 20$, misclassification is about 1-2%, the biases of the $G \geq G^0$ estimators are within one oracle standard deviation, and the standard deviation ratios are 1.14, 1.21 and 1.33 for $G = 3, 4, 5$; with $T = 50$ they are 1.00, 1.01 and 1.03, as in Figure I. The BIC selects $G < G^0$ in less than

1% of the samples in every design and $G > G^0$ in 7% of the samples when $T = 9$ and never otherwise, so that the BIC-selected estimator behaves like the estimator at $G = G^0$.

Table C.1: Overspecification of G in short panels

	Bias	sd	sd ratio	Mixing	$\hat{G}_{BIC}$ freq.
		$T = 9$			
oracle	0.001	0.040	1	–	–
$G = 2$	0.235	0.088	2.20	0.354	0.004
$G = 3$	0.081	0.079	1.96	0.127	0.925
$G = 4$	0.110	0.085	2.12	0.157	0.069
$G = 5$	0.140	0.089	2.23	0.181	0.002
$G = \hat{G}_{BIC}$	0.082	0.079	1.98	–	–
		$T = 20$			
oracle	-0.001	0.027	1	–	–
$G = 2$	0.216	0.059	2.15	0.307	0.008
$G = 3$	0.006	0.031	1.14	0.012	0.992
$G = 4$	0.014	0.033	1.21	0.018	0.000
$G = 5$	0.024	0.036	1.33	0.025	0.000
$G = \hat{G}_{BIC}$	0.007	0.032	1.18	–	–
		$T = 50$			
oracle	-0.000	0.017	1	–	–
$G = 2$	0.221	0.034	2.03	0.286	0.000
$G = 3$	-0.000	0.017	1.00	0.000	1.000
$G = 4$	0.002	0.017	1.01	0.000	0.000
$G = 5$	0.005	0.017	1.03	0.000	0.000
$G = \hat{G}_{BIC}$	-0.000	0.017	1.00	–	–

Notes: Design of Section 5.1, $N = 75$, $G^0 = 3$, $\beta^0 = 3$. "sd ratio" is the ratio of the standard deviation of $\hat{\beta}_G$ to that of the oracle $\tilde{\beta}$; "mixing" is the share of units whose true group is not the majority true group of their estimated group (at $G = G^0$, the misclassification rate); the last column reports the frequency with which the BIC selects each G ($G_{max} = 5$). Based on 1,000 MC replications.

C.2 Latent Group Structure in the Covariates

The second experiment lets the covariates carry a latent binary structure that is independent of the groups of the outcome equation:

$$y_{it} = x_{it}\beta^0 + \alpha^0_{g_i^0 t} + v_{it},$$
$$x_{it} = \alpha^0_{g_i^0 t} + \sigma_h\, \kappa_i\, h_t + u_{it},$$

with $\beta^0 = 1$, $G^0 \in \{1, 2\}$, $\kappa_i \in \{0, 1\}$ drawn with probability $1/2$ independently of g_i^0, $\sigma_h \in \{1, 3\}$, and α^0_{gt}, h_t, u_{it}, v_{it} independent standard normal. Thus, the joint structure of (y_{it}, x_{it}) has $2G^0$ cells. For $G^0 = 1$, this is the example of Remark 1 (with $f_t = \alpha^0_{1t}$). Besides the oracle, the benchmark is the "joint oracle", the least-squares estimator based on the $2G^0$ cells $\{g_i^0\} \times \{\kappa_i\}$, i.e., on the fixed informative refinement of Remark 1. Three sample sizes keep NT roughly constant while N/T grows, $(N, T) \in \{(75, 50), (150, 25), (300, 10)\}$, and two let N and T grow together, $(N, T) \in \{(150, 100), (300, 200)\}$.

The results turn out to depend on how the GFE estimator is computed, and three strategies are compared. Strategy A is Algorithm 1 in BM from 10 random starting values, as everywhere else in the paper: each starting value is a random assignment of the units to G groups, from which the slope and the group effects are obtained by least squares.[11] Strategy B is a "path-following" warm start: for each $G \geq 2$, the algorithm

[11] BM initialize instead with a slope drawn from a prespecified distribution and the group effects of G randomly selected units.

is initialized at the k-means classification (50 replicates) of the residuals $y_{it} - x_{it}'\hat{\beta}_{G-1}$ of the estimate obtained with $G-1$ groups, and iterated to convergence. Strategy C retains, for each G, the one of A and B with the smaller objective. Table C.2 reports, for $G = G^0 + 1$ and $G = G^0 + 3$, the standard deviation ratios of the three strategies, together with two diagnostics of the estimated partition: its κ-imbalance, the size-weighted average over the estimated cells of the absolute deviation of the share of units with $\kappa_i = 1$ from $1/2$, whose expected value for a partition that ignores κ_i is $(2\pi N/G)^{-1/2}$; and the covariate–error covariance $\sqrt{NT}\|\delta(\hat{\gamma})\|$ of Appendix A.

Table C.2: Latent group structure in the covariates: $x_{it} = \alpha^0_{g^0_i t} + \sigma_h \kappa_i h_t + u_{it}$, $\beta^0 = 1$

		sd ratio				κ-imbalance			$\sqrt{NT}\|\delta(\hat{\gamma})\|$		$\|\Omega(\hat{\gamma})\|$			
(G^0, σ_h, N, T)	G	Joint	A	B	C	A	B	Random	A	B	Joint	A	B	$\mathbb{P}(Q_B < Q_A)$
$(1, 1, 75, 50)$	2	1.11	1.02	1.01	1.01	0.065	0.066	0.065	0.18	0.18	0.26	0.02	0.02	0.82
	4		1.06	1.03	1.04	0.090	0.089	0.092	0.27	0.27		0.05	0.05	0.85
$(1, 1, 150, 25)$	2	1.13	1.02	1.02	1.02	0.046	0.045	0.046	0.19	0.19	0.25	0.01	0.01	0.85
	4		1.04	1.04	1.05	0.065	0.065	0.065	0.29	0.29		0.02	0.03	0.82
$(1, 1, 300, 10)$	2	1.12	1.04	1.03	1.03	0.032	0.032	0.033	0.27	0.26	0.25	0.00	0.00	0.80
	4		1.09	1.09	1.10	0.047	0.046	0.046	0.38	0.39		0.01	0.01	0.85
$(1, 3, 75, 50)$	2	1.76	1.02	1.02	1.02	0.065	0.066	0.065	0.27	0.27	2.23	0.04	0.04	0.81
	4		1.05	1.03	1.03	0.090	0.089	0.092	0.42	0.42		0.13	0.13	0.84
$(1, 3, 150, 25)$	2	1.79	1.02	1.02	1.01	0.046	0.045	0.046	0.30	0.28	2.23	0.02	0.02	0.83
	4		1.04	1.05	1.05	0.064	0.065	0.065	0.46	0.47		0.06	0.07	0.82
$(1, 3, 300, 10)$	2	1.74	1.03	1.02	1.03	0.032	0.032	0.033	0.38	0.37	2.24	0.01	0.01	0.81
	4		1.06	1.07	1.07	0.046	0.046	0.046	0.56	0.61		0.03	0.03	0.85
$(2, 1, 75, 50)$	3	1.15	1.03	1.02	1.03	0.089	0.079	0.080	0.23	0.16	0.27	0.02	0.02	0.84
	5		1.05	1.05	1.06	0.118	0.101	0.103	0.50	0.25		0.06	0.05	0.87
$(2, 1, 150, 25)$	3	1.12	1.04	1.03	1.04	0.059	0.055	0.056	0.23	0.16	0.26	0.01	0.01	0.80
	5		1.08	1.08	1.12	0.083	0.072	0.073	0.61	0.26		0.03	0.03	0.73
$(2, 1, 300, 10)$	3	1.14	1.32	1.25	1.27	0.040	0.039	0.040	0.30	0.24	0.25	0.01	0.01	0.75
	5		1.44	1.32	1.37	0.054	0.052	0.052	0.71	0.38		0.03	0.02	0.78
$(2, 3, 75, 50)$	3	1.83	1.05	1.01	1.02	0.101	0.078	0.080	0.45	0.22	2.21	0.11	0.04	0.83
	5		1.14	1.06	1.08	0.149	0.100	0.103	1.01	0.39		0.30	0.13	0.91
$(2, 3, 150, 25)$	3	1.74	1.03	1.03	1.04	0.063	0.055	0.056	0.44	0.23	2.22	0.04	0.02	0.84
	5		1.13	1.06	1.10	0.104	0.072	0.073	1.22	0.39		0.15	0.07	0.79
$(2, 3, 300, 10)$	3	1.74	1.19	1.16	1.17	0.042	0.039	0.040	0.48	0.35	2.16	0.03	0.03	0.76
	5		1.34	1.23	1.28	0.060	0.052	0.052	1.14	0.55		0.08	0.06	0.79
$(1, 3, 150, 100)$	2	1.92	1.01	1.01	1.01	0.045	0.045	0.046	0.19	0.19	2.25	0.02	0.02	0.81
	4		1.03	1.02	1.02	0.064	0.065	0.065	0.31	0.32		0.06	0.07	0.81
$(1, 3, 300, 200)$	2	1.73	1.01	1.01	1.01	0.032	0.032	0.033	0.13	0.13	2.25	0.01	0.01	0.84
	4		1.02	1.00	1.01	0.046	0.045	0.046	0.22	0.22		0.03	0.03	0.80
$(2, 3, 150, 100)$	3	1.87	1.08	1.01	1.02	0.114	0.056	0.056	0.78	0.17	2.22	0.19	0.02	0.84
	5		1.25	1.02	1.07	0.192	0.071	0.073	1.60	0.27		0.47	0.06	0.92
$(2, 3, 300, 200)$	3	1.76	1.18	1.00	1.01	0.156	0.040	0.040	0.76	0.11	2.24	0.36	0.01	0.92
	5		1.45	1.01	1.06	0.295	0.051	0.052	1.64	0.20		0.86	0.03	0.95

Notes: A: Algorithm 1 of BM from 10 random starting values; B: warm start at the k-means classification (50 replicates) of the residuals of the estimate with $G-1$ groups; C: the one of A and B with the smaller objective Q; the last column is the frequency with which B attains the smaller objective. "sd ratio" is the ratio of the standard deviation of the estimator to that of the oracle $\tilde{\beta}$ based on the true groups; "Joint" refers to the joint oracle, the least-squares estimator based on the $2G^0$ cells $\{g_i^0\} \times \{\kappa_i\}$, i.e., on the fixed informative refinement of Remark 1, and to the partition into those cells. "κ-imbalance" is the size-weighted average over the estimated cells of |share of units with $\kappa_i = 1 - 1/2$|; "Random" is its expected value for a partition that ignores κ_i, $(2\pi N/G)^{-1/2}$. $\sqrt{NT}\|\delta(\hat{\gamma})\|$ and $\|\Omega(\hat{\gamma})\|$ are the covariate–error covariance generated and the covariate variation absorbed by the estimated partition, relative to the true partition (Appendix A); for the joint partition, $\|\Omega\|$ is the covariate variation carried by κ_i. No mixing of true groups occurred with $T \geq 25$; with $T = 10$ and $G^0 = 2$ the mixing rate is 3–4%. Based on 1,000 MC replications.

Four findings stand out. First, the joint oracle displays the variance inflation of Remark 1: its standard deviation exceeds the oracle's by 11–15% when $\sigma_h = 1$ and by 73–92% when $\sigma_h = 3$ (first "sd ratio" column), the counterpart of the covariate variation carried by the joint partition ($\|\Omega\|$ of 0.25 and 2.2,

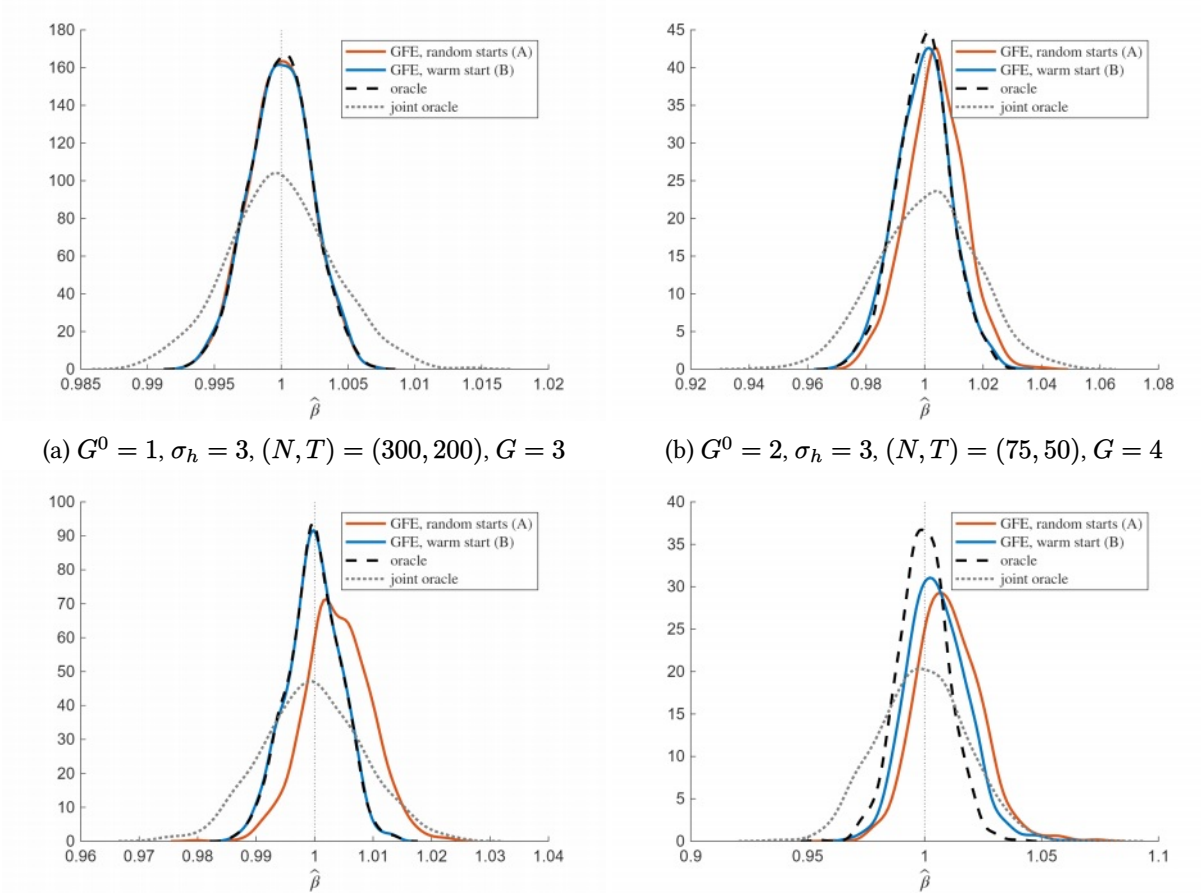
The mechanism described below operates in the same way, since what matters is that the initial slope is far from β^0.

respectively). Second, in the design of Remark 1, the GFE estimator with $G > G^0 = 1$ behaves like the oracle under every strategy, where the standard deviation ratios are at most 1.10, and equal to 1.00–1.03 at $(N, T) \in \{(150, 100), (300, 200)\}$, the κ-imbalance is at the level of a random split, $\sqrt{NT}\|\delta(\hat{\gamma})\|$ decreases with the sample size (from 0.42 at $(75, 50)$ to 0.22 at $(300, 200)$ for $G = 4$ and $\sigma_h = 3$), and the estimated partitions fit the idiosyncratic errors between 3 and 58 times better than the joint partition (in terms of $NT\, D(\gamma)$; not reported). This is the behavior that Assumption 5 describes.

Third, when the outcome equation has two groups and the covariate structure is strong ($G^0 = 2$, $\sigma_h = 3$), strategy A produces partitions aligned with the covariate structure, where the κ-imbalance rises to four to six times that of a random split when N and T grow together, $\sqrt{NT}\|\delta(\hat{\gamma})\|$ is of the order of that of the fixed informative refinement, and the standard deviation ratio grows with the sample size, from 1.14 at $(75, 50)$ to 1.45 at $(300, 200)$ for $G = G^0+3$ (from 1.08 to 1.41 for $G = G^0+2$, the case of Figure I(c))—the variance inflation of Remark 3. The mechanism is a path dependence of the algorithm. At a random initial partition the group effects are not absorbed and, since x_{it} contains $\alpha^0_{g_i^0 t}$, the initial slope estimate suffers from the omitted-variable bias of Section 3. The residuals $y_{it} - x_{it}'\hat{\beta}$ then carry the covariate structure $\sigma_h \kappa_i h_t$ scaled by this bias, the first classification aligns with it, and the alignment persists as a fixed point of the iterations once the bias has vanished, because such partitions fit the errors nearly as well as error-driven ones and gain through the covariate channel. Two variants confirm this reading: in the same design with the group effect removed from the covariates ($x_{it} = \sigma_h \kappa_i h_t + u_{it}$, so that the initial slope estimate is unbiased), the κ-imbalance of strategy A is at the level of a random split; and in the $G^0 = 1$ design the group effect is common to all units and absorbed by every partition, so that no imprinting occurs. A latent binary structure in the covariates alone therefore does not activate the mechanism. The group effect in the covariates matters only in combination with $G^0 > 1$, and what it generates is not a correlation between covariates and errors, which is zero in every design considered, but a component of the covariates that a random initial partition fails to absorb. Strategy B removes the path dependence, where its κ-imbalance is at the random-split level and its standard deviation ratio is at most 1.08 whenever $T \geq 25$. The objective discriminates between the two, but not sharply, since B attains the smaller objective in 73–95% of the replications when $G > G^0$, the differences being of the order of 10^{-3} of the objective, and in the remaining replications the covariate-aligned solution of A has the smaller objective. Strategy C, which retains the smaller objective, is therefore close to B (ratios of at most 1.06 at $(300, 200)$ and at most 1.12 whenever $T \geq 25$) but inherits part of the inflation of A. Since the global minimizer cannot be computed, the experiment cannot establish whether Assumption 5 holds for it in this design. What it establishes is that covariate-aligned and error-driven local minima coexist at nearly the same objective value, so that the finite-sample behavior of the computed estimator depends on the computational strategy, and that the inflation is a variance phenomenon. Indeed, the bias of strategy C is at most 0.25 oracle standard deviations whenever $T \geq 25$, in line with the $\sqrt{NT}$-consistency of Remark 3 (with $T = 10$ and $G^0 = 2$ the misclassification bias dominates, as in Appendix C.1).

Fourth, the covariate variation absorbed by the estimated partition is informative about which of the two situations one is in. Under Assumption 5, $\Omega(\hat{\gamma}) = o_p(1)$: the within-group covariate variation $\widehat{W}(\hat{\gamma}_G)$ stabilizes once $G \geq G^0$. The last block of columns of Table C.2 report $\|\Omega(\hat{\gamma})\|$. Under $G^0 = 1$, it is small and identical across strategies, and so is it under $G^0 = 2$ for strategy B, hence the additional groups absorb only the covariate variation that fitting the noise removes, which decreases like $1/N$, as Lemma B.4 predicts for the error-driven classification. Under $G^0 = 2$ and $\sigma_h = 3$, strategy A absorbs an order of magnitude

more: at $(N,T)=(300,200)$, $\|\Omega(\hat{\gamma})\|$ is 0.36 and 0.86 for $G=3$ and $G=5$ (0.75 for $G=4$, unreported), i.e., up to 40% of the covariate variation carried by the joint partition (2.24), whereas in strategy B it is below 0.04. Reporting $\widehat{W}(\hat{\gamma}_G)$ along with $\hat{\beta}_G$ across G thus provides a diagnostic for covariate-driven splits, which Section 6 uses.[12]



(a) $G^0=1$, $\sigma_h=3$, $(N,T)=(300,200)$, $G=3$

(b) $G^0=2$, $\sigma_h=3$, $(N,T)=(75,50)$, $G=4$

(c) $G^0=2$, $\sigma_h=3$, $(N,T)=(150,100)$, $G=4$

(d) $G^0=2$, $\sigma_h=3$, $(N,T)=(300,10)$, $G=4$

Figure C.2: Latent group structure in the covariates: Distribution of the GFE estimator under strategies A and B, the oracle and the joint oracle

Figure C.2 displays four of the cases in Table C.2 (the case $(300,200)$ with $G^0=2$ is Figure I(c)). Panel (a) is the second finding: with a single group in the outcome equation, the GFE estimator with $G=3$ coincides with the oracle under both strategies although the covariates carry a strong binary structure, and the joint oracle is visibly wider. Panels (b) and (c) are the third finding: as the sample grows from $(75,50)$ to $(150,100)$, the distribution of strategy A moves away from the oracle's and develops a second mode to the right of it—the replications in which the algorithm converged to a covariate-aligned partition—, whereas strategy B tracks the oracle. Panel (d) is a short panel: both strategies are distorted, by the misclassification of 3–4% of the units rather than by the mechanism above, as in Appendix C.1.

The dependence of Algorithm 1 on its starting values is well known, and BM (Supplemental Material, Section S.1) recommend many random starting values or their variable neighborhood search (Algorithm 2) to reach lower values of the objective. The point here is different. The covariate-aligned and the error-driven

[12]The second component of Assumption 5, $\delta(\hat{\gamma})$, is not observable, and its feasible analogue based on the GFE residuals is identically zero by the first-order condition for $\hat{\beta}_G$ at $\hat{\gamma}$. Its effect is visible only through the movement of $\hat{\beta}_G$ across G, see (B.9), which the stabilization rule of Section 6 tracks.

solutions have nearly the same objective value, so that a better optimizer does not remove the dependence on the initialization—in the replications in which the covariate-aligned solution has the smaller objective, it would select it—, and the two solutions have very different statistical properties. What discriminates between them is not the objective but the initialization. Two practical recommendations follow. First, when the covariates may carry latent structure of their own, the classification should be initialized at the k-means solution of residuals from a preliminary estimator whose partition already absorbs the group effects (strategy B), rather than at random assignments. Second, the within-group covariate variation should be monitored across G. Both are cheap.

C.3 Detectability of Under- and Overspecification

Let $\hat{\sigma}^2(G) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - x_{it}'\hat{\beta}^{(G)} - \hat{\alpha}_{\hat{g}_i t}^{(G)}\right)^2$ denote the GFE objective function when using G groups.

Table C.3 reports, in the design of Section 5.1 and for six panel sizes, the average reduction in the GFE objective from adding one group, the BIC penalty differential per additional group, and the BIC selection frequencies. Adding a group reduces the objective by about 0.20—one fifth of the error variance—for each $G \leq G^0$, and by 0.02–0.07 for $G > G^0$, the latter being the fit to the idiosyncratic errors $D(\hat{\gamma})$ of Appendix A, which Lemma B.4 bounds by $O_p(N^{-1/2} + T^{-1/2})$. Underspecification is thus a macroscopic event that the BIC misses in less than 1% of the samples in every design, whereas the gain from overspecification is of the order of the BIC penalty when T is short (0.067 against 0.087 for $(N,T) = (75,9)$), which is where the BIC overselects (7% of the samples). This is the asymmetry described in Section 3; see also Section S.3 in the Supplementary Appendix of BM.

Table C.3: Detectability of under- and overspecification

(N,T)	$\hat{\sigma}^2(G-1) - \hat{\sigma}^2(G)$				Penalty	$\mathbb{P}(\hat{G} < G^0)$	$\mathbb{P}(\hat{G} = G^0)$	$\mathbb{P}(\hat{G} > G^0)$
	$G=2$	$G=3$	$G=4$	$G=5$				
(75,9)	0.256	0.176	0.067	0.052	0.087	0.005	0.923	0.072
(75,20)	0.214	0.202	0.042	0.034	0.097	0.008	0.992	0.000
(75,50)	0.196	0.225	0.027	0.023	0.109	0.000	1.000	0.000
(150,20)	0.207	0.203	0.031	0.026	0.054	0.000	1.000	0.000
(300,20)	0.201	0.199	0.024	0.021	0.029	0.000	1.000	0.000
(150,50)	0.187	0.222	0.018	0.015	0.059	0.000	1.000	0.000

Notes: Design of Section 5.1, $G^0 = 3$. $\hat{\sigma}^2(G)$ is the GFE objective divided by NT; "Penalty" is the BIC penalty differential per additional group, $\hat{\sigma}^2 T\log(NT)/(NT)$, with $\hat{\sigma}^2$ estimated at $G_{max} = 5$. Based on 1,000 MC replications.

C.4 Group Similarity in the Application

Table C.4 reports three measures of similarity between the estimated partitions of Section 6 with G and $G+1$ groups, for $G = 1,\ldots,9$: the adjusted Rand index; the share of units that do not belong to the majority G-group of their $(G+1)$-group (units reshuffled beyond a pure split); and the share of unit pairs grouped together with $G+1$ groups that were separated with G groups, which is zero if the $(G+1)$-partition refines the G-partition, as Theorem 1(a) predicts when $G \geq G^0$. The transitions from $2 \to 3$ up to $7 \to 8$ reshuffle 8–14% of the units and join 7–21% of the pairs, i.e., the classification has not yet settled, whereas $8 \to 9$ and $9 \to 10$ are pure splits (no pair joined, no unit reshuffled, adjusted Rand indices of 0.98 and 0.99), in line with the stabilization of the estimates at $G = 8$ in Figure III. The partitions for $G = 9$ and

10 were estimated with the same algorithm as the others, with 200 starting values and, in addition, starting values obtained by splitting one group of the $(G-1)$-partition. With $T=7$, the objective is nearly flat across partitions that differ in a few units, so the reported transitions should be read together with this caveat.

The last two columns report the within-group second moments of lagged democracy and lagged income after residualizing on the estimated groups, $\widehat{W}_{kk}(\hat{\gamma}_G)$. The variation of lagged income drops sharply up to $G=4$ and is flat afterwards (0.54–0.61 for $4 \leq G \leq 10$), which is the behavior expected under Assumption 5 once the groups of the outcome equation are accounted for. The variation of lagged democracy keeps decreasing, as it must, since the lagged outcome contains the lagged group effects.

Table C.4: Similarity of the estimated group structures of Section 6 for consecutive G

$G \to G+1$	ARI	Reshuffled units	Pairs joined	Pairs split	$\widehat{W}_{11}(\hat{\gamma}_G)$	$\widehat{W}_{22}(\hat{\gamma}_G)$
$1 \to 2$	0.00	0.00	0.00	0.50	0.129	1.125
$2 \to 3$	0.48	0.12	0.15	0.42	0.067	0.724
$3 \to 4$	0.75	0.08	0.07	0.26	0.052	0.747
$4 \to 5$	0.78	0.09	0.07	0.24	0.035	0.607
$5 \to 6$	0.75	0.14	0.21	0.19	0.031	0.584
$6 \to 7$	0.82	0.10	0.08	0.20	0.035	0.581
$7 \to 8$	0.84	0.13	0.11	0.15	0.028	0.543
$8 \to 9$	0.98	0.00	0.00	0.03	0.024	0.571
$9 \to 10$	0.99	0.00	0.00	0.02	0.023	0.562
10					0.021	0.561

Notes: ARI is the adjusted Rand index between the partitions with G and $G+1$ groups; "Reshuffled units" is the share of units that do not belong to the majority G-group of their $(G+1)$-group; "Pairs joined" is the share of unit pairs grouped together with $G+1$ groups that were separated with G groups (zero if the $(G+1)$-partition refines the G-partition); "Pairs split" is the share of unit pairs grouped together with G groups that are separated with $G+1$ groups. $\widehat{W}_{kk}(\hat{\gamma}_G)$ is the within-(group,time) second moment of lagged democracy ($k=1$) and lagged income ($k=2$) after residualizing on the estimated groups with G groups.

C.5 Weak Separation with Heavy-tailed Errors

The rate condition of Theorem 2, $NT^{-r/2}c_T^{-r} = o(1)$, involves the number of moments r of the errors in Assumption 3(c). As discussed in Section 4, the admissible growth of N relative to T is pinned down by r and by the degree of separation, and a panel with $N \geq T$ is admissible only when the errors have enough moments. Table C.5 repeats the three designs of Section 5.2 ($\kappa = 0.45$, $G^0 = 2$, $\beta^0 = 1$) with the errors v_{it} of the outcome equation drawn from a standardized Student-t distribution with 5 and 3 degrees of freedom—finite moments of order below 5 and below 3, respectively—, all other shocks remaining standard Normal, while Figure C.3 shows the densities for the $t(3)$ case in the layout of Figure II(a).

With $t(5)$ errors, which satisfy the moment requirement of the $N \geq T$ regime, the results are indistinguishable from the Gaussian ones. With $t(3)$ errors, the picture changes at every panel size: misclassification no longer vanishes with the length of the panel (2% of the units at $(50, 150)$, where it is 0.1% with Gaussian errors), the bias is 0.3–0.5 oracle standard deviations, and the standard deviation of the GFE estimator is 1.4–1.7 times the oracle's. The misclassified units are those with the most extreme error paths—their average squared error is two to three times that of the correctly classified units, against a 10–30% excess with Gaussian or $t(5)$ errors (not reported in the table)—, and the distortion they induce does not shrink with T at the rate of the oracle's sampling error, which is why the ratio is largest in the longest panel. Heavy

tails thus interact with weak separation in the way the rate condition suggests: the number of moments determines whether misclassification can be made rare enough for oracle equivalence, and no choice of the panel dimensions among those considered compensates for their absence.

Table C.5: Weak separation with heavy-tailed errors

(N,T)	Bias	sd	sd of $\tilde{\beta}$	sd ratio	bias/sd of $\tilde{\beta}$	Misclass.
Gaussian errors						
$(50,150)$	0.0005	0.0117	0.0116	1.00	0.04	0.001
$(50,50)$	0.0076	0.0217	0.0196	1.11	0.39	0.039
$(150,50)$	0.0036	0.0120	0.0114	1.05	0.32	0.020
$t(5)$ errors						
$(50,150)$	0.0001	0.0120	0.0119	1.01	0.01	0.001
$(50,50)$	0.0073	0.0226	0.0201	1.13	0.36	0.037
$(150,50)$	0.0035	0.0119	0.0114	1.04	0.31	0.020
$t(3)$ errors						
$(50,150)$	0.0034	0.0203	0.0119	1.70	0.28	0.018
$(50,50)$	0.0093	0.0297	0.0188	1.58	0.49	0.057
$(150,50)$	0.0044	0.0161	0.0116	1.39	0.38	0.027

Notes: Design of Section 5.2, $\kappa = 0.45$, $G^0 = 2$, $\beta^0 = 1$. v_{it} is standard Normal or standardized Student-t with the indicated degrees of freedom (unit variance); "Bias" and "sd" refer to the GFE estimator with $G = G^0$, $\tilde{\beta}$ is the oracle based on the true groups, and "Misclass." is the misclassification rate. Based on 1,000 MC replications.

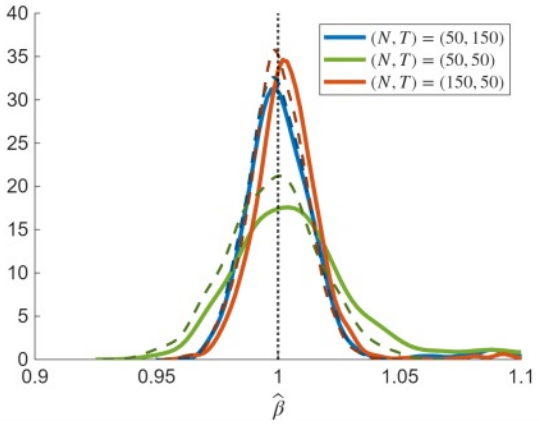


Figure C.3: Distribution of the GFE estimator (solid) and of the oracle (dashed) under weak separation with $t(3)$ errors