Beyond the Oracle: Misclassification-Robust Inference Under Latent Grouped Heterogeneity

Oriol González-Casasús*

Universitat Pompeu Fabra and Barcelona School of Economics

March 24, 2026

Please do not circulate

Abstract

This paper develops a framework for inference in models with latent grouped structures robust to imperfect membership estimation. Standard practice treats estimated group memberships as the ground truth, relying on the oracle property of classification algorithms. I show that, in finite samples, even high-accuracy classifiers may introduce systematic misclassification bias that distorts structural parameter estimates and invalidates standard inference. I characterize this non-standard bias using a local misclassification framework which provides a better approximation to small samples than oracle asymptotics, and enables the construction of bias-aware confidence intervals that are uniformly valid under a worst-case misclassification scheme. The method is easy to implement and allows for valid inference even in challenging contexts where groups are weakly separated or panels have a short time dimension. An interesting finding is that some classifiers automatically achieve bias reductions relative to others. Under some conditions, this new finding enables the construction of unbiased estimates despite classifiers are potentially imperfect and true groups remain unknown. I apply the framework to study investment-cash flow sensitivities during the 2008 financial crisis, which reveals that significant sensitivities historically documented in the literature may be largely artifacts of unobserved group-level heterogeneity and classification bias rather than pervasive financing frictions.

JEL Classification: C12, C23, C38, G31.

Key words: Latent Grouped Heterogeneity, Robust Econometrics, Local Misspecification, Grouped Fixed-Effects.

*Department of Economics and Business, Universitat Pompeu Fabra. Email: oriol.gonzalez@upf.edu. I thank Kirill Evdokimov, Simon Freyaldenhoven and Elena Manresa for helpful comments.

1 Introduction

The estimation of models with latent group structures has become a cornerstone of modern economic analysis. From capturing unobserved heterogeneity in panel data to identifying regimes in time series and categorizing agent types in networks, the ability to classify units into latent clusters is essential. The literature has seen a proliferation of methods to estimate these structures, ranging from parametric group formation models to advanced machine learning classifiers (e.g., Lin and Ng, 2012; Bonhomme and Manresa, 2015; Su, Shi and Phillips, 2016). The standard inferential practice in this domain relies on the "oracle property," the theoretical result that, asymptotically, latent groups are recovered with probability approaching one exponentially fast. Consequently, researchers routinely treat estimated group memberships as ground truth and proceed with standard inference, ignoring the uncertainty inherent in the classification step (see, also, Okui and Wang, 2021; Lumsdaine, Okui and Wang, 2023).

However, this reliance on oracle asymptotics is increasingly challenged (cf. Andrews, Kitagawa and McCloskey, 2021, 2024). In finite samples, classification is rarely perfect. When units are misclassified, the estimated group memberships function as a regressor subject to non-classical measurement error. Unlike standard measurement error, misclassification error is correlated with the true group identity and often the covariates, leading to complex biases in the estimators of structural parameters. Current asymptotic frameworks that assume perfect classification fail to capture these finite-sample distortions, leading to confidence intervals that severely undercover in practice. This problem is conceptually distinct from but related to the classic analysis of regression with a binary variable subject to observation error, dating back to Aigner (1973).

In this paper, I develop a framework to study the effects of classification errors on the structural parameters of a model. The approach in this paper treats the estimated group structure as the primitive of the analysis, rather than the outcome of a specific optimization problem. By abstracting away from the particular algorithm used to recover the groups, whether it be a parametric estimator from a group formation model or a black-box machine learning classifier, I ensure the results hold virtually regardless of the classification method employed provided an exogeneity condition is satisfied. Broadly speaking, the conditions allow for two general classes of classifiers: residual-based and external-proxy-based. I then proceed in two steps: first, I characterize how the noise inherent in these imperfect group assignments propagates to the structural parameters of the model; second, I develop an inference procedure that explicitly adjusts for this propagation. I focus the analysis on linear models with grouped fixed-effects, a setting that balances flexibility with analytical tractability. This choice allows me to derive explicit, interpretable expressions for the asymptotic bias, transparently revealing the mechanism through which misclassification distorts inference and providing the foundation for robust confidence intervals. Moreover, these models have become the de facto standard in empirical practice for capturing discrete unobserved heterogeneity, widely applied in fields ranging from labor economics to industrial organization; see, e.g., Bonhomme and Denis (2025) for a comprehensive survey.

A central theoretical contribution of this paper is the development of a local misclassification framework. In standard asymptotic analyses, classification errors are typically assumed to vanish exponentially fast (the oracle property) or remain fixed constants (leading to global inconsistency). I instead employ a local asymptotic device where the probability of misclassification decays at a rate on par with sampling variability. This framework allows me to analytically characterize the first-order bias induced by imperfect

grouping, a distortion that is empirically prevalent in finite samples but theoretically invisible under traditional oracle assumptions. By formalizing this intermediate regime, I provide a rigorous approximation of the behavior of estimators for structural parameters when the group structure is recovered with high, yet not perfect accuracy.

A first important finding is that misclassification does not lead to simple attenuation, but to a systematic mixing of group-specific objects that propagates into structural parameter estimates and can go in either direction. Crucially, unlike classical measurement error which is typically modeled as additive, symmetric, continuous noise, the error here arises from misclassification of a discrete group-membership indicator. Because the mismeasured covariate is a dummy variable, the error is inherently discrete and mechanically (negatively) correlated with the true covariate, the group-membership indicator. A second important finding is that, under some conditions, residual-based classifiers automatically achieve a reduction of the bias compared to proxy-based classifiers due to endogenous misclassification patterns. This enhances the appeal of the former methods, like the GFE estimator in Bonhomme and Manresa (2015), and supports their increasing popularity in the literature. In some particular cases, this novel insight can be used to exploit information from different classifiers to construct a combined estimator that does not suffer from misclassification distortions.

A computationally attractive feature of the proposed framework is that it operates as an ex post adjustment to existing inference procedures, requiring no modification to the underlying estimation strategy, hence making the method computationally simple and easy to implement. Practitioners can continue to employ their preferred classification algorithms to obtain point estimates, thereby preserving established empirical workflows and requiring minimal programming overhead. The practical contribution intervenes strictly at the inference stage: rather than relying on standard critical values, which are valid only under the often unrealistic assumption of perfect classification, I derive robust critical values that explicitly account for the bias induced by misclassification. This approach allows researchers to retain their standard point estimates while adjusting the length of confidence intervals to ensure valid coverage.

The present work provides a constructive solution to the weak group separation problem in models with latent group structures. It is well documented that when groups are not well-separated, classification accuracy degrades, and standard inference for structural parameters breaks down (Armstrong, Weidner and Zeleneev, 2022; González-Casasús, 2025). In the literature, strong separation between group-specific parameters is typically required to ensure that the probability of misclassification decays sufficiently fast—the oracle property. When separation is weak, misclassification probabilities may only decay slowly—or not at all—rendering standard inference invalid. The present results do not solve the identification problem of weak separation; rather, they provide inference that is robust to the consequences of it. Because weak separation manifests strictly as an increased probability of misclassification, and the method herein explicitly accounts for the maximum bias generated by a given rate of misclassification, the resulting confidence intervals remain valid even when the groups are difficult to distinguish. Indeed, the present local misclassification framework can be thought of as an artifact of local weak separation. This sidesteps the complex issue of developing weak separation tests as discussed in González-Casasús (2025). Interestingly, the strength of group separation explicitly shows up in this asymptotic theory, making transparent the required adjustment of the critical values due to it.

A distinguishing feature of the current analysis is the treatment of the latent grouped structure primarily

as a nuisance parameter. In many empirical applications—ranging from production function estimation to program evaluation—the structural parameter of scientific interest is the vector of common slope coefficients, while the group heterogeneity serves as a necessary control to prevent omitted variable bias. The present robust inference framework is specifically designed to protect these structural estimates from the propagation of classification errors, effectively profiling out the uncertainty associated with the latent group assignments. However, it is worth noting that this focus does not come at the expense of the incidental parameters; the theoretical machinery developed is general, and the characterization of the worst-case bias extends naturally to grouped fixed-effects and other group-level objects, offering a unified approach to inference under classification uncertainty. This work then complements inferential methods for group membership estimates as proposed by Dzemski and Okui (2024).

The asymptotic framework used in this paper conceptually parallels the literature on "small" measurement error in continuous variables, most notably the local approximations developed by Chesher (1991); see, more recently, Evdokimov and Zeleneev (2023). In that setting, the variance of the measurement error is modeled as shrinking to zero as the sample size increases, typically at a rate of $1/\sqrt{n}$. This device prevents the measurement error from dominating the asymptotic distribution while ensuring it does not vanish so quickly that its finite-sample distortion is lost in the limit. Similarly, the local misclassification framework models the probability of group misclassification as a shrinking sequence. Unlike fixed misclassification asymptotics, where bias would persist and lead to inconsistency, or the oracle asymptotics, where misclassification vanishes exponentially fast such that it becomes negligible, the proposed approach balances the decay rate of classification errors and sampling noise. This allows to derive an asymptotic distribution that explicitly captures the trade-off between sampling variability and the bias induced by imperfect group recovery, much like small measurement error approximations capture the first-order effects of small covariate contamination.

To the best of my knowledge, this is the first paper to derive the explicit asymptotic bias of structural parameters induced by misclassification, moving beyond the binary distinction between perfect and imperfect recovery found in the existing literature. This paper also departs from the existing literature by addressing the prevalent scenario where no auxiliary information—such as a validation subsample or repeated measurements—is available to estimate the matrix of misclassification probabilities or false error rates (e.g., Battaglia, Christensen, Hansen and Sacher, 2024). The proposed robust inference only relies on bounds for each group's total misclassification rate, which can be obtained off the shelf from the output of classifiers or can be calibrated by the user. However, if such auxiliary data were available, the results derived in this paper provide all necessary information to debias the estimates and construct correctly centered confidence sets. I also show that, under some conditions, using different classifiers sidesteps the need to estimate the matrix of misclassification probabilities to construct an asymptotically unbiased estimator.

Few papers study the consequences of misclassification, and do so differently than this paper. Recently, Dzemski and Okui (2021) consider a stylized grouped panel data model and show that standard $\sqrt{nT}$-consistency still holds provided misclassification is sufficiently small. This yields an asymptotic approximation that is indistinguishable from the case of perfect classification, rendering it unable to guide practitioners on how to handle the uncertainty present in finite samples. Therefore, the implications for structural inference remain bounded by the standard asymptotic paradigm. Bonhomme and Manresa (2015, Section S2.2) consider possible misclassification in grouped panel models with fixed T. However, they consider estimation of a pseudo-true value, whereas I consider inference on the true structural parameters. While

they identify the conditions for consistency and acknowledge the persistence of misclassification error in fixed-T settings, they do not characterize the asymptotic bias induced by this error on structural parameters. My contribution is to derive this explicit bias term and provide robust inference methods that remain valid even when the oracle recovery of groups fails, potentially because of a small T.

The results here contribute to the robust inference literature that guarantees validity by explicitly accounting for the worst-case bias induced by local parameter drift (Armstrong and Kolesár, 2018; Bonhomme and Weidner, 2022). Similar to how Conley, Hansen and Rossi (2012) assess sensitivity to local violations of exclusion restrictions, Kitamura, Otsu and Evdokimov (2013) analyze local perturbations in moment conditions, Rambachan and Roth (2023) protect against local violations of parallel trends, or González-Casasús and Schorfheide (2025) deal with local dynamic misspecification in the context of forecasting and dynamic causal effects, I derive the maximum bias arising from local deviations from perfect classification and study the consequences thereof for inference.

Finally, I illustrate the practical relevance of the theoretical results through both simulations and an empirical application. Monte Carlo evidence demonstrates that standard inference leads to severe size distortions when the group separation is weak or the time dimension is short, whereas the proposed robust confidence intervals maintain correct coverage. When I apply the method to revisit the debate on investment-cash flow sensitivities around the 2008 financial crisis, I find that allowing for a flexible grouped structure and acknowledging potential misclassification fundamentally alters established empirical narratives and renders cash flow sensitivities economically modest.

The remainder of the paper is organized as follows. Section 2 presents the results for the grouped cross-sectional model, while Section 3 extends these results to panel data. Section 4 develops the robust inference procedure. Section 5 provides numerical illustrations, and Section 6 applies the results to evaluate investment-cash flow sensitivities during the 2008 financial crisis. Section 7 concludes. All proofs, additional details on the empirical application and practical recommendations are collected in the Appendix.

2 Grouped Cross-Sectional Model

I start by considering a linear cross-sectional model with grouped fixed-effects:[1]

$$y_i = \alpha_{g_i} + x_i'\beta + \epsilon_i, \quad i \in [n],$$

where α_{g_i} is group's g fixed-effect, x_i is a K-dimensional vector of covariates and ϵ_i is an idiosyncratic error term. We may write $\alpha_{g_i} = \alpha' m_i$, where α is a G-dimensional vector of parameters, $m_i = e_{g_i}$ is a G-dimensional random vector of zeros with a one in position g_i indicating that unit i belongs to group $g_i \in [G]$.[2] Crucially, no restrictions are placed on the relationship between the fixed-effects and the covariates; the group-specific effects α_{g_i} serve as genuine fixed-effects capable of absorbing arbitrary correlation patterns between unobserved heterogeneity and the regressors.

In a standard setting, the variable m_i is latent, hence is replaced by $\hat{m}_i = e_{\hat{g}_i}$, where $\hat{g}_i$ is the estimated group

[1] The case of a model with common slope parameters is interesting to illustrate that the dangers of group misclassification even extend to such parameters. Clearly, these dangers are only exacerbated when the slope is allowed to vary across groups. Such an extension is worked out in Appendix F.

[2] For any natural number a, I define $[a] = \{1, 2, \ldots, a\}$.

of unit i. I list several common examples of classifiers, which easily generalize in more complex models. Examples 1-4 can be categorized as residual-based classifiers, while Examples 5-8 are instead classifiers based on external proxies. As the references in the different examples suggest, proxy-based classifiers are more common in cross-sectional applications, whereas residual-based classifiers dominate in panel data settings, discussed later in Section 3.

Example 1 (Finite Mixture Model) *Assume the error term follows a distribution with density $f(\cdot;\vartheta)$. One can define the classifier by the maximum a posteriori probability (Keane and Wolpin, 1997; Deb and Trivedi, 1997; Heckman and Singer, 1984):*

$$\hat{g}_i = \arg\max_{g\in[G]} \hat{\pi}_g f\left(y_i - x_i'\hat{\beta} - \hat{\alpha}_g \mid \hat{\vartheta}_g\right),$$

where $f(\cdot \mid \vartheta_g)$ is the probability density function characterized by group-specific parameters ϑ_g (e.g., variance or degrees of freedom), and $\hat{\pi}_g$ represents the estimated unconditional probability of belonging to group g, i.e., the mixing proportion.

Example 2 (Group or C-LASSO) *One could determine the group assignment by the convergence of individual effects (Ke, Fan and Wu, 2015; Su et al., 2016; Wang, Phillips and Su, 2018). First, minimize the objective*

$$Q(\beta,\tilde{\alpha},\alpha) = \sum_{i\in[n]} (y_i - \tilde{\alpha}_i - x_i'\beta)^2 + \frac{\lambda}{n}\sum_{i=1}^{n}\prod_{g=1}^{G}|\tilde{\alpha}_i - \alpha_g|,$$

where λ is a tuning parameter. The classifier is defined post-estimation by clustering the estimated $\hat{\tilde{\alpha}}_i$ values that have converged to the same local point:

$$\hat{g}_i = \arg\min_{k\in[G]} \|\hat{\tilde{\alpha}}_i - \hat{\alpha}_k\|^2.$$

Example 3 (k-Means or GFE) *The classifier can be defined by jointly minimizing parameter values and group assignments (Bonhomme and Manresa, 2015; Okui and Wang, 2021; Lumsdaine et al., 2023):*

$$\hat{g}_i = \arg\min_{g\in[G]} (y_i - x_i'\hat{\beta} - \hat{\alpha}_g)^2.$$

One may similarly apply k-means to a vector of external proxies $W_i = f(\alpha_{g_i}) + u_i$ to define the group memberships (Hubmer, Chan, Ozkan, Salgado and Hong, 2024; Gregory, Menzio and Wiczer, 2025).

Example 4 (Rank-Based) *This approach assumes that the latent groups are ordered monotonically with respect to the outcome, e.g., "High", "Medium" or "Low" types. This corresponds to the standard practice of sorting or portfolio formation in asset pricing (Fama and French, 1993; Almeida, Campello and Weisbach, 2004). Let $\hat{u}_i = y_i - x_i'\hat{\beta}$ be the proxy for the unobserved heterogeneity. One can define empirical quantiles $\hat{\tau}_0,\ldots,\hat{\tau}_G$ such that they partition the support of $\hat{u}$ into G intervals with probability masses $\pi_1,\ldots,\pi_G$ (often assumed $1/G$). The classifier is defined as:*

$$\hat{g}_i = \sum_{k=1}^{G} k\cdot \mathbb{1}\{\hat{\tau}_{k-1} < \hat{u}_i \leqslant \hat{\tau}_k\}.$$

Example 5 (k-Nearest Neighbor) *This method applies when group membership is determined by a vector*

of external proxies W_i. This is typically a supervised learning approach, implying the existence of a training set $\mathcal{L}$ (a subset of data where groups are known) or a previous iteration where pseudo-labels were generated (Abadie and Imbens, 2006).

Let (W_i, g_i) be the data in a training set $\mathcal{L}$, and let $d(\cdot, \cdot)$ be a metric on the covariate space (e.g., Euclidean distance). For a target observation $i \notin \mathcal{L}$, let $\mathcal{N}_k(i)$ be the set of indices of the k observations in $\mathcal{L}$ closest to W_i. The classifier estimates the local class probabilities:

$$\hat{p}_{ig} = \frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} \mathbb{1}\{g_j = g\}.$$

The discrete classification is then the mode of the local neighborhood:

$$\hat{g}_i = \arg\max_{g \in [G]} \hat{p}_{ig}.$$

In the absence of a labeled training set (unsupervised context), kNN is often used to construct an adjacency matrix A where $A_{ij} = 1$ if $j \in \mathcal{N}_k(i)$, which is then fed into a spectral clustering algorithm to minimize a normalized cut objective (Mantegna, 1999; Hoberg and Phillips, 2016).

Example 6 (Random Forests) *Machine learning methods are typically used when group membership is believed to be determined by a set of observable external characteristics W_i rather than just the outcome trajectory (Bajari, Nekipelov, Ryan and Yang, 2015).*

Let $\mathcal{T}$ be an ensemble of decision trees trained to predict latent types based on features W_i. The classifier is defined as the mode of the predictions of individual trees T_b:

$$\hat{g}_i = \text{mode}\left\{T_b(W_i)\right\}_{b=1}^{B}.$$

If the true groups are latent and unobserved, unsupervised random forests can be used to generate a similarity matrix between units based on how often they end up in the same leaf, which is then fed into a clustering algorithm (like k-means).

Example 7 (Neural Networks) *Similar to random forests, neural networks can model complex, non-linear relationships between external proxies W_i and group membership probabilities (Gu, Kelly and Xiu, 2020). The classifier is a function $f_\theta : \mathbb{R}^d \to [0,1]^G$ parameterized by weights θ (the network architecture). The network outputs a vector of probabilities (via a softmax layer), and the classifier selects the group with the highest probability:*

$$\hat{g}_i = \arg\max_{k \in [G]} \left[f_\theta(W_i)\right]_k$$

where $[v]_k$ denotes the k-th element of vector v, representing the predicted probability $P(g_i = k | W_i)$.

Example 8 (Large Language Models) *With the increasing popularity of artificial intelligence, one could also rely on a large language model to define the classifier. E.g., this has been successfully done to classify paragraphs of the Federal Open Market Committee announcements into groups to study monetary policy shocks (Hansen and Kazinnik, 2024; Gambacorta, Kwon, Park, Patelli and Zhu, 2024).*

One option could be to use an embedding approach, where the model is treated as a highly non-linear function that maps raw text into a fixed-dimensional vector space, and then apply a standard clustering algorithm.

When group membership is determined by unstructured text $\mathcal{D}_i$ (e.g., corporate filings), the classifier uses a pre-trained transformer $\varphi(\cdot)$ to map documents to embeddings $W_i = \varphi(\mathcal{D}_i) \in \mathbb{R}^d$. The groups are then defined by clustering these semantic vectors, e.g.,

$$\hat{g}_i = \arg\min_{k \in [G]} \|W_i - \hat{\mu}_k\|^2.$$

An alternative approach could be to utilize the instructional capabilities of modern large language models to directly output the classification label based on a prompt. Let $\mathcal{V}$ be the model's vocabulary. Let $f_\theta(t|context)$ be the probability distribution over the next token $t \in \mathcal{V}$ given a context sequence, parameterized by weights θ. Let $\mathcal{P}$ be a prompt designed to elicit the group structure (e.g., "Read the following financial report. Classify the firm as Constrained (Group 1) or Unconstrained (Group 2) based on its liquidity discussion"). The classifier is defined by the token (or sequence of tokens) that maximizes the conditional probability generated by the model:

$$\hat{y}_i = \underset{w \in \{\text{"Group 1"}, \ldots, \text{"Group } G\text{"}\}}{\arg\max} \prod_{j=1}^{|w|} f_\theta(w_j | \mathcal{P}, \mathcal{D}_i, w_{1:j-1})$$

The output string $\hat{y}_i$ is then mapped to the integer set:

$$\hat{g}_i = M(\hat{y}_i)$$

where M is a bijective mapping from the text labels to $[G]$.

Evidently, there may be a measurement error problem due to treating the estimates $\hat{m}_i$ as regular data. The goal in this section is to characterize the bias from this procedure—in particular, the bias in the estimates of common slope parameters β due to group misclassification.

Let $\psi = (\alpha', \beta')'$ and define $z_i = (m_i', x_i')'$ and $\hat{z}_i = (\hat{m}_i', x_i')'$. The least squares estimator of the feasible model is

$$\hat{\psi} = \left(\frac{1}{n} \sum_{i=1}^n \hat{z}_i \hat{z}_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^n \hat{z}_i y_i.$$

The least squares estimators $\hat{\alpha}$ and $\hat{\beta}$ of α and β are the upper and lower blocks of $\hat{\psi}$, respectively.

I use asymptotic theory to approximate the finite-sample distribution of $\hat{\psi}$. Both measurement error in $\hat{m}_i$ and downstream sampling error determine the finite-sample distribution. To guarantee that asymptotics deliver a useful approximation, I adopt a framework in which both sources of error remain present as the sample size n grows large. This is achieved by allowing the accuracy of the classifier $\hat{m}_i$ to increase with n at an appropriate rate. Formally, I consider the following high-level assumption:

Assumption 1 *(a) $\mathbb{E}[\|z_i\|^2] < \infty$, $\mathbb{E}[\|\epsilon_i z_i\|^2] < \infty$ and $\mathbb{E}[z_i z_i']$ has full rank.*

(b) $\frac{1}{n} \sum_{i=1}^n \hat{z}_i \hat{z}_i' \xrightarrow{p} \mathbb{E}[z_i z_i']$ as $n \to \infty$, and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{z}_i (\hat{m}_i - m_i)' \xrightarrow{p} \begin{bmatrix} \Omega \\ \Gamma \end{bmatrix} := \begin{bmatrix} \sum_g \pi_g \sum_h q_{h|g} e_h (e_h - e_g)' \\ \sum_g \pi_g \mathbb{E}[x_i | g_i = g] \sum_h q_{h|g} (e_h - e_g)' \end{bmatrix}.$$

(c) $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{z}_i\epsilon_i \xrightarrow{d} N(\Xi, \mathbb{E}[\epsilon_i^2 z_i z_i'])$ *and* $\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2\hat{z}_i\hat{z}_i' \xrightarrow{p} \mathbb{E}[\epsilon_i^2 z_i z_i']$ *as* $n \to \infty$, *where* $\hat{\epsilon}_i = y_i - \hat{z}_i'\hat{\psi}$ *and*

$$\Xi := \begin{bmatrix} \sum_g \pi_g \sum_{h \neq g} q_{h|g} \mathbb{E}[\epsilon_i | \hat{g}_i = h, g_i = g](e_h - e_g) \\ 0 \end{bmatrix}.$$

(d) *The error* ϵ_i *has finite third moments. Moreover, for any* $h \neq g \in [G]$, *the random variable* $R_i = \epsilon_i(\alpha_h - \alpha_g)$, *with upper support* ω_R *and density* f_R, *possesses a mean residual life function* $m_R(c) = \mathbb{E}[R_i - c \mid R_i > c]$ *such that* $m_R(c_n) = o(c_n)$ *and* $f_R(c_n) = o(n^{-1/4})$ *for any sequence* $c_n \to \omega_R$ *as* $n \to \infty$.

Technically, I consider a sequence of populations in which the distribution of (y_i, x_i, m_i) is held fixed and the conditional distribution of $\hat{m}_i$ given (y_i, x_i, m_i) varies with n so that Assumption 1 is satisfied. While this local misclassification framework is only a technical device not to be taken literally, a heuristic interpretation is that it captures the usual practice whereby grouped models are estimated by increasingly accurate, yet imperfect classifiers.

Assumption 1(a) is a standard moments condition. The key requirements for the local misclassification framework are captured in Assumptions 1(b)-(d).

To better understand the current setup, note that for a unit with true group g, if correctly classified $\hat{m}_i - m_i = 0$, otherwise if classified to h, $\hat{m}_i - m_i = e_h - e_g$. Thus each misclassified unit contributes the rank-one matrix

$$\hat{m}_i(\hat{m}_i - m_i)' = e_h(e_h - e_g)' = e_h e_h' - e_h e_g'.$$

Summing over units and normalizing gives proportions of units assigned to each (h, g) pair (assigned h but true g). If the raw misclassification proportions (for fixed g) were of order 1, i.e., do not vanish as $n \to \infty$, the sum $\sum_{i=1}^{n}\hat{m}_i(\hat{m}_i - m_i)'$ would be $O(n)$ and after dividing by $\sqrt{n}$ would diverge—which is incompatible with a finite limit. In this case, misclassification bias would dominate and drive the asymptotics. If misclassification proportions were of order smaller than $n^{-1/2}$ they would vanish at the $\sqrt{n}$-scale. In this case, misclassification bias would be negligible with respect to sampling variance in the limit. Thus, the assumption that the sum divided by $\sqrt{n}$ converges to a finite nonzero matrix requires that the misclassification probability decays at the rate $n^{-1/2}$. Then, the limit Ω captures the directional structure of misclassification: which pairs $g \mapsto h$ are happening more than others. This is the nondegenerate case studied in the paper.

In particular, if misclassification is random with small probabilities $p_{h|g}^{(n)}$ (probability that a unit whose true group is g is assigned to group h), then for each true group g the expected contribution to $\hat{m}(\hat{m} - m)'$ from such units is

$$\pi_g \sum_h p_{h|g}^{(n)} e_h(e_h - e_g)'.$$

If $p_{h|g}^{(n)}$ scales so that $\sqrt{n}p_{h|g}^{(n)} \to q_{h|g}$ for some limiting numbers $q_{h|g}$ as $n \to \infty$, then we recover the form for Ω in Assumption 1(b). It is important to stress that the sequence $p_{h|g}^{(n)}$ is not intended to represent a literal sequence of data-generating processes (DGPs) but is rather employed as a purely theoretical device, akin to local-to-unity (Phillips, 1987) or weak instrument asymptotics (Staiger and Stock, 1997). In finite samples, misclassification probabilities are strictly positive and fixed, which would theoretically dominate

the asymptotics and yield a degenerate, biased limit. By modeling the misclassification rate as a drifting sequence that balances the order of the estimation error, we obtain a non-degenerate asymptotic distribution that captures the first-order finite-sample distortions of classification noise. This provides a more reliable approximation for finite-sample inference than standard oracle asymptotics.[3]

The other condition in Assumption 1(b), $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i(\hat{m}_i - m_i)' \xrightarrow{p} \Gamma$, characterizes the local asymptotic cross-moment between the covariates and the group assignment error. The underlying requirement for this limit to be properly defined as a function of unconditional group moments is a strict exogeneity condition: $\mathbb{P}(\hat{g}_i = h|g_i = g, x_i) = \mathbb{P}(\hat{g}_i = h|g_i = g)$. Under this condition, the covariate mean of the misclassified subpopulation equals the unconditional mean of their true group, $\mathbb{E}[x_i|\hat{g}_i = h, g_i = g] = \mathbb{E}[x_i|g_i = g]$. This exogeneity is naturally satisfied by classification algorithms, like those listed in the examples above, that partition units based on trajectories of residuals, when $\epsilon_i \perp x_i|g_i$, or that are based on external proxies $W_i = f(\alpha_{g_i}) + u_i$, when $u_i \perp x_i|g_i$.[4]

Crucially, while the misclassification event is conditionally mean independent of x_i, the limit matrix Γ is generally non-zero. This is because the latent grouped structure operates as a fixed-effect; true group assignments are inherently correlated with the covariates. When units from group g are misclassified into group h, they carry their distinct covariate distribution into the estimated group h, and Γ captures this compositional shift. Consequently, misclassification distorts $\hat{\beta}$ not solely through the attenuation of the estimated group intercepts, but by inducing a spurious correlation between the covariates and the mismeasured group assignments. Finally, it must be noted that this strict exogeneity condition—and thus the validity of the subsequent robust inference—would be systematically violated by classifiers that directly partition the included regressor space x_i.

Assumption 1(c) requires that the feasible score satisfies a central limit theorem. The asymptotic bias in the feasible score Ξ is due to potential endogenous misclassification, and depends on the conditional distribution of the error term for the misclassified subpopulation, $\mathbb{E}[\epsilon_i|\hat{g}_i = h, g_i = g]$. Unlike the exogeneity condition imposed above on the covariates, this term cannot be simplified for residual-based classifiers.

It is useful to define $\rho_{hg} = \mathbb{E}[\epsilon_i|\hat{g}_i = h, g_i = g]/(\alpha_h - \alpha_g)$, the error mean in the misclassified subpopulation normalized by the distance between groups. For residual-based classifiers, Assumption 1(d) is employed to streamline the endogenous misclassification patterns, where the exact value of ρ_{hg} will depend on the specific classifier used. Without Assumption 1(d), the derivation of the asymptotic distribution of $\hat{\psi}$ remains unchanged, yet the exact bias reduction cannot be cleanly characterized. Appendix H argues that in our framework, generally $0 \leqslant \rho_{hg} < 1$ with high probability.[5] For classifiers that rely on external proxies, ρ_{hg} will be exactly zero due to the orthogonality of ϵ_i and u_i. On the other hand, ρ_{hg} becomes non-negligible for classifiers that exploit the residuals in the model, since in that case the tails of the error distribution characterize misclassification patterns. For instance, when using the GFE classifier, Proposition A.1 shows that with high probability $\rho_{hg} = 1/2$.

[3] Although this local misclassification framework can be explicitly micro-founded via sequences of data generating processes with specific growing signal-to-noise ratios, I abstract from these underlying primitives, as the drifting parameters serve purely as theoretical devices to focus on the finite-sample impact of the resulting misclassification. Moreover, these micro-foundations would necessarily be classifier-specific, hence less general than the approach used here.

[4] The conditional mean independence assumption is also valid for residual-based classifiers. While the error term is conditionally orthogonal to the covariates, the residuals are not. However, per the arguments in the proof of Proposition A.1, with high probability $\mathbb{E}[x_i|\hat{g}_i = h, g_i = g] = \mathbb{E}[x_i|\tilde{g}_i = h, g_i = g]$, where $\tilde{g}_i$ is the oracle (infeasible) classifier. Since $\tilde{g}_i$ only relies on the true error terms, $\mathbb{E}[x_i|\hat{g}_i = h, g_i = g] = \mathbb{E}[x_i|g_i = g]$ with high probability.

[5] Consistent with this, the bounds $0 \leqslant \rho_{hg} < 1$ will be implicitly assumed throughout this paper.

Assumption 1(d) naturally encompasses a large class of widely used error distributions with sufficiently light tails. For instance, it is satisfied by the Gaussian distribution, distributions with strictly bounded support (such as the Uniform), and distributions with exponential tails (such as the Laplace). The tail rate can be relaxed at the expense of assuming finiteness of higher order moments of the error. However, this assumption explicitly rules out heavy-tailed distributions governed by power laws, such as the Student-t, Pareto or Cauchy. Since the only purpose of the assumption is to yield a tractable characterization of the bias reduction factor ρ_{hg}, those limitations are not deemed worrisome.

In the current framework, it is possible to show the following:

Theorem 1 *Under Assumptions 1(a)-(c), as* $n \to \infty$,

$$\sqrt{n}(\hat{\psi} - \psi) \xrightarrow{d} N\left(-\mathbb{E}[z_i z_i']^{-1}\left\{\begin{bmatrix}\Omega & 0\\ \Gamma & 0\end{bmatrix}\psi - \Xi\right\}, V\right),$$

where $V = \mathbb{E}[z_i z_i']^{-1}\mathbb{E}[\epsilon_i^2 z_i z_i']\mathbb{E}[z_i z_i']^{-1}$. *Moreover,*

$$\hat{V} = \left(\frac{1}{n}\sum_{i=1}^{n}\hat{z}_i\hat{z}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2\hat{z}_i\hat{z}_i'\right)\left(\frac{1}{n}\sum_{i=1}^{n}\hat{z}_i\hat{z}_i'\right)^{-1} \xrightarrow{p} V.$$

Intuitively, Theorem 1 establishes that, while the estimator $\hat{\psi}$ remains consistent and asymptotically normal under local misclassification, it is no longer centered at the true parameter value ψ. Instead, the imperfect recovery of latent groups introduces a systematic bias term, shifting the sampling distribution away from zero even as the sample size approaches infinity. The estimator's variance, however, remains unaffected by misclassification. This asymptotic paradigm is confirmed in Figure I, calibrated to the simulation results of Section 5, where misclassification translates into a location shift of the estimator's distribution. The figure confirms how the local misclassification asymptotics are able to provide an accurate approximation of the finite-sample behavior of estimators under good, yet imperfect membership estimation of grouped structures, in a way that oracle asymptotics cannot.

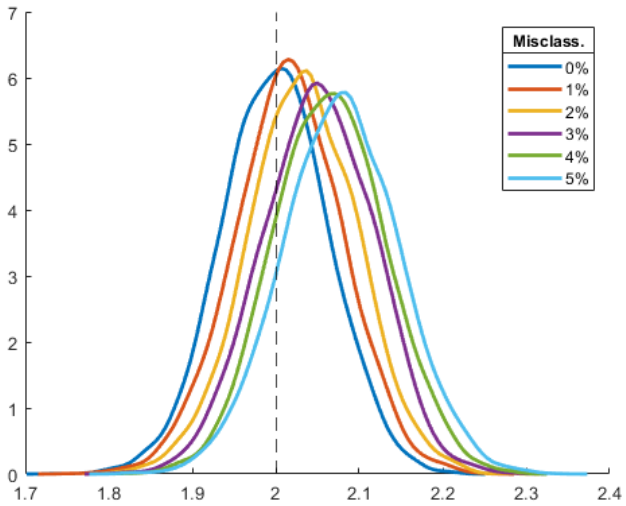


Figure I: Finite sample distribution of $\hat{\beta}$ for varying degrees of misclassification in $\hat{g}$

One key implication from Theorem 1 is the following:

Corollary 1 *Under Assumption 1, as $n \to \infty$, the asymptotic bias of $\hat{\beta}$ is*

$$ABias(\hat{\beta}) = \sum_{a=1}^{G} \sum_{b \neq a} \kappa_{ab} q_{a|b},$$
$$\kappa_{ab} = \left(\sum_{g=1}^{G} \pi_g \text{Var}(x_i | g_i = g) \right)^{-1} \pi_b (1 - \rho_{ab}) (\mathbb{E}[x_i | g_i = a] - \mathbb{E}[x_i | g_i = b]) (\alpha_a - \alpha_b)$$

Corollary 1 provides an explicit expression of how group shares π_b, (local) misclassification rates $q_{a|b}$, group-specific covariate mean contrasts $\mathbb{E}[x_i|g_i = a] - \mathbb{E}[x_i|g_i = b]$, and fixed-effects contrasts $\alpha_a - \alpha_b$ enter the asymptotic bias, and how the inverse average within-group covariance weights those vectors. It aligns with the intuition that misclassification bias is negligible in several specific scenarios. First, if grouped fixed-effects are homogeneous, misallocating units across groups generates no bias because the unobserved heterogeneity is effectively constant, hence fully captured under any grouping scheme. The result is also consistent with the findings in González-Casasús (2025) whereby, under weak group separation, the consequences on parameter estimates are negligible regardless of classification accuracy. Second, the bias contribution vanishes if the regressors in the misclassified groups have little heterogeneity in means or if these groups represent a negligible share of the total sample. A trivial case where bias would be negligible is a random effects specification, where naturally $\mathbb{E}[x_i|g_i = a] = \mathbb{E}[x_i]$. Finally, asymptotic unbiasedness is also preserved if the misclassification probability is of a lower order than $n^{-1/2}$, meaning the classification error dissipates faster than the parametric sampling noise.

An interesting pattern in the bias captured by Corollary 1 is due to the type of classifier employed. It turns out that residual-based classifiers achieve an automatic reduction of the bias compared to proxy-based classifiers, since the former imply $0 < \rho_{hg} < 1$ whereas the latter have $\rho_{hg} = 0$. While the exact bias reduction depends on the symmetry of the classifier, a powerful rule of thumb is $\rho_{hg} = 1/2$, i.e., 50% bias reduction, achieved for instance by the popular GFE estimator of Bonhomme and Manresa (2015). Upon inspection of the proof of Corollary 1, one can see that the underlying mechanisms are due to the endogenous misclassification patterns; the Ξ term. Heuristically, it happens because the endogeneity of the classification error actively offsets the classical measurement error. In proxy-based classifiers, the misclassification event is orthogonal to the distortion that the least squares estimate $\hat{\psi}$ has to absorb. However, in proxy-based classifiers, the large shock that generates misclassification in first place, attenuates the error that the least squares estimate $\hat{\psi}$ has to absorb; see Appendix H for more detailed intuition and supporting simulation evidence. This novel finding supports the increasing shift towards residual-based classifiers observed in the literature (e.g., Bonhomme and Manresa, 2015; Su et al., 2016). If errors fail to satisfy Assumption 1(d), the exact relative bias reduction is unknown and harder to characterize.

Given the above insight on the relative bias due to different classifiers, it immediately becomes possible to construct a debiased estimator. Provided there are available external proxies with the same group signal as the residuals (i.e., same implied misclassification rates $q_{a|b}$) and under the rule of thumb $\rho_{hg} = 1/2$, Corollary 1 implies that the estimator $\tilde{\beta} = 2\hat{\beta}_R - \hat{\beta}_P$ will be asymptotically unbiased, where $\hat{\beta}_R$ refers to the least squares estimator that relies on a residual-based classifier, and $\hat{\beta}_P$ that which uses a proxy-based classifier. This is an important observation that permits debiasing the estimator without auxiliary data to

estimate the entire misclassification matrix, namely we do not require knowledge of the true groups. While asymptotically unbiased, in practice $\tilde{\beta}$ may suffer from larger mean squared error that either $\hat{\beta}_P$ or $\hat{\beta}_R$.[6]

Corollary 1 captures an interesting trade-off related to the degree of group separation: as separation lowers and groups become more similar, misclassification probability rises, but the penalty for misclassification shrinks. Another important observation from a practical perspective is that it is without loss of generality to assume that regressors are demeaned, since the fixed-effects contrasts would cancel out any scaling.[7] Therefore, the group-specific mean $\mathbb{E}[x_i|g_i = a]$ contribution to the sensitivity weights κ_{ab} may be understood relative to the grand mean, $\mathbb{E}[x_i|g_i = a] - \mathbb{E}[x_i]$. This is a crucial observation for accurate inference, and Proposition 3 below exploits it.

It is crucial to underscore that the misclassification bias derived above is fundamentally non-classical, as already hinted in Figure I. In standard errors-in-variables models, measurement error is typically assumed to be independent of the true regressors, leading to well-known attenuation bias, shrinking estimates toward zero. In contrast, the error induced by misclassifying latent groups is structural and endogenous: the measurement error in the fixed-effects is perfectly correlated with the true group identity and, by extension, correlated with the covariates whenever the regressor distribution varies across groups. Moreover, the mismeasured covariate is not continuous and centered around the true covariate as in the classical setting, but is binary for indicating group membership. As a result, the sign of the asymptotic bias is indeterminate ex ante. The bias is a linear combination of misclassification probabilities $q_{a|b}$ weighted by the sensitivity parameters κ_{ab}. The sensitivity weights matrix κ plays a very important role for estimation and inference, as later illustrated in Section 6. Since κ_{ab}, which captures the cross-term between regressors and group differences, can be positive or negative, misclassification can lead to either attenuation or amplification of the estimated slope coefficients. A coefficient could even change signs entirely if the misclassification is severe and systematic enough.

One special instance where the sign of the bias can be easily characterized is in the case of binary grouping, as Corollary 2 shows. This result highlights a fundamental divergence from classical measurement error, and illustrates that amplification bias happens if and only if the group with the larger fixed-effects also has a higher average covariate value. Since the textbook result of attenuation bias in the classical measurement error framework is typically derived for a scalar parameter of interest (see, e.g., Wooldridge, 2010, pp. 73-76), Corollary 2 is simplified accordingly.

Corollary 2 *Suppose $G = 2$ and $K = 1$. The asymptotic bias of $\hat{\beta}$ is positive if and only if*

$$(\mathbb{E}[x_i|g_i = 1] - \mathbb{E}[x_i|g_i = 2])\,(\alpha_1 - \alpha_2) > 0.$$

[6] More specifically, Proposition A.2 shows that $\sqrt{n}(\tilde{\beta} - \beta) \xrightarrow{d} N(0, V_\beta)$. Under the local misclassification asymptotics, there is no efficiency cost of debiasing because misclassification is not large enough to affect the variance. In particular, Assumption 1(c) implies that misclassification only affects the first moment of the feasible scores. Thus, least squares estimates under either classifier will be asymptotically driven by the same score, hence will be perfectly correlated, such that

$$\text{Var}(\tilde{\beta}) = 4\text{Var}(\hat{\beta}_R) + \text{Var}(\hat{\beta}_P) - 4\text{Cov}(\hat{\beta}_P, \hat{\beta}_R) = (4 + 1 - 4)V_\beta = V_\beta.$$

In practice, however, the additional noise through the debiasing structure of $\tilde{\beta}$ may be non-negligible.

[7] It is a standard result that fixed-effects capture any level shifts. More specifically, the model under consideration is equivalent to $y_i = \tilde{\alpha}_{g_i} + \tilde{x}_i'\beta + \epsilon_i$, where $\tilde{\alpha}_{g_i} = \alpha_{g_i} + \mathbb{E}[x_i]'\beta$ and $\tilde{x}_i = x_i - \mathbb{E}[x_i]$. Since the bias of $\hat{\beta}$ only depends on fixed-effects and covariate mean contrasts, and $(\mathbb{E}[x_i|g_i = a] - \mathbb{E}[x_i]) - (\mathbb{E}[x_i|g_i = b] - \mathbb{E}[x_i]) = \mathbb{E}[x_i|g_i = a] - \mathbb{E}[x_i|g_i = b]$ and $\tilde{\alpha}_g - \tilde{\alpha}_{g'} = \alpha_g - \alpha_{g'}$, it is without loss of generality for inference on β to assume that x_i in the original model is mean zero.

Furthermore, misclassification distortion is not only contained within the slope vector β; it propagates directly to the grouped fixed-effects, $\alpha = (\alpha_1, \ldots, \alpha_G)'$. The estimator for the grouped fixed-effects, $\hat{\alpha}_g = \bar{y}_g - \bar{x}_g'\hat{\beta}$, suffers from a dual source of bias. First, there is the indirect transmission from the biased slope estimate $\hat{\beta}$. Second, there is a direct composition effect: even if β were known effectively, $\hat{\alpha}_g$ is estimated using a sample of units assigned to group g, which inevitably includes misclassified units from other groups with different intercepts. Consequently, inference on group rankings or value-added measures is equally susceptible to the robustness concerns raised here, as the next corollary formalizes.

Corollary 3 *Under Assumption 1, as $n \to \infty$, the asymptotic bias of $\hat{\alpha}$ is*

$$ABias(\hat{\alpha}_a) = -\frac{1}{\pi_a}\sum_{b \neq a} \pi_b(1-\rho_{ab})q_{a|b}(\alpha_a - \alpha_b) - (\mathbb{E}[x_i|g_i = a] - \mathbb{E}[x_i])'ABias(\hat{\beta}).$$

Corollary 3 shows that the relative bias reduction for different classifiers equally extends to group-specific objects.

Before concluding this section, I present a result that will prove helpful for constructing valid inference.

Proposition 1 *Define*

$$\hat{\kappa}_{ab} = \left(\sum_{g=1}^{G} \hat{\pi}_g \widehat{\text{Var}}(x_i|\hat{g}_i = g)\right)^{-1} \hat{\pi}_b(1-\hat{\rho}_{ab})(\hat{\mathbb{E}}[x_i|\hat{g}_i = a] - \hat{\mathbb{E}}[x_i|\hat{g}_i = b])(\hat{\alpha}_a - \hat{\alpha}_b),$$

where hat objects are the estimated or sample counterparts. Under Assumption 1, $\hat{\kappa}_{ab} \xrightarrow{p} \kappa_{ab}$.

A key practical implication of Proposition 1 is that the sensitivity of the estimator to misclassification, captured by κ_{ab}, is consistently estimable from the data. This implies that even without knowledge of the true groups or classification error rates, researchers can assess the potential vulnerability of their estimates by examining the magnitude of these sensitivity weights. This becomes crucial in practice to understand the source and direction of the bias; see Section 6.

3 Grouped Panel Data Model

In this section I extend the model studied earlier to include a time dimension:

$$y_{it} = \alpha_{g_i t} + x_{it}'\beta + \epsilon_{it}, \quad i \in [n],\ t \in [T].$$

It is well-known that panels with large T tend to generate little misclassification (Bonhomme and Manresa, 2015; González-Casasús, 2025), hence a framework with small T is needed to understand the forces at play in finite samples.[8] At the same time, the widely used assumption of time-invariant groups is more credible in a framework with fewer time periods. The setting where the time dimension of the panel is small is empirically relevant, as it is commonplace in many economic applications (see, e.g., Bonhomme

[8] One technical advantage of assuming fixed T is that we avoid dealing with an infinite-dimensional nuisance parameter α. An analogous finite-dimensional central limit theorem shall still apply in a setting where $T \to \infty$ because we are only interested in a finite subset of the parameter vector ψ, namely β (Cattaneo, Jansson and Newey, 2018), yet the resulting asymptotic bias would be harder to interpret and operationalize since it would depend on the infinite-dimensional nuisance parameter.

and Manresa, 2015; Grunewald, Klasen, Martínez-Zarzoso and Muris, 2017; Sanogo, 2019; Gómez-Puig, Sosvilla-Rivero and Martínez-Zarzoso, 2022).

A rigorous analysis of grouped panel data models with fixed T faces a dilemma. Since the number of time periods is finite, latent groups may not be perfectly recovered, implying that the misclassification probability is strictly non-zero in the limit. Under standard asymptotics with fixed misclassification, the estimator $\hat{\psi}$ is inconsistent, converging to a pseudo-true value. However, viewing the estimator merely as inconsistent provides no guidance on the magnitude of the distortion or the reliability of inference in finite samples.

To resolve this, I retain the local misspecification framework as a technical device to approximate the finite-sample distribution. In the panel setting, the local misclassification rates $q_{h|g,T}$ would naturally depend on the time dimension T. Under the idea that misclassification errors are small, they should have the property that $q_{h|g,T} \to 0$ as $T \to \infty$. However, in finite samples where T is small, $q_{h|g,T}$ will be a non-negligible positive constant. Since T will be fixed in our framework, for economy of notation the subscript T will be suppressed, and I will continue to write $q_{h|g}$.

I stress again that the drifting framework is not a claim about the true DGP; rather, it is a theoretical experiment that prevents the bias term from dominating the asymptotics. By balancing the order of the bias and the variance, this framework yields a non-degenerate asymptotic distribution that captures the first-order effect of classification noise. This allows us to characterize the sensitivity of the estimator to imperfect grouping in a way that fixed-parameter asymptotics cannot, providing a theoretical foundation for robust inference that remains valid even when T is small.

Let us then fix T and rewrite the grouped panel model as

$$y_i = \alpha_{g_i} + x_i'\beta + \epsilon_i,$$

where now $y_i = (y_{i1}, \ldots, y_{iT})'$, likewise for α_{g_i} and ϵ_i, and x_i is the $K \times T$ matrix of regressors for unit i. Following the literature, group memberships are assumed to be time-invariant, so that we may write

$$m_i = e_{g_i} \otimes \mathrm{I}_T,$$

and express $\alpha_{g_i} = m_i'\alpha$, where α is a GT-dimensional vector and $\otimes$ denotes the Kronecker product operator. It is then easy to see that the cross-sectional results generalize to the panel data setting in a straightforward way:

$$y_i = Z_i'\psi + \epsilon_i,$$

where $Z_i = \begin{bmatrix} m_i' & x_i' \end{bmatrix}'$ is a $(GT + K) \times T$ matrix and $\psi = (\alpha', \beta')'$ is a $(GT + K)$-dimensional parameter vector. Standard least squares algebra gives

$$\hat{\psi} = \left(\sum_{i=1}^{n} \hat{Z}_i \hat{Z}_i' \right)^{-1} \sum_{i=1}^{n} \hat{Z}_i y_i.$$

To characterize the asymptotic distribution of $\hat{\psi}$ in the panel case, I consider the following generalization of Assumption 1. Because T is fixed, we can allow arbitrary dependence across time and consider limit

results that depend on the cross-sectional distribution.

Assumption 2 *(a)* $\mathbb{E}[\|Z_i\|^2] < \infty$, $\mathbb{E}[\|Z_i\epsilon_i\|^2] < \infty$ *and* $\mathbb{E}[Z_iZ_i']$ *has full rank.*

(b) $\frac{1}{n}\sum_{i=1}^n \hat{Z}_i\hat{Z}_i' \xrightarrow{p} \mathbb{E}[Z_iZ_i']$ *as* $n \to \infty$, *and*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{Z}_i(\hat{m}_i - m_i)' \xrightarrow{p} \begin{bmatrix} \Omega \otimes \mathrm{I}_T \\ \bar{\Gamma} \end{bmatrix},$$

where $\bar{\Gamma} := \sum_g \pi_g \sum_h q_{h|g}\mathbb{E}[x_i|g_i = g]((e_h - e_g)' \otimes \mathrm{I}_T)$.

(c) $\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{Z}_i\epsilon_i \xrightarrow{d} N(\bar{\Xi}, \mathbb{E}[Z_i\epsilon_i\epsilon_i'Z_i'])$ *and* $\frac{1}{n}\sum_{i=1}^n \hat{Z}_i\hat{\epsilon}_i\hat{\epsilon}_i'\hat{Z}_i' \xrightarrow{p} \mathbb{E}[Z_i\epsilon_i\epsilon_i'Z_i']$ *as* $n \to \infty$, *where*

$$\bar{\Xi} := \begin{bmatrix} \sum_g \pi_g \sum_{h \neq g} q_{h|g}((e_h - e_g) \otimes \mathrm{I}_T)\mathbb{E}[\epsilon_i|\hat{g}_i = h, g_i = g] \\ 0 \end{bmatrix}.$$

(d) The error vector ϵ_i *has a distribution such that* $\mathbb{E}[\epsilon_i \mid a'\epsilon_i = r] = \frac{a}{\|a\|^2}r$ *for any* $a \in \mathbb{R}^T$, *and possesses finite third moments. Moreover, for any* $h \neq g \in [G]$, *the random variable* $R_i = \sum_{t=1}^T \epsilon_{it}(\alpha_{ht} - \alpha_{gt})$, *with upper support* ω_R *and density* f_R, *possesses a mean residual life function* $m_R(c) = \mathbb{E}[R_i - c \mid R_i > c]$ *such that* $m_R(c_n) = o(c_n)$ *and* $f_R(c_n) = o(n^{-1/4})$ *for any sequence* $c_n \to \omega_R$ *as* $n \to \infty$.

Assumption 2 is a natural extension of Assumption 1 with panel data. The conditions in the present grouped panel data framework follow immediately from the cross-sectional analysis because group memberships are assumed to be time-invariant. In this case,

$$\hat{m}_i(\hat{m}_i - m_i)' = \left(e_{\hat{g}_i}(e_{\hat{g}_i} - e_{g_i})'\right) \otimes \mathrm{I}_T,$$

so that the time-aggregated misclassification errors are equivalent to the cross-sectional errors repeated over time. The conditions in Assumptions 2(b)-(c) leverage on that structure to generalize the analogous conditions in Assumptions 1(b)-(c). In particular, if the system is stationary, the relationships simplify to:

$$\bar{\Gamma} = \Gamma \otimes \mathbf{1}_T', \quad \bar{\Xi} = \Xi \otimes \mathrm{I}_T.$$

Assumption 2(d) reveals that the panel data framework introduces an extra, fundamental requirement: the linear projection condition $\mathbb{E}[\epsilon_i \mid a'\epsilon_i = r] = \frac{a}{\|a\|^2}r$. This is because now residual-based classifiers rely on the entire vector ϵ_i. Such a geometric property essentially restricts the error terms to the family of spherically symmetric distributions. When intersected with the light-tail requirement, the allowed class of distributions remains broad but specific. It perfectly accommodates the isotropic Multivariate Normal distribution, the Spherical Laplace distribution, and variables distributed uniformly over a T-dimensional hypersphere. However, heavy-tailed spherical distributions (e.g., Multivariate Student-t or Multivariate Cauchy) or non-spherical light-tailed distributions are ruled out.[9] If the error vector does not fulfill Assumption 2(d), the characterization of the asymptotic distribution of $\hat{\psi}$ remains unchanged, yet it may

[9]By Theorem 5.3.3 of Kagan, Linnik and Rao (1973), if the errors $\epsilon_{i1}, \ldots, \epsilon_{iT}$ are mutually independent, the condition that the conditional expectation $\mathbb{E}[\epsilon_{it} \mid a'\epsilon_i = z]$ is linear and equal to the orthogonal projection $a_t z/\|a\|^2$ holds if and only if the errors are Gaussian. Therefore, we must relax independence across T if we want to encompass the broader class of spherically symmetric distributions. In other words, for non-Gaussian errors, independence across T breaks the spherical symmetry required by the linear projection condition.

not be possible to cleanly characterize the exact bias, hence the relative bias reduction, of residual-based classifiers.

More generally, Assumption 2 is meant to capture common scenarios faced by practitioners, where groups are close such that the distance between time paths of fixed-effects is small relative to the error variance, or panels have a short time dimension such that the accumulation of information over time is insufficient to perfectly distinguish the groups.

Similarly to Theorem 1, we obtain the following result.

Theorem 2 *Under Assumptions 2(a)-(c), as* $n \to \infty$,

$$\sqrt{n}(\hat{\psi} - \psi) \xrightarrow{d} N\left(-\mathbb{E}[Z_i Z_i']^{-1} \left\{ \begin{bmatrix} \Omega \otimes \mathrm{I}_T & 0 \\ \bar{\Gamma} & 0 \end{bmatrix} \psi - \bar{\Xi} \right\}, V\right),$$

where $V = \mathbb{E}[Z_i Z_i']^{-1}\mathbb{E}[Z_i \epsilon_i \epsilon_i' Z_i']\mathbb{E}[Z_i Z_i']^{-1}$. *Moreover,*

$$\hat{V} = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{Z}_i \hat{Z}_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \hat{Z}_i \hat{\epsilon}_i \hat{\epsilon}_i' \hat{Z}_i'\right) \left(\frac{1}{n}\sum_{i=1}^{n} \hat{Z}_i \hat{Z}_i'\right)^{-1} \xrightarrow{p} V.$$

Theorem 2 confirms that the bias mechanism identified in the cross-sectional model persists in the panel setting. The estimator $\hat{\psi}$ is consistent and asymptotically normal but shifted by a bias that reflects the time-aggregated misclassification errors, scaled by the within-group variation over time.

In the panel data framework, the normalized ratio ρ_{hg} becomes a T-dimensional vector, with generic element $\rho_{hg,t} = \mathbb{E}[\epsilon_{it}|\hat{g}_i = h, g_i = g]/(\alpha_{ht} - \alpha_{gt})$. While the notation is meant to encompass great generality, we may still assume the unit interval bounds on each $\rho_{hg,t}$ and the rule of thumb $\rho_{hg,t} = 1/2$. The latter is again justified by arguably the most popular residual-based classifier, the GFE estimator of Bonhomme and Manresa (2015), and implies that $\rho_{hg,t}$ is constant across h, g, t.

Therefore, with some abuse of notation:

Corollary 4 *Under Assumption 2, as* $n \to \infty$, *the asymptotic bias of* $\hat{\beta}$ *is*

$$ABias(\hat{\beta}) = \sum_{a=1}^{G} \sum_{b \neq a} \kappa_{ab} q_{a|b},$$

$$\kappa_{ab} = \left(\sum_{g=1}^{G}\sum_{t=1}^{T} \pi_g \mathrm{Var}(x_{it}|g_i = g)\right)^{-1} \sum_{t=1}^{T} \pi_b (1 - \rho_{ab,t})(\mathbb{E}[x_{it}|g_i = a] - \mathbb{E}[x_{it}|g_i = b])(\alpha_{at} - \alpha_{bt}).$$

Corollary 4 reveals that, in the panel context, the bias is driven by the interplay between misclassification rates and the time-averaged contrasts in group-specific parameters. The time dimension averages the sensitivity weights, meaning that while transient misclassification might be dampened, systematic classification errors across time will continue to distort the structural estimates. Like in the cross-sectional case, when the model has two groups we can easily characterize the sign of the bias.

Corollary 5 *Suppose* $G = 2$, $K = 1$ *and* $\rho_{hg,t}$ *is time-invariant. The asymptotic bias of* $\hat{\beta}$ *is positive if and*

only if

$$\sum_{t=1}^{T} \left(\mathbb{E}[x_{it}|g_i = 1] - \mathbb{E}[x_{it}|g_i = 2]\right)(\alpha_{1t} - \alpha_{2t}) > 0.$$

The result in Corollary 5 demonstrates that a necessary and sufficient condition for the bias of $\hat{\beta}$ to be positive is that the time-path of the between-group mean regressor difference points in the same direction (in an inner-product sense) as the time-path of grouped fixed-effects contrasts. In particular, Group 1 must have higher fixed-effects than Group 2 specifically in those time periods where it also has a higher regressor value than Group 2. One can interpret this condition as synchronized divergence: Imagine Group 1 is "high skill" and Group 2 is "low skill", then positive bias occurs if in years where the education gap widens (changes in x), the ability gap also widens (changes in α). This structure also reveals that time may serve as a bias cancellation mechanism: If the group with the larger fixed-effects has larger covariates in period 1 but smaller covariates in period 2, the net bias from this specific misclassification flow could be zero, even if classification errors persist in both periods. This is yet another reason why to expect long panels to exhibit small biases.

The inner product structure in Corollary 4 also hints that the asymptotic bias of $\hat{\beta}$ is crucially limited by the strength of group separation, as Corollary 6 illustrates. Such a separation is only relevant to the extent that it may lead to misclassification, since at the extreme of strong separation $q_{a|b} = 0$, and at the extreme of no separation $\|\alpha_a - \alpha_b\| = 0$. The intermediate regime is the one empirically relevant and the one mainly addressed by this paper.

Corollary 6 *Under Assumption 2, as $n \to \infty$, the asymptotic bias of $\hat{\beta}$ satisfies*

$$ABias(\hat{\beta}) = O\left(\max_{a,b} q_{a|b} \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\alpha_{at} - \alpha_{bt})^2}\right).$$

One can also study the consequences of misclassification on group-specific objects.

Corollary 7 *Under Assumption 2, as $n \to \infty$, the asymptotic bias of $\hat{\alpha}$ is*

$$ABias(\hat{\alpha}_{at}) = -\frac{1}{\pi_a}\sum_{b \neq a} \pi_b(1 - \rho_{ab,t})q_{a|b}(\alpha_{at} - \alpha_{bt}) - (\mathbb{E}[x_{it}|g_i = a] - \mathbb{E}[x_{it}])' ABias(\hat{\beta}).$$

Corollary 7 confirms that the bias for group-specific objects is equally relevant in panel data settings. Like in the cross-sectional case, the bias is driven by two factors: a direct effect, due to accounting for the incorrect heterogeneous intercept; and an indirect effect, capturing the propagation of the slope bias into the intercepts. These two distinct channels may be equally problematic in empirical settings and cause distortions in the estimated time paths of heterogeneity.

Proposition 2 *Define*

$$\hat{\kappa}_{ab} = \left(\sum_{g=1}^{G}\sum_{t=1}^{T} \hat{\pi}_g \widehat{\mathrm{Var}}(x_{it}|\hat{g}_i = g)\right)^{-1} \sum_{t=1}^{T} \hat{\pi}_b(1 - \hat{\rho}_{ab,t})(\hat{\mathbb{E}}[x_{it}|\hat{g}_i = a] - \hat{\mathbb{E}}[x_{it}|\hat{g}_i = b])(\hat{\alpha}_{at} - \hat{\alpha}_{bt}),$$

where hat objects are the estimated or sample counterparts. Under Assumption 2, $\hat{\kappa}_{ab} \xrightarrow{p} \kappa_{ab}$.

Similar to the cross-sectional case, Proposition 2 implies that the susceptibility of panel estimates to classification error is identifiable from the data and consistently estimable. Researchers can compute the sample analog $\hat{\kappa}_{ab}$ to quantify exactly how much a hypothetical misclassification rate would translate into parameter bias, and thus learn about its direction. This will be the underlying mechanism used in the next section.

4 Robust Inference

To employ the results above and debias the estimator for the structural parameter β, we would need to replace all the population objects in Corollaries 1 and 4 with sample counterparts. This is inherently complicated due to the fact bias arises because of misclassification and depends on it. If we had access to auxiliary data and we could estimate the matrix of misclassification rates $q_{a|b}$, then the above results provide the necessary information to construct debiased estimates; see Appendix G (cf. Battaglia et al., 2024). However, the availability of such auxiliary data is not assumed in this paper.[10]

4.1 Bias-Aware Inference

The approach proposed in this paper is to construct bias-aware confidence intervals that guarantee uniformly valid inference under a worst-case misclassification scheme (cf. Armstrong and Kolesár, 2018). The general idea works as follows. By Theorems 1 and 2, the estimator $\hat{\beta}$ is normally distributed with variance that does not depend on the misclassification. Its bias, however, does depend on it. To control size in finite samples, it is necessary to subtract the largest possible bias of $\hat{\beta}$ under the null hypothesis. Under some worst-case misclassification scheme $\mathcal{Q}$, we can upper-bound the bias by

$$\overline{B} = \sup_{\mathcal{Q}} |ABias(\hat{\beta})|.$$

Once the worst-case bias has been calculated, it can be used to compute a $100(1-\alpha)\%$ CI. Because the asymptotic variance does not depend on misclassification probabilities, this can be done by using a quantile of the $|N(\overline{B}/\sqrt{\hat{V}}, 1)|$ distribution as the critical value (Armstrong and Kolesár, 2018):

$$\hat{\beta} \pm cv_\alpha\left(\overline{B}/\sqrt{\hat{V}}\right)\sqrt{\hat{V}/n},$$

where $cv_\alpha(t)$ is the $1-\alpha$ quantile of the absolute value of a $N(t,1)$ distribution.

Propositions 1 and 2 show that κ_{ab} is identified and consistently estimable; the only unknown component in the bias is due to $q_{a|b}$. Let us define the admissible set of misclassification rates:

Assumption 3 *The admissible set of misclassification rates is given by*

$$\mathcal{Q} = \left\{ q \geqslant 0 : \sum_{a \neq b} q_{a|b} \leqslant \bar{q}_b \ \forall b \in [G] \right\},$$

where $\bar{q}_b$ *are some known or calibrated bounds.*

[10] It would be interesting to explore whether the misclassification matrix can be estimated using resampling techniques (cf. Monti, Tamayo, Mesirov and Golub, 2003; Hennig, 2007; Shamir and Tishby, 2010).

The scheme in Assumption 3 supposes that we can bound the total misclassification for each group by a known or consistently estimable constant. Assuming a bound on the total misclassification rate is a practical necessity in the absence of auxiliary data, but it is also a justifiable strategy in applied work. While the exact misclassification matrix is unknown, researchers often have access to prior economic knowledge, cross-validation results, or posterior classification probabilities from the clustering algorithm used (cf. Conley et al., 2012). These internal diagnostics provide a baseline for the expected total error rate of each group. Furthermore, the proposed method allows for sensitivity analyses. A researcher can report confidence intervals for a range of plausible error rates: e.g., "if 5% of the sample is misclassified, the confidence interval adjusts accordingly." This transparently communicates the fragility of the results to classification noise, unlike the standard oracle inference. In Appendix B, I provide some empirical recommendations to calibrate the bounds $\bar{q}_b$.

Because the bias is linear in q and $\mathcal{Q}$ is convex and compact, the worst-case bias is attained at an extreme point of $\mathcal{Q}$. Crucially, the constraints are separable across groups. Therefore the optimization decomposes group by group:

$$\sup_{q\in\mathcal{Q}}\left|\sum_{b=1}^{G}\sum_{a\neq b}\kappa_{ab}q_{a|b}\right| = \sum_{b=1}^{G}\sup_{\substack{q_{a|b}\geqslant 0\\ \sum_{a\neq b}q_{a|b}\leqslant\bar{q}_b}}\left|\sum_{a\neq b}\kappa_{ab}q_{a|b}\right|.$$

This is a linear program for each b. Therefore, the sharp worst-case absolute bias is:

$$\overline{B} = \sum_{b=1}^{G}\bar{q}_b \max_{a\neq b}|k_{ab}|.$$

The above equation shows that worst-case misclassification bias out of Group b concentrates all local misclassification mass on the single destination Group a that generates the largest bias per unit of misassignment. This bound is sharp, and allows us to construct robust inference as the next theorem shows.

Theorem 3 *Let $c \in \mathbb{R}^K$ be a vector of interest. Define the worst-case bias for the estimate of $c'\beta$ as $\hat{\overline{B}}_c = \sum_{b=1}^{G}\bar{q}_b \max_{a\neq b}|c'\hat{k}_{ab}|$, and let $\hat{V}_c = c'\hat{V}_\beta c$, where $\hat{V}_\beta$ is the block of $\hat{V}$ corresponding to β. Define*

$$\mathcal{C}_c = \left[c'\hat{\beta} - \chi, c'\hat{\beta} + \chi\right], \quad \text{where } \chi = cv_\alpha\left(\hat{\overline{B}}_c/\sqrt{\hat{V}_c}\right)\sqrt{\hat{V}_c/n}.$$

Let $\mathcal{P}_n$ denote the set of distributions P for the data $(y_i, x_i, \hat{g}_i)$ that satisfy Assumptions 1 (or 2 for a panel model) and 3. Then,

$$\liminf_{n\to\infty}\inf_{P\in\mathcal{P}_n} P(c'\beta \in \mathcal{C}_c) \geqslant 1-\alpha.$$

Theorem 3 establishes that the proposed confidence interval $\mathcal{C}_c$ is uniformly valid over a large class of DGPs characterized by local misclassification errors. The uniformity framework in this paper is constructed by fixing the underlying distribution of the true data (outcomes, covariates, and true group memberships) while allowing the conditional distribution of the estimated group memberships, namely the specific patterns of misclassification, to vary across all admissible configurations defined by $\mathcal{Q}$. In other words, the space of distributions $\mathcal{P}_n$ is indexed entirely by the finite-dimensional parameter vector q. This decoupling ensures that the asymptotic results are not driven by a specific, favorable error structure but hold for the entire class of classification algorithms with high yet potentially imperfect accuracy. For empirical

practice, this implies that researchers do not need to know the precise transition probabilities, i.e., exactly which groups are being confused with one another; they only need to defend a credible upper bound on the total misclassification rate for each group. The resulting inference is robust: the confidence intervals automatically account for the worst-case bias scenario compatible with those bounds, guaranteeing correct coverage even if the unknown misclassification structure is the most damaging one possible.

Under the admissible set $\mathcal{Q}$, the worst-case bias is finite, analytic, and easy to estimate. Thus, robust confidence intervals can be constructed with no numerical optimization. The confidence interval length scales linearly with group-wise misclassification intensities $\bar{q}_b$ and maximal group-to-group bias slopes $|\hat{\kappa}_{ab}|$. Furthermore, despite the latter are estimated on the observed misclassified units, the confidence interval remains valid.

It is important to emphasize that the potential conservatism of the proposed robust inference is strictly bounded by the ambiguity of the classification errors, not the structural parameters of the model. While the misclassification rates $q_{a|b}$ are treated as unknown nuisance parameters confined within the set $\mathcal{Q}$, the sensitivity of the estimator to these errors, captured by κ_{ab}, is consistently estimated (Propositions 1 and 2). This distinction is crucial for the practical utility of the method. In generalized robust inference problems, one often faces ambiguity about both the magnitude of the error and the local curvature of the moment condition. Here, because the observed group means and fixed-effects converge to their population counterparts despite the local misclassification, we do not need to take a worst-case supremum over the values of κ_{ab}. Instead, we treat the sensitivity vectors as fixed at their sample estimates $\hat{\kappa}_{ab}$. This effectively shrinks the optimization problem: the confidence interval is widened only enough to account for the unknown permutation of units, without paying an additional penalty for the estimation uncertainty of the structural slopes or group contrasts. Consequently, the resulting intervals are significantly narrower than they would be under a fully agnostic worst-case bound where all data moments were treated as uncertain.

There exists a specific instance where inference is guaranteed to be asymptotically exact:

Proposition 3 *Under the conditions of Theorem 3, if $G = 2$ and $(\bar{q}_1, \bar{q}_2)$ satisfy oracle calibration, then*

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}_n} P(c'\beta \in \mathcal{C}_c) = 1 - \alpha.$$

Proposition 3 highlights a fundamental appeal of binary grouping. The result proves that in the case of two groups (e.g., treatment vs. control, or movers vs. stayers), the robust confidence interval is tight: the worst-case bias bound used to construct the interval coincides perfectly (in absolute value) with the actual bias in the data's least favorable misclassification configuration. Consequently, researchers working in a setting with two groups do not trade off efficiency for robustness, as inference remains sharp. This is true under oracle calibration, which represents a scenario where the researcher knows a sufficiently accurate estimate of the misclassification bound $\bar{q}_b$. This is reasonable in a setting with only two groups because such a bound equals the misclassification probability, i.e., units can only be misclassified to a unique group, hence the sum in the definition of $\mathcal{Q}$ is redundant.

The driving mechanism underlying Proposition 3 is that all the sensitivity weights κ_{ab} automatically have the same sign when $G = 2$, which guarantees that the absolute value of the sum of bias contributions from each group equals the sum of the absolute value contributions from each group. More generally,

as G increases, the assumption that all misclassification biases reinforce each other becomes increasingly unlikely due to the triangle inequality logic. In practice, biases between different pairs of groups will often partially cancel out, meaning that for $G > 2$, both oracle calibration and sharpness are harder to guarantee.

4.2 Debiasing Through Different Classifiers

When external proxy data are available, it is common to rely on them for the classification step. The results in this paper show that doing so may become undesirable because using residual-based classifiers automatically dominates in terms of mean squared error: they achieve a bias reduction while the variance remains the same. Importantly, the latter argument crucially hinges on the condition that the accuracy of classification in either case is comparable. In other words, the signal-to-noise ratio in either source of group information must be similar, so that the implied misclassification rates $q_{a|b}$ are the same. Under the condition of equivalent misclassification rates, and using the rule of thumb for residual-based classifiers of $\rho = 1/2$, Proposition A.2 shows that the combined estimator $\tilde{\beta} = 2\hat{\beta}_R - \hat{\beta}_P$ is asymptotically normal, unbiased and with variance V_β. While in finite samples the variance of the estimator $\tilde{\beta}$ may be larger than that of $\hat{\beta}_R$ and $\hat{\beta}_P$ due to its debiasing structure, this result implies that confidence intervals based on it are correctly centered despite misclassification. That is, using $\tilde{\beta}$ completely sidesteps the need for robust inference methods, and the standard oracle confidence bands around it turn out to guarantee valid inference.

This arguably surprising result shows that it is possible to avoid the need to estimate the matrix of misclassification by exploiting estimation results under two different classifiers. Despite classifiers are not perfect and true groups remain unknown, hence the matrix of misclassification errors cannot be learned, the estimator $\tilde{\beta}$ will not suffer from any effects due to misclassification.

Theorem 4 *Let $c \in \mathbb{R}^K$ be a vector of interest. Define*

$$\tilde{\mathcal{C}}_c = \left[c'\tilde{\beta} - \tilde{\chi}, c'\tilde{\beta} + \tilde{\chi}\right], \quad \text{where } \tilde{\chi} = z_{1-\alpha/2}\sqrt{\hat{V}_c/n},$$

where $z_{1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of a $N(0,1)$ distribution. Let $\mathcal{P}_n$ denote the set of distributions P for the data $(y_i, x_i, \hat{g}_{i,R}, \hat{g}_{i,P})$ that satisfy Assumptions 1 (or 2 for a panel model). Then, under the rule of thumb for residual-based classifiers of $\rho = 1/2$ and the assumption that residual- and proxy-based classifiers have comparable error rates,

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}_n} P(c'\beta \in \tilde{\mathcal{C}}_c) = 1 - \alpha.$$

Despite the estimator $\tilde{\beta}$ and confidence interval $\tilde{\mathcal{C}}_c$ rely on several strong assumptions to work exactly as intended, practitioners are strongly encouraged to employ them as an additional robustness check, along with the robust confidence interval described above.

5 Numerical Illustration

This section presents Monte Carlo simulations designed to evaluate the numerical performance of the proposed robust inference procedure. The analysis serves two primary objectives. First, I assess the accuracy of the local asymptotic approximations derived in Theorems 1 and 2, verifying whether the theoretical bias expressions correctly capture the distortions induced by classification error in finite samples. Sec-

ond, I systematically compare the coverage properties of standard oracle confidence intervals against the robust confidence intervals developed in Section 4. By varying the misclassification rates, I illustrate the fragility of standard post-selection inference, which rapidly loses validity as classification error increases, and demonstrate the ability of the robust procedure to maintain nominal coverage rates even in the presence of substantial classification uncertainty. In Appendix C, I provide further numerical results under alternative designs, and some additional empirical recommendations.

5.1 Varying Degrees of Misclassification

The design in this subsection focuses on the cross-sectional case and evaluates the coverage probability of the confidence intervals as the misclassification intensity increases.

I consider a sample of size $n = 250$ with $G = 2$ groups. The finite-sample distribution is approximated by 5,000 Monte Carlo replications. The true group memberships g_i are drawn uniformly at random. The DGP is given by:

$$y_i = \alpha_{g_i} + x_i\beta + \epsilon_i,$$

where the slope is $\beta = 2$ and the idiosyncratic error is $\epsilon_i \sim N(0, \sigma^2)$. The regressors x_i are generated conditionally on the true group membership: $x_i \sim N(\mu_{g_i}, 4)$. Crucially, to induce misclassification bias, I specify distinct fixed-effects and regressor means across groups, and set $\alpha = (-1.5, 2)'$ and $\mu = (-1, 1)'$. This separation ensures that group misclassification generates correlation between the regressor and the effective error term. I set $\sigma = 2$ to control the signal-to-noise ratio, allowing for a clearer comparison of robust versus naive inference approaches, and capturing the scheme addressed in this paper where bias is of the same order as the standard error; see Figure I.

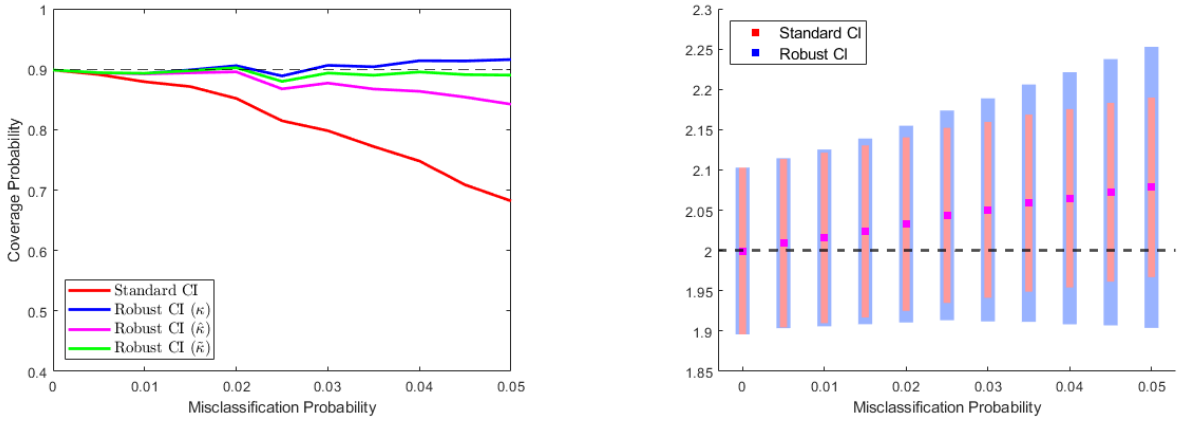
Following the asymptotic framework, I model local misclassification where the probability of error scales with sample size. I define the misclassification probability $p = q/\sqrt{n}$. I consider a cyclic misclassification structure where units in group g are misclassified into group $(g \mod 2) + 1$ with probability p. This represents a random (non-adversarial) error structure, hence $\rho_{hg} = 0$. Since there are only two groups in this setup, units can only be misclassified to the group they are not truly assigned.

I compare two methods. On the one hand we have standard inference: standard least squares estimation of $\hat{\beta}$ using the estimated group memberships $\hat{g}_i$, with conventional standard errors. On the other hand we have robust inference: the fixed-length confidence interval described in Section 4.

Figure II compares standard versus robust inference 90% confidence intervals as a function of the group misclassification probability p, ranging from 0% to 5%, i.e., for different values of q. The left plot shows the difference in coverage probabilities, while the right plot shows the average confidence interval (across Monte Carlo iterations).

The results in Figure IIa highlight two key findings. First, we observe a clear failure of standard inference. The coverage of the standard inference method (red line) deteriorates monotonically. Even with moderate noise, misclassification rates as low as 5% for each group create a bias sufficient to distort the test statistics,[11] causing coverage to drop significantly below the nominal level. This confirms that small deviations from perfect classification can invalidate standard inference results. Second, robust inference is able to restore the correct coverage by effectively identifying the maximum possible bias consistent with

[11] See Figure I for the corresponding finite-sample distributions of $\hat{\beta}$.



(a) Coverage Probability

(b) Average CI

Figure II: Standard versus robust inference for different degrees of misclassification

the local misclassification error rate q and adjust the length of the confidence interval accordingly. The blue line in Figure IIa represents the infeasible confidence interval that uses the true sensitivity weights κ. In practice, these weights must be replaced by its estimate $\hat{\kappa}$. While theoretically a valid replacement according to Propositions 1 and 2, in small samples these estimated weights may slightly undercover for moderate misclassification rates due to the attenuation of group-specific estimates based on misclassified groups (magenta line). One successful solution is to inflate the estimated bias bound based on the expected misclassification to offset the attenuation (green line), which brings back the empirical coverage to the nominal level and guarantees valid inference robust to misclassification. In Appendix C, I discuss in more detail the attenuation bias in $\hat{\kappa}$ and the inflation mechanism. In the current setting, inflation becomes more relevant because we are using oracle calibration for the misclassification bounds. The results in Appendix C suggest that, in practice, the potential conservativeness in the empirical calibration of the bounds will likely offset the attenuation bias in the estimates of the sensitivity weights such that the empirical coverage of the robust confidence intervals is automatically close to the nominal for any misclassification rate.

The linear decrease in the coverage probability of standard inference observed in Figure II is a direct consequence of the bias expression's linearity in the misclassification rates: as the error rate increases, the drift of the estimator's distribution shifts linearly, moving the true parameter further into the tails of the standard confidence interval, as observed in Figure IIb as well. The steepness of the red line depends on the bias-to-noise ratio, being increasingly larger as misclassification bias further dominates the noise. This emphasizes the importance of the sensitivity weights κ_{ab} and further motivates Propositions 1 and 2.

Figure IIb demonstrates that under no misclassification—both actual and expected—, standard and robust confidence intervals are equivalent. In either case, the confidence interval is centered at the least squares estimate (magenta square), which increases with misclassification hinting at a positive bias. A clear message from the graph is that only the robust confidence interval adjusts its length along with misclassification to guarantee the right coverage. This confirms that by explicitly incorporating the maximal asymptotic bias into the critical values, the proposed procedure effectively immunizes inference against the failure of the oracle property.

5.2 GFE and its non-oracle performance

To evaluate the finite-sample performance of the proposed robust inference method, I also conduct a Monte Carlo experiment in a panel data setting. This setup departs from the static cross-sectional model to illustrate that the theoretical results hold in more complex environments where group membership is estimated from short time series.

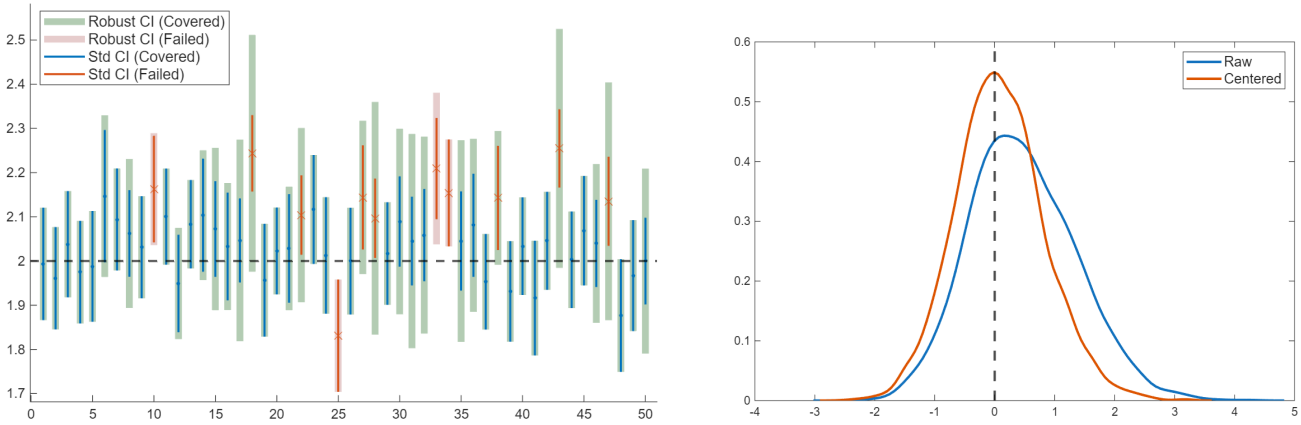
I consider a linear panel data model with grouped patterns of time-varying heterogeneity:

$$y_{it} = \alpha_{g_i t} + x_{it}\beta + \epsilon_{it}, \quad i \in [n],\ t \in [T]\ g \in [G].$$

I set $n = 80$, $T = 10$ and $G = 2$, exhibiting the short panel structure commonly found applications. The finite-sample distribution is approximated by 5,000 Monte Carlo replications. The structural parameters are generated as follows. The common slope parameter is set to $\beta = 2$. To induce realistic heterogeneity patterns between groups, the group-specific time paths follow random walk processes separated by a fixed distance. Specifically, $\alpha_{gt} = \alpha_{g,t-1} + \eta_{gt}$ with $\eta_{gt} \sim N(0,1)$. To ensure the groups are locally identifiable but prone to misclassification, I set the initial conditions to -0.5 and 0.5 for each group, respectively. To generate bias, the regressor must be correlated with group membership, so I generate $x_{it} \sim N(\alpha_{g_i t}, 1)$. The error term is $\epsilon_{it} \sim N(0, \sigma^2)$, with $\sigma = 2$. The combination of a short time horizon, proximate group trends, and a moderate noise level is calibrated to generate an observed total misclassification rate around 6%, which is the value found in the empirical application in Section 6.

I estimate the group memberships $\hat{g}_i$ and the structural parameters $(\hat{\beta}, \hat{\alpha})$ using the GFE estimator by Bonhomme and Manresa (2015), hence $\rho_{hg,t} = 1/2$. I compare two confidence intervals: the standard confidence interval, treating the estimated groups $\hat{g}_i$ as the truth, and the robust confidence interval defined in Theorem 3, adapted for the panel setting.

Constructing the robust interval requires defining the sensitivity parameters κ_{ab} and the admissible set $\mathcal{Q}$ for the panel context. The sensitivity parameter κ_{ab} is computed using the estimated $\hat{\kappa}_{ab}$ in Corollary 4. I exploit the fact that true groups are known to us to calibrate the bounds $\bar{q}_b$ using the realized misclassification rates from the simulation (an oracle calibration).



(a) Subsample of CIs

(b) Finite-sample distribution of $\hat{\beta}$

Figure III: Standard versus robust inference under small T

The results in Figure III show that standard inference breaks down under this design. Figure IIIa plots the

confidence intervals based on the standard and robust method for a subsample of Monte Carlo replications. We can see that the robust intervals are often wider than the standard counterparts, only enough to cover the true parameter value and retain nominal coverage. Because the estimated groups contain errors correlated with the regressor, $\hat{\beta}$ suffers from bias. Figure IIIb shows that the finite-sample distribution of the GFE estimator $\hat{\beta}$ is slightly shifted to the right (blue line). First, this confirms once more the results in Theorem 2 where misclassification largely affects the center but not the dispersion of the estimator; the red line is the infeasible debiased distribution based on the sample estimate of the asymptotic bias formula. Second, one can verify that in this design the necessary and sufficient condition for positive bias in Corollary 5 indeed holds. Third, the shift in the distribution of $\hat{\beta}$ causes the coverage of the standard 90% confidence interval to drop significantly until 68.1%. In contrast, the robust confidence interval successfully restores coverage to the nominal level by accounting for the worst-case bias within the local neighborhood of the estimated groups, yielding 89.4% coverage. This confirms that consistent inference is achievable even when T is small and misclassification is prevalent.

A clear takeaway from Figure III is that standard confidence intervals for structural parameters appear highly precise due to the large sample size, yet this precision is illusory as it ignores the uncertainty in latent group recovery. While the proposed robust confidence intervals are necessarily wider, they remain informative and, in cases of low sensitivity, have the same length as standard intervals, providing rigorous assurance rather than false confidence.

6 Latent Financial Constraints: Re-evaluating Investment-Cash Flow Sensitivities

A central question in corporate finance is the extent to which investment is driven by the availability of internal funds (cash flow) after controlling for investment opportunities (Tobin's Q); see, e.g., Fazzari, Hubbard and Petersen (1988); Kaplan and Zingales (1997). A positive sensitivity is often interpreted as evidence of financial constraints.

The standard literature typically pre-classifies firms into "Constrained" and "Unconstrained" groups using ex-ante observable proxies (e.g., firm size, dividend payout ratio, or bond ratings). The problem is that these observable proxies are imperfect (Moyen, 2004). A small firm might be flush with venture capital (unconstrained), while a large firm might be highly leveraged and distressed (constrained).

The constraint status is a latent state derived from behavior, not just size (Farre-Mensa and Ljungqvist, 2016; Lian and Ma, 2021). Many firms likely lie in the "grey area" of financial health. Standard oracle theory assumes we can perfectly separate these firms asymptotically. However, with a short panel (e.g., 10-15 years) and many firms on the margin, misclassification is inevitable. In this section I employ the proposed method to do robust inference on the cash flow sensitivity by accounting for the possible bias introduced by these marginal firms.

I use a balanced panel of U.S. manufacturing firms for the period 2000-2015, obtained from Compustat; see Appendix D for a description of the dataset construction. This yields a panel of $n = 917$ firms and $T = 16$ time periods. The specification follows the standard Q-theory of investment with grouped heterogeneity:

$$I_{it} = \alpha_{g_i t} + \beta_1 Q_{it-1} + \beta_2 CF_{it} + \epsilon_{it},$$

where I_{it} is the ratio of investment to lagged capital, Q_{it-1} is the lagged Tobin's Q, CF_{it} is the ratio of cash flow to lagged capital, and $\alpha_{g_i t}$ captures group-specific time trends. The model considered allows, for example, constrained firms to have a different trend during the 2008 crisis than unconstrained firms, which a simple firm or time fixed-effects specification cannot capture.

I set the number of groups to $G = 5$ following the classification scheme in Kaplan and Zingales (1997): not, likely not, potentially, likely, and financially constrained.[12] The parameter of interest is β_2, capturing the sensitivity of investment to cash flow. If misclassification correlates with cash flow (which is likely, as low cash flow firms are harder to classify), the standard least squares estimator is biased. More generally, this is because fixed-effects specifically capture unobserved heterogeneity, such as managerial quality or unmeasured productivity shocks, that is likely correlated with input choices.

In the baseline specification, I impose homogeneity on the slope parameter β_2 while allowing for heterogeneous group-level intercepts α_{gt}.[13] This modeling choice is motivated by two primary considerations. First, it allows us to recover the average structural sensitivity across the entire corporate sector. If we were to allow β_2 to vary by group, the resulting estimates would be susceptible to the critique raised by Moyen (2004), as the differences in sensitivities might simply reflect the small sample properties of specific clusters rather than a generalized financing friction. Second, by allowing for group-specific intercepts, we effectively purge the estimate of unobserved factors, such as management quality or industry-specific cycles, that are common to firms within a latent group.[14] Thus, the present approach provides a conservative and robust estimate of the role of internal funds that is less sensitive to the noise of individual group dynamics.

The standard practice is to use the qualitative information in the annual reports, together with quantitative information in the companies' financial statements and notes, to classify each firm into one group (see, e.g., Kaplan and Zingales, 1997; Hadlock and Pierce, 2010; Campello, Graham and Harvey, 2010; Hoberg and Maksimovic, 2015). These classification schemes are arguably subject to error if relevant proxies are not accounted for or managers do not always report truthfully (Moyen, 2004; Bakke and Whited, 2010). Moreover, these approaches are labor-intensive and difficult to scale across extensive panel datasets. The estimation approach I use recovers these latent groups statistically and endogenously from the data. This allows for the identification of unobserved heterogeneity at the scale of the full Compustat universe, effectively circumventing the trade-off between the depth of qualitative classification and the breadth of quantitative analysis. Moreover, it enjoys the additional automatic bias reduction due to endogenous misclassification patterns.

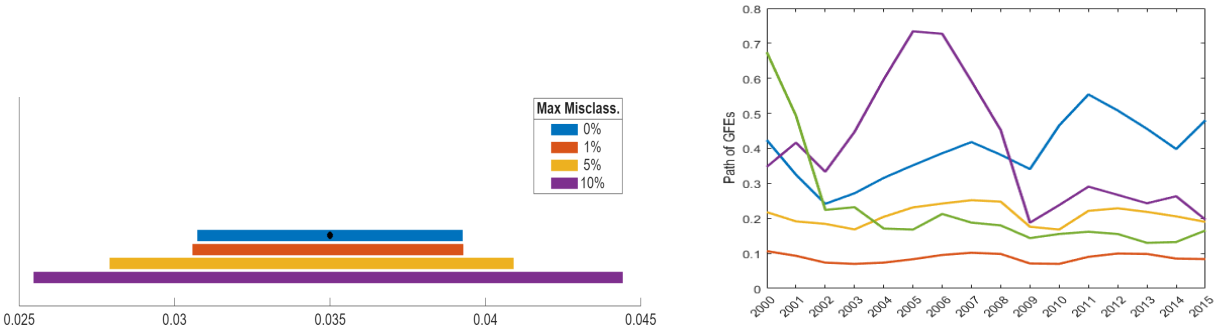
Figure IV plots the inference objects based on the results in this paper. The GFE estimator for β_2 is 0.0356, represented by a black dot in Figure IVa. The colored bars represent different 95% confidence intervals for β_2 under various levels of potential group misclassification. In all cases, the maximum misclassification $\bar{q}_b$ is set equal to a constant across groups.

Since the total sample size is large, the confidence intervals in Figure IVa are relatively narrow. When we assume no misclassification errors, the standard confidence interval coincides with the robust one. How-

[12] This is more groups than the literature typically allows for. For our purposes, this offers further robustness as González-Casasús (2025) shows. In this context where $n \gg T$, information criteria are inconsistent, but in Appendix E I explore other approaches to assess the choice of the number of groups.

[13] See Appendix F for an analysis of a model with group-specific slopes, and Appendix E for different robustness checks.

[14] The grouped fixed-effects approach might also partially alleviate the Erickson and Whited (2000, 2012) measurement error critique by controlling for unobserved, group-level investment opportunity trajectories.



(a) Robust confidence intervals for β_2

(b) Time paths of unobserved heterogeneity

Figure IV: Robust inference on cash flow sensitivity

ever, group memberships are only estimated on the information from a small time dimension. As we start to acknowledge potential misclassification, the robust confidence intervals increasingly widen. Overall, the confidence intervals include small yet positive values of cash flow sensitivity. I find that zero does not belong to the robust confidence interval even after allowing for unrealistically large misclassification, which provides sufficient evidence to conclude that there is indeed positive cash flow sensitivity. However, while statistically significant, such sensitivity is economically modest.

Following the empirical recommendations in Appendix B, I also calibrate the misclassification bound in a data-driven way, namely using the relative distances approach. Doing so, I obtain the following estimated misclassification bound:

$$\hat{\tilde{q}} = (1.14, 1.55, 3.16, 1.07, 2.27).$$

This calibrated bound, based on the classification output from the underlying k-means algorithm, shows that one should indeed allow for different levels of potential misclassification across groups. In particular, the corresponding maximal misclassification probabilities are roughly $4\%, 5\%, 10\%, 4\%, 8\%$, respectively. Based on the data-driven bound, the implied robust confidence interval for β_2 is $[0.028, 0.043]$, whereas the standard oracle interval is $[0.031, 0.039]$. The calibrated robust interval roughly coincides with the yellow interval in Figure IVa. The results seem consistent because the average misclassification rate implied by $\hat{\tilde{q}}$ is 6%.

Figure IVb depicts the grouped time paths of unobserved heterogeneity captured by the fixed-effects. We indeed observe large diversity of behavior across groups of firms, especially around the 2008 crisis. The observed heterogeneity in these paths validates the classification approach. For instance, there is a group that clearly exhibits a sharper decline during the 2008 financial crisis. This asymmetric response may identify it as the financially constrained cluster. In contrast, the group represented by a blue line sees an important increase in the latent heterogeneity following the 2008 crisis, hinting that this is likely the unconstrained group. At the same time, groups represented in yellow and orange (or even green) follow similar paths, which could result in more likely misclassification of their units in finite samples, stressing the relevance of the results derived in this paper. Indeed, the calibrated bound $\hat{\tilde{q}}$ shows that Group 3 (yellow line in Figure IVb) is potentially suffering the most misclassification. All in all, Figure IVb confirms that the estimated groups reflect structural economic differences in resilience to aggregate shocks.

To understand the potential effects of misclassification for estimation and inference on β_2, it is helpful to

explore the corresponding estimated sensitivity weights matrix with generic element κ_{ab}:

$$\hat{\kappa} = \frac{1}{2} \times 10^{-2} \times \begin{pmatrix} 0 & 5.05 & 1.77 & -0.03 & 0.31 \\ 0.53 & 0 & -0.01 & 0.1 & 0.83 \\ 0.33 & -0.02 & 0 & 0.05 & 0.48 \\ -0.04 & 1.26 & 0.4 & 0 & 0.07 \\ 0.18 & 4.41 & 1.46 & 0.03 & 0 \end{pmatrix}.$$

The structure of the estimated misclassification sensitivity matrix $\hat{\kappa}$ provides crucial insights into the mechanics of the bias correction. The first factor, 1/2, comes from the residual-based nature of the classifier and captures the automatic bias reduction. The matrix factor reveals a striking asymmetry: the bias is predominantly driven by the second column. Specifically, the large positive coefficients $\hat{\kappa}_{12}$, $\hat{\kappa}_{42}$ and $\hat{\kappa}_{52}$ indicate that misclassifying units from Group 2 (orange line in Figure IVb) into Group 1 (blue line), Group 4 (purple line) or Group 5 (green line) introduces substantial upward bias into the estimates. In contrast, misclassification from other groups (e.g., Groups 1 or 4) yields negligible sensitivity, with coefficients close to zero or even negative. This pattern suggests that Group 2 likely consists of firms with distinct structural characteristics, such as high unobserved volatility or aggressive investment shocks, that, when conflated with other types, may artificially inflate the cash-flow sensitivity. Consequently, robust estimation and inference needs to account for these specific, high-cost classification errors associated with Group 2. The combination of the calibrated misclassification rates $\hat{\bar{q}}$ and sensitivity weights $\hat{\kappa}$ yields an estimated maximum bias $\hat{\bar{B}}/\sqrt{n}$ for $\hat{\beta}_2$ of 0.004, which represents a relative bias of approximately 10%. This magnitude is non-negligible and emphasizes the need for robust inference methods.

The finding of an economically modest cash flow sensitivity stands in contrast to the higher estimates often reported in the literature, but this discrepancy highlights the value of the robust data-driven approach employed. As demonstrated by Moyen (2004), regression results in this domain are often sensitive to the arbitrary ex-ante criteria used to classify firms. By identifying latent groups endogenously, I sidestep this classification critique, ensuring the results are not an artifact of proxy selection. Furthermore, by estimating a homogeneous slope parameter β_2, I capture the representative sensitivity of the average firm in the sample.

The sharp reduction in the magnitude compared to standard least squares estimates reported in the literature may have a dual source, as the additional empirical results in Appendix E further demonstrate. First, the commonly observed sensitivity is largely driven by complex forms of latent heterogeneity which standard methods used in this literature, such as individual or time fixed-effects are unable to fully capture. Once this latent structure is properly accounted for via the robust inference procedure, the residual role of internal funds effectively vanishes. This is in line with the recent results by Leng, Sojli, Tham and Wang (2025), who show that allowing for rich latent heterogeneity patterns in finance applications proves to be crucial. Second, comparing the more recent 2000-2015 sample used here versus the earlier samples used in the literature, such as the 1970-1984 period, makes it evident that sensitivities have diminished in recent years. But even in earlier periods, the robust inference method is less optimistic about the role of cash flow on investment than historically reported.

7 Conclusion

This paper exposes a fundamental fragility in the estimation of models with latent group structures. While the oracle property has long reassured practitioners that classification errors can be safely ignored, I show that this result is deceptive in practice. In finite samples, even high-accuracy classifiers induce a systematic misclassification bias that is empirically relevant. This bias invalidates conventional inference, leading to spurious significance and distorted policy conclusions.

I introduce a local asymptotic framework that formally characterizes this non-standard misclassification bias, showing it depends on the subtle interplay between group separation, misclassification probabilities, and covariate distributions. Leveraging this theory, I develop a robust inference procedure that is straightforward to implement and requires no auxiliary data to estimate the misclassification matrix. Unlike standard methods that undercover when groups are weakly separated or the time dimension of the panel is small, the proposed confidence intervals guarantee validity by adjusting their length to account for the bias under a worst-case misclassification scheme. In some special instances, combining information from different (potentially imperfect) classifiers can automatically solve the distortions due to misclassification.

The empirical relevance of these contributions is underscored by the evaluation of investment-cash flow sensitivities during the 2008 financial crisis. Where standard empirical practice often reports large sensitivities by relying on ad hoc, ex-ante firm classifications, the proposed framework reveals a more nuanced reality. By modeling constraint status as an endogenous latent structure and adjusting inference for potential misclassification flows, I demonstrate that the average structural reliance on internal funds is economically modest. Although the robust confidence intervals strictly exclude zero, confirming a positive role for cash flow, the sharp reduction in the estimated magnitude suggests that historical estimates may be substantially inflated by either the systematic misclassification of firm constraint status, latent constraint-status and time-specific heterogeneity, or both. Moreover, cash flow sensitivities appear to have diminished in recent years.

The results in this paper may serve as a cautionary tale for a larger class of models. In an era of increasing reliance on machine learning for economic measurement, treating estimated labels as data is a recipe for false discoveries. By providing a tool to quantify and correct for this uncertainty, this paper paves the way for more credible inference. Future work may extend this framework to nonlinear models or dynamic discrete choice settings, where misclassification issues are equally pervasive. Moreover, the current results do not apply to classifiers that directly partition the included regressor space. In such cases, the misclassified subpopulation is systematically located near decision boundaries, rendering the unconditional mean substitution used here invalid and leaving the maximum bias potentially unbounded without auxiliary data.

References

Abadie, A. and Imbens, G. W. (2006), 'Large sample properties of matching estimators for average treatment effects', *econometrica* **74**(1), 235–267.

Aigner, D. J. (1973), 'Regression with a binary independent variable subject to errors of observation', *Journal of Econometrics* **1**(1), 49–59.

Almeida, H., Campello, M. and Weisbach, M. S. (2004), 'The cash flow sensitivity of cash', *The journal of finance* **59**(4), 1777–1804.

Andrews, I., Kitagawa, T. and McCloskey, A. (2021), 'Inference after estimation of breaks', *Journal of Econometrics* **224**(1), 39–59.

Andrews, I., Kitagawa, T. and McCloskey, A. (2024), 'Inference on winners', *The Quarterly Journal of Economics* **139**(1), 305–358.

Armstrong, T. B. and Kolesár, M. (2018), 'Optimal inference in a class of regression models', *Econometrica* **86**(2), 655–683.

Armstrong, T. B., Weidner, M. and Zeleneev, A. (2022), 'Robust estimation and inference in panels with interactive fixed effects', *arXiv preprint arXiv:2210.06639* .

Bajari, P., Nekipelov, D., Ryan, S. P. and Yang, M. (2015), 'Machine learning methods for demand estimation', *American Economic Review* **105**(5), 481–485.

Bakke, T.-E. and Whited, T. M. (2010), 'Which firms follow the market? an analysis of corporate investment decisions', *The Review of Financial Studies* **23**(5), 1941–1980.

Battaglia, L., Christensen, T., Hansen, S. and Sacher, S. (2024), 'Inference for regression with variables generated from unstructured data', *arXiv preprint arXiv:2402.15585* .

Bishop, C. M. and Nasrabadi, N. M. (2006), *Pattern recognition and machine learning*, Vol. 4, Springer.

Bonhomme, S. and Denis, A. (2025), 'Fixed effects and beyond. bias reduction, groups, shrinkage and factors in panel data', *Documentos de trabajo-Banco de España* (26), 1.

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

Bonhomme, S. and Weidner, M. (2022), 'Minimizing sensitivity to model misspecification', *Quantitative Economics* **13**(3), 907–954.

Breiman, L. (2001), 'Random forests', *Machine learning* **45**(1), 5–32.

Campello, M., Graham, J. R. and Harvey, C. R. (2010), 'The real effects of financial constraints: Evidence from a financial crisis', *Journal of financial Economics* **97**(3), 470–487.

Cattaneo, M. D., Jansson, M. and Newey, W. K. (2018), 'Inference in linear regression models with many covariates and heteroscedasticity', *Journal of the American Statistical Association* **113**(523), 1350–1361.

Chesher, A. (1991), 'The effect of measurement error', *Biometrika* **78**(3), 451–462.

Conley, T. G., Hansen, C. B. and Rossi, P. E. (2012), 'Plausibly exogenous', *Review of Economics and Statistics* **94**(1), 260–272.

Deb, P. and Trivedi, P. K. (1997), 'Demand for medical care by the elderly: a finite mixture approach', *Journal of applied Econometrics* **12**(3), 313–336.

Dzemski, A. and Okui, R. (2021), 'Convergence rate of estimators of clustered panel models with misclassification', *Economics Letters* **203**, 109844.

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

Erickson, T. and Whited, T. M. (2000), 'Measurement error and the relationship between investment and q', *Journal of political economy* **108**(5), 1027–1057.

Erickson, T. and Whited, T. M. (2012), 'Treating measurement error in tobin's q', *The Review of Financial Studies* **25**(4), 1286–1329.

Evdokimov, K. S. and Zeleneev, A. (2023), 'Simple estimation of semiparametric models with measurement errors', *arXiv preprint arXiv:2306.14311* .

Fama, E. F. and French, K. R. (1993), 'Common risk factors in the returns on stocks and bonds', *Journal of financial economics* **33**(1), 3–56.

Farre-Mensa, J. and Ljungqvist, A. (2016), 'Do measures of financial constraints measure financial constraints?', *The review of financial studies* **29**(2), 271–308.

Fazzari, S., Hubbard, R. G. and Petersen, B. (1988), 'Investment, financing decisions, and tax policy', *The American economic review* **78**(2), 200–205.

Gal, Y. and Ghahramani, Z. (2016), Dropout as a bayesian approximation: Representing model uncertainty in deep learning, *in* 'international conference on machine learning', PMLR, pp. 1050–1059.

Gambacorta, L., Kwon, B., Park, T., Patelli, P. and Zhu, S. (2024), *CB-LMs: language models for central banking*, Bank for International Settlements, Monetary and Economic Department Basel ....

Gómez-Puig, M., Sosvilla-Rivero, S. and Martínez-Zarzoso, I. (2022), 'On the heterogeneous link between public debt and economic growth', *Journal of International Financial Markets, Institutions and Money* **77**, 101528.

González-Casasús, O. (2025), 'Grouped fixed-effects estimation is more robust than you think', *Working Paper* .

González-Casasús, O. and Schorfheide, F. (2025), Misspecification-robust shrinkage and selection for var forecasts and irfs, Technical report, National Bureau of Economic Research.

Gregory, V., Menzio, G. and Wiczer, D. (2025), 'The alpha beta gamma of the labor market', *Journal of Monetary Economics* **150**, 103695.

Grunewald, N., Klasen, S., Martínez-Zarzoso, I. and Muris, C. (2017), 'The trade-off between income inequality and carbon dioxide emissions', *Ecological Economics* **142**, 249–256.

Gu, S., Kelly, B. and Xiu, D. (2020), 'Empirical asset pricing via machine learning', *The Review of Financial Studies* **33**(5), 2223–2273.

Hadlock, C. J. and Pierce, J. R. (2010), 'New evidence on measuring financial constraints: Moving beyond the kz index', *The review of financial studies* **23**(5), 1909–1940.

Hansen, A. L. and Kazinnik, S. (2024), 'Can chatgpt decipher fedspeak?', *Available at SSRN 4399406* .

Heckman, J. and Singer, B. (1984), 'A method for minimizing the impact of distributional assumptions in econometric models for duration data', *Econometrica: Journal of the Econometric Society* pp. 271–320.

Hennig, C. (2007), 'Cluster-wise assessment of cluster stability', *Computational Statistics & Data Analysis* **52**(1), 258–271.

Hoberg, G. and Maksimovic, V. (2015), 'Redefining financial constraints: A text-based analysis', *The Review of Financial Studies* **28**(5), 1312–1352.

Hoberg, G. and Phillips, G. (2016), 'Text-based network industries and endogenous product differentiation', *Journal of political economy* **124**(5), 1423–1465.

Hubmer, J., Chan, M., Ozkan, S., Salgado, S. and Hong, G. (2024), Scalable versus productive technologies, Technical report, Penn Institute for Economic Research, Department of Economics, University of ....

Kagan, A. M., Linnik, I. V. and Rao, C. R. (1973), *Characterization problems in mathematical statistics*, Wiley series in probability and mathematical statistics, Wiley, New York. Translated from the Russian by B. Ramachandran.

Kaplan, S. N. and Zingales, L. (1997), 'Do investment-cash flow sensitivities provide useful measures of financing constraints?', *The quarterly journal of economics* **112**(1), 169–215.

Ke, Z. T., Fan, J. and Wu, Y. (2015), 'Homogeneity pursuit', *Journal of the American Statistical Association* **110**(509), 175–194.

Keane, M. P. and Wolpin, K. I. (1997), 'The career decisions of young men', *Journal of political Economy* **105**(3), 473–522.

Kitamura, Y., Otsu, T. and Evdokimov, K. (2013), 'Robustness, infinitesimal neighborhoods, and moment restrictions', *Econometrica* **81**(3), 1185–1201.

Leng, X., Sojli, E., Tham, W. W. and Wang, W. (2025), 'Time-varying group unobserved heterogeneity in finance', *Journal of Business & Economic Statistics* pp. 1–14.

Lian, C. and Ma, Y. (2021), 'Anatomy of corporate borrowing constraints', *The Quarterly Journal of Economics* **136**(1), 229–291.

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

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.

MacKay, D. J. (2003), *Information theory, inference and learning algorithms*, Cambridge university press.

Mantegna, R. N. (1999), 'Hierarchical structure in financial markets', *The European Physical Journal B-Condensed Matter and Complex Systems* **11**(1), 193–197.

Monti, S., Tamayo, P., Mesirov, J. and Golub, T. (2003), 'Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data', *Machine learning* **52**(1), 91–118.

Moyen, N. (2004), 'Investment–cash flow sensitivities: Constrained versus unconstrained firms', *The Journal of finance* **59**(5), 2061–2092.

Okui, R. and Wang, W. (2021), 'Heterogeneous structural breaks in panel data models', *Journal of Econometrics* **220**(2), 447–473.

Phillips, P. C. (1987), 'Towards a unified asymptotic theory for autoregression', *Biometrika* **74**(3), 535–547.

Rambachan, A. and Roth, J. (2023), 'A more credible approach to parallel trends', *Review of Economic Studies* **90**(5), 2555–2591.

Sanogo, T. (2019), 'Does fiscal decentralization enhance citizens' access to public services and reduce poverty? evidence from côte d'ivoire municipalities in a conflict setting', *World development* **113**, 204–221.

Shamir, O. and Tishby, N. (2010), 'Stability and model selection in k-means clustering', *Machine learning* **80**(2), 213–243.

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

Su, L., Shi, Z. and Phillips, P. C. (2016), 'Identifying latent structures in panel data', *Econometrica* **84**(6), 2215–2264.

Wang, W., Phillips, P. C. and Su, L. (2018), 'Homogeneity pursuit in panel data models: Theory and application', *Journal of Applied Econometrics* **33**(6), 797–815.

Wooldridge, J. M. (2010), *Econometric analysis of cross section and panel data*, MIT press.

Appendix

A Proofs

Proof of Theorem 1. Observe that

$$\begin{aligned}\sqrt{n}(\hat{\psi}-\psi) &= \left(\frac{1}{n}\sum_{i=1}^{n}\hat{z}_i\hat{z}_i'\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{z}_i(y_i-\hat{z}_i'\psi)\\ &= \left(\frac{1}{n}\sum_{i=1}^{n}\hat{z}_i\hat{z}_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{z}_i(m_i-\hat{m}_i)'\alpha\right)+\left(\frac{1}{n}\sum_{i=1}^{n}\hat{z}_i\hat{z}_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{z}_i\epsilon_i\right)\\ &=: A+B.\end{aligned}$$

By Assumption 1(b),

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{z}_i(m_i-\hat{m}_i)'\alpha = -\begin{bmatrix}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{m}_i(\hat{m}_i-m_i)'\alpha\\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_i(\hat{m}_i-m_i)'\alpha\end{bmatrix}\xrightarrow{p}\begin{bmatrix}-\Omega\alpha\\ -\Gamma\alpha\end{bmatrix},$$

hence, by Assumptions 1(a) and 1(b) and Slutsky's theorem,

$$A\xrightarrow{p}-\mathbb{E}[z_iz_i']^{-1}\begin{bmatrix}\Omega & 0\\ \Gamma & 0\end{bmatrix}\psi.$$

Assumptions 1(a)-(c) and the Continuous Mapping Theorem (CMT) imply that $B\xrightarrow{d}N(\mathbb{E}[z_iz_i']^{-1}\Xi, V)$.

Finally, Assumptions 1(b)-(c) and the CMT imply that $\hat{V}\xrightarrow{p}V$. ■

Proof of Corollary 1. Let

$$\mathbb{E}[z_iz_i']^{-1}=\begin{bmatrix}B_{11} & B_{12}\\ B_{21} & B_{22}\end{bmatrix}$$

so that B_{21} is the lower-left block and B_{22} is the lower-right block.

From the formula for inversion of block matrices,

$$\begin{aligned}B_{21} &= -\left(\mathbb{E}[x_ix_i']-\mathbb{E}[x_im_i']\mathbb{E}[m_im_i']^{-1}\mathbb{E}[m_ix_i']\right)^{-1}\mathbb{E}[x_im_i']\mathbb{E}[m_im_i']^{-1},\\ B_{22} &= \left(\mathbb{E}[x_ix_i']-\mathbb{E}[x_im_i']\mathbb{E}[m_im_i']^{-1}\mathbb{E}[m_ix_i']\right)^{-1}.\end{aligned}$$

Thus, for A defined in the proof of Theorem 1,

$$\begin{aligned}A\xrightarrow{p}&\left(\mathbb{E}[x_ix_i']-\mathbb{E}[x_im_i']\mathbb{E}[m_im_i']^{-1}\mathbb{E}[m_ix_i']\right)^{-1}\mathbb{E}[x_im_i']\mathbb{E}[m_im_i']^{-1}\Omega\alpha\\ &-\left(\mathbb{E}[x_ix_i']-\mathbb{E}[x_im_i']\mathbb{E}[m_im_i']^{-1}\mathbb{E}[m_ix_i']\right)^{-1}\Gamma\alpha.\end{aligned}$$

The term $\mathbb{E}[m_im_i']$ is a diagonal matrix with the population probabilities that a randomly drawn unit truly belongs to group g on the diagonal: $\mathbb{E}[m_im_i']=\text{diag}(\pi_1,\ldots,\pi_G)$. The term $\mathbb{E}[x_im_i']$ is the matrix of group-specific averages of x—its (k,g)-entry is $\mathbb{E}[x_{ik}|g_i=g]\pi_g$. Therefore, $\mathbb{E}[x_im_i']\mathbb{E}[m_im_i']^{-1}$ is the

matrix with gth column equal to $\mathbb{E}[x_i|g_i = g]$. Moreover, Assumption 1 implies that the (a, b) entry of Ω is given by

$$\Omega_{ab} = \mathbb{1}\{a = b\} \sum_{g=1}^{G} \pi_g q_{a|g} - \pi_b q_{a|b}.$$

Thus, the gth element of the vector $\Omega\alpha$ is given by

$$\sum_{b=1}^{G} \pi_b q_{g|b}(\alpha_g - \alpha_b),$$

so we deduce that

$$\mathbb{E}[x_i m_i']\mathbb{E}[m_i m_i']^{-1}\Omega\alpha = \sum_{a=1}^{G}\sum_{b=1}^{G} \mathbb{E}[x_i|g_i = a]\pi_b q_{a|b}(\alpha_a - \alpha_b).$$

Similarly, from the definition of Γ we obtain

$$\Gamma\alpha = \sum_{g} \pi_g \mathbb{E}[x_i|g_i = g] \sum_{h} q_{h|g}(\alpha_h - \alpha_g).$$

It then follows that

$$\mathbb{E}[x_i m_i']\mathbb{E}[m_i m_i']^{-1}\Omega\alpha - \Gamma\alpha = \sum_{a=1}^{G}\sum_{b=1}^{G} \left(\mathbb{E}[x_i|g_i = a] - \mathbb{E}[x_i|g_i = b]\right) \pi_b(\alpha_a - \alpha_b)q_{a|b}.$$

On the other hand, $\mathbb{E}[x_i x_i'] - \mathbb{E}[x_i m_i']\mathbb{E}[m_i m_i']^{-1}\mathbb{E}[m_i x_i']$ is the population covariance matrix of x_i conditional on m_i averaged over groups—equivalently, the residual variance of x_i after projecting out group dummies. Using previous arguments, we can write

$$\mathbb{E}[x_i x_i'] - \mathbb{E}[x_i m_i']\mathbb{E}[m_i m_i']^{-1}\mathbb{E}[m_i x_i'] = \mathbb{E}[x_i x_i'] - \sum_{g=1}^{G} \pi_g \mathbb{E}[x_i|g_i = g]\mathbb{E}[x_i|g_i = g]'.$$

But $\mathbb{E}[x_i x_i']$ decomposes by the law of total expectation

$$\mathbb{E}[x_i x_i'] = \sum_{g=1}^{G} \pi_g \mathbb{E}[x_i x_i'|g_i = g],$$

so that

$$\mathbb{E}[x_i x_i'] - \mathbb{E}[x_i m_i']\mathbb{E}[m_i m_i']^{-1}\mathbb{E}[m_i x_i'] = \sum_{g=1}^{G} \pi_g \mathrm{Var}(x_i|g_i = g).$$

Using the previous results and the definition of ρ_{hg}, we obtain

$$\begin{aligned}\mathbb{E}[z_i z_i']^{-1}\Xi &= B_{21}\sum_g \pi_g \sum_{h\neq g} q_{h|g}\mathbb{E}[\epsilon_i|\hat{g}_i = h, g_i = g](e_h - e_g)\\ &= B_{21}\sum_g \pi_g \sum_{h\neq g} q_{h|g}\rho_{hg}(\alpha_h - \alpha_g)(e_h - e_g)\\ &= -\left(\sum_{g=1}^{G}\pi_g \text{Var}(x_i|g_i = g)\right)^{-1}\sum_g \pi_g \sum_{h\neq g} q_{h|g}\rho_{hg}\left(\mathbb{E}[x_i|g_i = h] - \mathbb{E}[x_i|g_i = g]\right)(\alpha_h - \alpha_g).\end{aligned}$$

The result then follows. ■

Proof of Corollary 2. Observe that, when $G = 2$, the asymptotic bias of $\hat{\beta}$ equals

$$ABias(\hat{\beta}) = \kappa_{12}q_{1|2} + \kappa_{21}q_{2|1}.$$

Moreover, $q_{1|2}$ and $q_{2|1}$ are both non-negative, and the sensitivity weights satisfy

$$\kappa_{ab} \propto (\mathbb{E}[x_i|g_i = a] - \mathbb{E}[x_i|g_i = b])(\alpha_a - \alpha_b),$$

so that letting $\Delta = \alpha_1 - \alpha_2$ and $\mu = \mathbb{E}[x_i|g_i = 1] - \mathbb{E}[x_i|g_i = 2]$,

$$\kappa_{12} \propto \mu\Delta \quad \text{and} \quad \kappa_{21} \propto \mu\Delta.$$

Clearly, κ_{21} and κ_{12} automatically point in the same direction. Thus, we can focus on κ_{12} without loss of generality and conclude that asymptotic bias will be positive if and only if $\kappa_{12} \geqslant 0$. In turn, $\kappa_{12} \geqslant 0$ if and only if μ and Δ have the same sign. This shows the result. ■

Proof of Corollary 3. The proof follows similar steps as for Corollary 1, but now focuses on the upper block of $\hat{\psi}$. Thus, using the formula for block matrix inversion,

$$\begin{aligned}B_{11} &= \mathbb{E}[m_i m_i']^{-1} + \mathbb{E}[m_i m_i']^{-1}\mathbb{E}[m_i x_i']S^{-1}\mathbb{E}[x_i m_i']\mathbb{E}[m_i m_i']^{-1},\\ B_{12} &= -\mathbb{E}[m_i m_i']^{-1}\mathbb{E}[m_i x_i']S^{-1}.\end{aligned}$$

where $S = \mathbb{E}[x_i x_i'] - \mathbb{E}[x_i m_i']\mathbb{E}[m_i m_i']^{-1}\mathbb{E}[m_i x_i']$ is the expected within-group covariance matrix (the denominator in Corollary 1). Following the same arguments as in the proof of Corollary 1, note that $\mathbb{E}[m_i m_i']^{-1}\mathbb{E}[m_i x_i']$ simplifies to the matrix of group-conditional means, call it μ, where the gth row is $\mu_g' = \mathbb{E}[x_i|g_i = g]$. Thus,

$$\begin{aligned}B_{11} &= \mathbb{E}[m_i m_i']^{-1} + \mu S^{-1}\mu',\\ B_{12} &= -\mu S^{-1}.\end{aligned}$$

Let Ξ_1 be the upper block of Ξ. The asymptotic bias of $\hat{\alpha}$ is the upper block of the bias term in Theorem 1:

$$ABias(\hat{\alpha}) = -B_{11}\Omega\alpha - B_{12}\Gamma\alpha + B_{11}\Xi_1.$$

Substituting the expressions for B_{11} and B_{12}:

$$\begin{aligned} ABias(\hat{\alpha}) &= -(\mathbb{E}[m_i m_i']^{-1} + \mu S^{-1}\mu')(\Omega\alpha - \Xi_1) + \mu S^{-1}\Gamma\alpha \\ &= \underbrace{-\mathbb{E}[m_i m_i']^{-1}(\Omega\alpha - \Xi_1)}_{\text{Direct Effect}} - \underbrace{\mu S^{-1}(\mu'(\Omega\alpha - \Xi_1) - \Gamma\alpha)}_{\text{Indirect Effect}} \end{aligned}$$

Because $\mathbb{E}[m_i m_i'] = \text{diag}(\pi_1, \ldots, \pi_G)$ is the diagonal matrix of group probabilities and each row g of $\Omega\alpha - \Xi_1$ is given by

$$\sum_{b=1}^{G} \pi_b(1 - \rho_{gb})q_{g|b}(\alpha_g - \alpha_b),$$

the direct effect is given by

$$-\frac{1}{\pi_a}\sum_{g \neq a} \pi_g(1 - \rho_{ag})q_{a|g}(\alpha_a - \alpha_g).$$

Moreover, we know from the proof of Corollary 1 that the term $S^{-1}(\mu'(\Omega\alpha - \Xi_1) - \Gamma\alpha)$ in the indirect effect is exactly the asymptotic bias of the slope coefficients, $ABias(\hat{\beta})$. The fact that we can replace $\mathbb{E}[x_i|g_i = a]$ by $\mathbb{E}[x_i|g_i = a] - \mathbb{E}[x_i]$ without loss of generality is explained in footnote 7. We then deduce the result. ■

Proof of Proposition 1. Each term is analyzed in order:

- Group shares: Decompose

$$\hat{\pi}_g = e_g' \frac{1}{n}\sum_{i=1}^{n} \hat{m}_i = e_g' \frac{1}{n}\sum_{i=1}^{n} m_i + e_g' \frac{1}{n}\sum_{i=1}^{n} (\hat{m}_i - m_i)\,.$$

  The first term on the right-hand side converges to π_g by the LLN and CMT. The second term is $O_p(n^{-1/2})$ by Assumption 1(b). Thus, $\hat{\pi}_g \xrightarrow{p} \pi_g$.

- Group-specific means: The estimator can be written as

$$\hat{\mathbb{E}}[x_i|\hat{g}_i = g] = \frac{n^{-1}\sum_{i=1}^{n} e_g'\hat{m}_i x_i}{n^{-1}\sum_{i=1}^{n} e_g'\hat{m}_i}.$$

  The denominator equals $\pi_g + o_p(1)$ by the previous bullet point. The numerator can be written as

$$\frac{1}{n}\sum_{i=1}^{n} e_g'\hat{m}_i x_i = \frac{1}{n}\sum_{i=1}^{n} e_g' m_i x_i + \frac{1}{n}\sum_{i=1}^{n} e_g'\left(\hat{m}_i - m_i\right) x_i.$$

  The first term on the right-hand side converges to $\pi_g \mathbb{E}[x_i|g_i = g]$ by the LLN. The second term is $o_p(1)$ by Assumption 1(b). Thus, $\hat{\mathbb{E}}[x_i|\hat{g}_i = g] \xrightarrow{p} \mathbb{E}[x_i|g_i = g]$.

- Coefficients: It follows from Theorem 1 that $\hat{\psi} - \psi = O_p(n^{-1/2})$. Thus, $\hat{\alpha} \xrightarrow{p} \alpha$.

- Group-specific variances: Letting $\Delta = \mathbb{E}[x_i|g_i = g] - \hat{\mathbb{E}}[x_i|\hat{g}_i = g]$, we can write the estimator as

$$\begin{aligned}
\widehat{\mathrm{Var}}(x_i|\hat{g}_i = g) &= \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{\mathbb{E}}[x_i|\hat{g}_i = g]\right)\left(x_i - \hat{\mathbb{E}}[x_i|\hat{g}_i = g]\right)' \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mathbb{E}[x_i|g_i = g] + \Delta\right)\left(x_i - \mathbb{E}[x_i|g_i = g] + \Delta\right)' \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mathbb{E}[x_i|g_i = g]\right)\left(x_i - \mathbb{E}[x_i|g_i = g]\right)' + \Delta\Delta' \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\Delta\left(x_i - \mathbb{E}[x_i|g_i = g]\right)' + \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mathbb{E}[x_i|g_i = g]\right)\Delta'.
\end{aligned}$$

  Since $\Delta = o_p(1)$ by the bullet point above and $\frac{1}{n}\sum_{i=1}^{n} x_i - \mathbb{E}[x_i|g_i = g] = O_p(n^{-1/2})$, the second, third and fourth terms on the right-hand side are $o_p(1)$. Thus, $\widehat{\mathrm{Var}}(x_i|\hat{g}_i = g) \xrightarrow{p} \mathrm{Var}(x_i|g_i = g)$ by the LLN.

- Relative bias factor: For most of the classifiers considered, ρ_{ab} is a deterministic constant, hence so is $\hat{\rho}_{ab}$. In some cases, e.g., the finite mixture model, it depends on the objects analyzed above, hence consistency follows by the CMT.

The result then follows by the CMT. ■

Proof of Theorem 2. The proof follows that in Theorem 1 verbatim, but is repeated here for completeness. Observe that

$$\begin{aligned}
\sqrt{n}(\hat{\psi} - \psi) &= \left(\frac{1}{n}\sum_{i=1}^{n}\hat{Z}_i\hat{Z}_i'\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{Z}_i(y_i - \hat{Z}_i'\psi) \\
&= \left(\frac{1}{n}\sum_{i=1}^{n}\hat{Z}_i\hat{Z}_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{Z}_i(m_i - \hat{m}_i)'\alpha\right) + \left(\frac{1}{n}\sum_{i=1}^{n}\hat{Z}_i\hat{Z}_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{Z}_i\epsilon_i\right) \\
&=: A + B.
\end{aligned}$$

By Assumption 2(b),

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{Z}_i(m_i - \hat{m}_i)'\alpha = -\begin{bmatrix}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{m}_i(\hat{m}_i - m_i)'\alpha \\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_i(\hat{m}_i - m_i)'\alpha\end{bmatrix} \xrightarrow{p} \begin{bmatrix}-(\Omega\otimes \mathrm{I}_T)\alpha \\ -\bar{\Gamma}\alpha\end{bmatrix},$$

hence, by Assumptions 2(a) and 2(b) and Slutsky's theorem,

$$A \xrightarrow{p} -\mathbb{E}[Z_iZ_i']^{-1}\begin{bmatrix}(\Omega\otimes \mathrm{I}_T) & 0 \\ \bar{\Gamma} & 0\end{bmatrix}\psi.$$

Assumptions 2(a)-(c) and the Continuous Mapping Theorem (CMT) imply that $B \xrightarrow{d} N(\mathbb{E}[Z_iZ_i']^{-1}\bar{\Xi}, V)$.

Finally, Assumptions 2(b)-(c) and the CMT imply that $\hat{V} \xrightarrow{p} V$. ■

Proof of Corollary 4. The proof proceeds analogously to Corollary 1, with the necessary modifications to account for the vector-valued nature of the panel data.

Here I only clarify the main changes. First, since x_i is now a $K \times T$ matrix, the denominator term satisfies

$$\mathbb{E}[x_i x_i'] - \mathbb{E}[x_i m_i']\mathbb{E}[m_i m_i']^{-1}\mathbb{E}[m_i x_i'] = \sum_{g=1}^{G}\sum_{t=1}^{T} \pi_g \text{Var}(x_{it}|g_i = g).$$

Second, using the definition of Ω and the fact that $\alpha = (\alpha_1', \ldots, \alpha_G')'$ where each α_g is T-dimensional, we obtain that $(\Omega \otimes \mathrm{I}_T)\alpha$ is now a GT-dimensional vector, with hth block (corresponding to group h, of size T) equal to

$$[(\Omega \otimes \mathrm{I}_T)\alpha]_h = \sum_{g \neq h} \pi_g q_{h|g}(\alpha_h - \alpha_g).$$

Next, using the definition of $\bar{\Gamma}$ and the fact that $((e_h - e_g)' \otimes \mathrm{I}_T)\alpha = \alpha_h - \alpha_g$,

$$\bar{\Gamma}\alpha = \sum_{g=1}^{G} \pi_g \sum_{h=1}^{G} q_{h|g}\mathbb{E}[x_i|g_i = g](\alpha_h - \alpha_g) = \sum_{t=1}^{T}\sum_{g=1}^{G}\sum_{h \neq g} \pi_g q_{h|g}\mathbb{E}[x_{it}|g_i = g](\alpha_{ht} - \alpha_{gt}).$$

Finally, using the definition of $\rho_{hg,t}$,

$$\begin{aligned}\mathbb{E}[Z_i Z_i']^{-1}\bar{\Xi} = -&\left(\sum_{g=1}^{G}\sum_{t=1}^{T} \pi_g \text{Var}(x_{it}|g_i = g)\right)^{-1} \times \\ &\sum_{t=1}^{T}\sum_{g} \pi_g \sum_{h \neq g} q_{h|g}\rho_{hg,t}\left(\mathbb{E}[x_{it}|g_i = h] - \mathbb{E}[x_{it}|g_i = g]\right)(\alpha_{ht} - \alpha_{gt}).\end{aligned}$$

The result then follows. ■

Proof of Corollary 5. The proof proceeds exactly as in Corollary 2, but now $\mu\Delta$ is an inner product of T-dimensional vectors. The terms $(1 - \rho_{hg,t})$ can be safely ignored because they are time invariant (hence factor out of the t-sum) and are non-negative. ■

Proof of Corollary 6. Using Corollary 4 and the triangle inequality,

$$\left\|ABias(\hat{\beta})\right\| = \left\|\sum_{a=1}^{G}\sum_{b \neq a} \kappa_{ab} q_{a|b}\right\| \leqslant G(G-1)\max_{a,b \neq a} \|\kappa_{ab}\| \, q_{a|b}.$$

Moreover, by the Cauchy Schwarz inequality, and using the fact that $0 \leqslant \rho_{ab,t} < 1$ and $0 < \pi_b < 1$,

$$\begin{aligned}\|\kappa_{ab}\| &\leqslant \left\|\left(\sum_{g=1}^{G}\sum_{t=1}^{T} \pi_g \text{Var}(x_{it}|g_i = g)\right)^{-1}\right\| \|\mathbb{E}[x_i|g_i = a] - \mathbb{E}[x_i|g_i = b]\| \, \|\alpha_a - \alpha_b\| \\ &= \left\|\left(\frac{1}{T}\sum_{g=1}^{G}\sum_{t=1}^{T} \pi_g \text{Var}(x_{it}|g_i = g)\right)^{-1}\right\| \left\|\frac{1}{\sqrt{T}}(\mathbb{E}[x_i|g_i = a] - \mathbb{E}[x_i|g_i = b])\right\| \left\|\frac{1}{\sqrt{T}}(\alpha_a - \alpha_b)\right\|,\end{aligned}$$

showing that

$$\kappa_{ab} = O\left(\sqrt{\frac{1}{T}\sum_{t=1}^{T}(\alpha_{at} - \alpha_{bt})^2}\right),$$

from what we deduce the result. ■

Proof of Corollary 7. The proof proceeds analogously to Corollary 3, with the necessary modifications to account for the vector-valued nature of the panel data.

Here I only discuss the main modifications. Following the same arguments as in the proof of Corollary 3, we can write

$$ABias(\hat{\alpha}) = \underbrace{-\mathbb{E}[m_i m_i']^{-1}((\Omega \otimes \mathrm{I}_T)\alpha - \bar{\Xi}_1)}_{\text{Direct Effect}} - \underbrace{\mu S^{-1}(\mu'[(\Omega \otimes \mathrm{I}_T) - \bar{\Xi}_1] - \bar{\Gamma}\alpha)}_{\text{Indirect Effect}}.$$

Now $\mathbb{E}[m_i m_i']$ is a block-diagonal matrix where the gth block is $\pi_g \mathrm{I}_T$. We found in the proof of Corollary 4 that the hth block of $(\Omega \otimes \mathrm{I}_T)\alpha - \bar{\Xi}_1$ was $\sum_{g \neq h} \pi_g q_{h|g}(1 - \rho_{hg}) \odot (\alpha_h - \alpha_g)$, where $\odot$ denotes the Hadamard product, hence the hth block of the direct effect is given by

$$-\frac{1}{\pi_h} \sum_{g \neq h} \pi_g q_{h|g}(1 - \rho_{hg}) \odot (\alpha_h - \alpha_g).$$

As in the proof of Corollary 3, we still identify $S^{-1}(\mu'[(\Omega \otimes \mathrm{I}_T) - \bar{\Xi}_1] - \bar{\Gamma}\alpha)$ in the indirect effect as the asymptotic bias of the slope coefficients, $ABias(\hat{\beta})$. Finally, the fact that we can recenter $\mathbb{E}[x_{it}|g_i = g]$ follows again by footnote 7. ■

Proof of Proposition 2. The proof follows the same logic as Proposition 1. The summation over $t \in [T]$ is finite. Thus, by the LLN and the CMT applied to the fixed-T sums, the consistency of the time-averaged sample moments to their population counterparts holds. ■

Proof of Theorem 3. Consider the scalar parameter $\theta = c'\beta$ and its estimator $\hat{\theta} = c'\hat{\beta}$. From Theorem 1 (or Theorem 2 for a panel model) and their respective corollaries, under any $q \in \mathcal{Q}$,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\left(\sum_{a=1}^{G} \sum_{b \neq a} c'\kappa_{ab} q_{a|b}, c'V_\beta c\right).$$

By the definition, the (estimated) maximum absolute bias is bounded by $\hat{\overline{B}}_c$. Let $Z \sim N(0,1)$. The t-statistic is $T_n = \sqrt{n}(\hat{\theta} - \theta)/\sqrt{\hat{V}_c}$. Under Assumption 1 (or 2 for a panel model), we have

$$T_n \xrightarrow{d} Z + \frac{\sum_{a=1}^{G} \sum_{b \neq a} c'\kappa_{ab} q_{a|b}}{\sqrt{c'V_\beta c}}.$$

The event $c'\beta \in \mathcal{C}_c$ is then equivalent to $|T_n| \leqslant cv_\alpha\left(\hat{\overline{B}}_c/\sqrt{\hat{V}_c}\right)$ under any $q \in \mathcal{Q}$. Since

$$\left|\frac{\sum_{a=1}^{G} \sum_{b \neq a} c'\hat{\kappa}_{ab} q_{a|b}}{\sqrt{\hat{V}_c}}\right| \leqslant \frac{\hat{\overline{B}}_c}{\sqrt{\hat{V}_c}}$$

almost surely for all $q \in \mathcal{Q}$, $\hat{\overline{B}}_c \xrightarrow{p} \overline{B}_c$ by Proposition 1 (or Proposition 2 for a panel), and given that the coverage probability of a fixed-length confidence interval centered at a biased estimator is minimized when

the bias is maximal (Armstrong and Kolesár, 2018), it follows that:

$$\lim_{n\to\infty}\inf_{q\in\mathcal{Q}}\mathbb{P}\left(c'\beta\in\mathcal{C}_c\right)\geqslant\mathbb{P}\left(\left|Z+\frac{\overline{B}_c}{\sqrt{c'V_\beta c}}\right|\leqslant cv_\alpha\left(\frac{\overline{B}_c}{\sqrt{c'V_\beta c}}\right)\right)=1-\alpha.$$

Thus, the asymptotic coverage is at least $1-\alpha$ uniformly over $\mathcal{P}_n$. ■

Proof of Proposition 3. The statement is equivalent to showing that the bias bound is sharp, i.e., attained. For notational simplicity I focus on the cross-sectional and scalar case, although the same calculations go through more generally; think of objects as T-dimensional vectors and replace κ_{ab} by $c'\kappa_{ab}$. I also ignore ρ without loss of generality. Observe that the magnitude of the asymptotic bias of $\hat{\beta}$ when $G=2$ equals

$$\left|ABias(\hat{\beta})\right|=\left|\kappa_{12}q_{1|2}+\kappa_{21}q_{2|1}\right|,$$

while the worst-case bias bound is given by

$$\bar{B}=\bar{q}_2|\kappa_{12}|+\bar{q}_1|\kappa_{21}|.$$

We need two conditions for both to be equal:

(a) No bias cancellation: the sensitivity weights must point to the same direction. Observe that these in turn satisfy

$$\kappa_{ab}\propto(\mathbb{E}[x_i|g_i=a]-\mathbb{E}[x_i|g_i=b])(\alpha_a-\alpha_b).$$

Thus, letting $\Delta=\alpha_1-\alpha_2$ and $\mu=\mathbb{E}[x_i|g_i=1]-\mathbb{E}[x_i|g_i=2]$,

$$\kappa_{12}\propto\mu\Delta\quad\text{ and }\quad\kappa_{21}\propto\mu\Delta.$$

Thus, it is automatically satisfied that both κ_{21} and κ_{12} will have the same sign, hence

$$\left|ABias(\hat{\beta})\right|=\left|\kappa_{12}q_{1|2}+\kappa_{21}q_{2|1}\right|=|\kappa_{12}|q_{1|2}+|\kappa_{21}|q_{2|1}.$$

(b) Accurate calibration: the worst-case misclassification scheme $\mathcal{Q}$ must be sharp. In the case where $G=2$, the admissible set of misclassification rates in Assumption 3 is equivalent to

$$\left\{q\geqslant 0:q_{1|2}\leqslant\bar{q}_2\text{ and }q_{2|1}\leqslant\bar{q}_1\right\}.$$

Thus, by step (a), under oracle calibration of $\bar{q}_1$ and $\bar{q}_2$,

$$\left|ABias(\hat{\beta})\right|=\bar{B}.$$

This completes the proof. ■

Proof of Theorem 4. The result is immediate from Proposition A.2. ■

Proposition A.1 *Let*

$$\hat{g}_i=\underset{k\in[G]}{\arg\min}\sum_{t=1}^{T}(y_{it}-x_{it}'\hat{\beta}-\hat{\alpha}_{kt})^2,$$

and suppose the following conditions hold:

(a) $\hat{\psi} - \psi = O_p(n^{-1/2})$.

(b) The error vector ϵ_i *has a distribution such that* $\mathbb{E}[\epsilon_i | a'\epsilon_i = r] = \frac{a}{\|a\|^2} r$ *for any* $a \in \mathbb{R}^T$ *and finite third moments. Moreover, for any* $h \neq g \in [G]$ *the scalar random variable* $R_i = \sum_{t=1}^{T} \epsilon_{it}(\alpha_{ht} - \alpha_{gt})$*, with upper support* ω_R*, possesses a sub-linear mean residual life function* $m_R(c) = \mathbb{E}[R_i - c \mid R_i > c]$*, such that* $m_R(c_n) = o(c_n)$ *and* $f_R(c_n) = o(n^{-1/4})$ *for any sequence* $c_n \to \omega_R$ *as* $n \to \infty$.

(c) Misclassification is local such that $\mathbb{P}(\hat{g}_i = h | g_i = g) = O(n^{-1/2})$.

Then, for any true group g *and destination group* h*:*

$$\mathbb{E}[\epsilon_{it} | \hat{g}_i = h, g_i = g] = \frac{\alpha_{ht} - \alpha_{gt}}{2}(1 + o_p(1)).$$

Proof of Proposition A.1. I first show that the estimation error in $(\hat{\beta}, \hat{\alpha})$ has a negligible first-order effect on the truncated mean. Define the infeasible (oracle) classification boundary variable $R_i = \sum_{t=1}^{T} \epsilon_{it}(\alpha_{ht} - \alpha_{gt})$ and the threshold $C = \sum_{t=1}^{T} (\alpha_{ht} - \alpha_{gt})^2/2$. The infeasible assignment rule is $\tilde{g}_i = h$ if and only if $R_i > C$.

The feasible classifier $\hat{g}_i = h$ satisfies:

$$\sum_{t=1}^{T} (y_{it} - x_{it}'\hat{\beta} - \hat{\alpha}_{ht})^2 < \sum_{t=1}^{T} (y_{it} - x_{it}'\hat{\beta} - \hat{\alpha}_{gt})^2.$$

Substituting y_{it} and expanding terms, we can rewrite the feasible condition as:

$$\sum_{t=1}^{T} \epsilon_{it}(\hat{\alpha}_{ht} - \hat{\alpha}_{gt}) - \sum_{t=1}^{T} x_{it}'(\hat{\beta} - \beta)(\hat{\alpha}_{ht} - \hat{\alpha}_{gt}) > \frac{1}{2}\sum_{t=1}^{T} (\hat{\alpha}_{ht} - \hat{\alpha}_{gt})(\hat{\alpha}_{ht} + \hat{\alpha}_{gt} - 2\alpha_{gt}),$$

or, after some manipulations, $R_i + \delta_i > C$, where

$$\begin{aligned}\delta_i = &\sum_{t=1}^{T} \epsilon_{it}(\hat{\alpha}_{ht} - \hat{\alpha}_{gt} - (\alpha_{ht} - \alpha_{gt})) - \sum_{t=1}^{T} x_{it}'(\hat{\beta} - \beta)(\hat{\alpha}_{ht} - \hat{\alpha}_{gt}) \\ &- \frac{1}{2}\sum_{t=1}^{T} \left(2(\alpha_{ht} - \alpha_{gt})(\hat{\alpha}_{ht} - \alpha_{ht}) + (\hat{\alpha}_{ht} - \alpha_{ht})^2 - (\hat{\alpha}_{gt} - \alpha_{gt})^2\right).\end{aligned}$$

Under our assumptions, estimated parameters are $\sqrt{n}$-consistent, hence $\delta_i = O_p(n^{-1/2})$.

The symmetric difference between the feasible and infeasible misclassification sets is $S_i = \mathbb{I}(R_i + \delta_i > C) \oplus \mathbb{I}(R_i > C)$. This indicator is non-zero only if R_i falls in the margin $[C - |\delta_i|, C + |\delta_i|]$. Because the parameter estimates are $\sqrt{n}$-consistent, standard asymptotic arguments allow us to bound the symmetric difference unconditionally with probability approaching 1. Under our local misclassification framework C diverges into the tail, which by assumption satisfies $f_{R_i}(C) = o(n^{-1/4})$. The probability measure of this margin is then:

$$\mathbb{P}(S_i = 1) = o(n^{-1/4}) \times O_p(n^{-1/2}) = o_p(n^{-3/4}).$$

While the margin boundary δ_i depends on ϵ_{it}, we can bypass this correlation by applying Hölder's inequal-

ity to bound the expectation via unconditional norms:

$$\mathbb{E}[|\epsilon_{it}|\mathbb{I}(S_i = 1)] \leqslant \mathbb{E}[|\epsilon_{it}|^3]^{1/3} \left(\mathbb{P}(S_i = 1)\right)^{2/3} = K \cdot o_p(n^{-1/2}).$$

Now, dividing by the misclassification probability $O(n^{-1/2})$ yields $o_p(1)$. This proves that replacing the feasible boundary with the infeasible one incurs an asymptotically negligible penalty in the conditional mean. Therefore, we may substitute the feasible condition with the infeasible oracle condition for the conditional first moment:

$$\mathbb{E}[\epsilon_{it} \mid \hat{g}_i = h, g_i = g] = \mathbb{E}[\epsilon_{it} \mid R_i > C] + o_p(1).$$

Applying the Law of Iterated Expectations and using the linearity of the conditional mean with $a = \alpha_h - \alpha_g$:

$$\mathbb{E}[\epsilon_{it} \mid R_i > C] = \mathbb{E}[\mathbb{E}[\epsilon_{it}|R_i] \mid R_i > C] = \frac{(\alpha_{ht} - \alpha_{gt})}{\sum_{t=1}^{T}(\alpha_{ht} - \alpha_{gt})^2}\mathbb{E}[R_i \mid R_i > C].$$

By definition of the mean residual life function $m_R(C)$, the truncated mean is exactly:

$$\mathbb{E}[R_i \mid R_i > C] = C + m_R(C).$$

Because $C \to \omega_R$ asymptotically, the sub-linear mean residual life condition implies $m_R(C) = C \cdot o(1)$. Therefore:

$$\mathbb{E}[R_i \mid R_i > C] = C(1 + o(1)).$$

Substituting C back yields:

$$\mathbb{E}[R_i \mid R_i > C] = \frac{1}{2}\sum_{t=1}^{T}(\alpha_{ht} - \alpha_{gt})^2(1 + o(1)).$$

Plugging this back into the earlier expression:

$$\mathbb{E}[\epsilon_{it} \mid R_i > C] = \frac{(\alpha_{ht} - \alpha_{gt})}{\sum_{t=1}^{T}(\alpha_{ht} - \alpha_{gt})^2}\left(\frac{1}{2}\sum_{t=1}^{T}(\alpha_{ht} - \alpha_{gt})^2(1 + o(1))\right) = \frac{\alpha_{ht} - \alpha_{gt}}{2}(1 + o(1)).$$

Recalling that $\mathbb{E}[\epsilon_{it} \mid \hat{g}_i = h, g_i = g] = \mathbb{E}[\epsilon_{it} \mid R_i > C] + o_p(1)$, the result follows. ■

Proposition A.2 *Let Assumption 1 hold for $(y_i, x_i, \hat{g}_{i,R}, \hat{g}_{i,P})$ and define $\tilde{\beta} = 2\hat{\beta}_R - \hat{\beta}_P$, where the R and P subscripts refer to estimation results under residual- and proxy-based classifiers, respectively. Suppose that $\rho_{hg} = 1/2$ and classification errors between residual- and proxy-based classifiers are equivalent. Then, as $n \to \infty$,*

$$\sqrt{n}(\tilde{\beta} - \beta) \xrightarrow{d} N(0, V_\beta).$$

Proof of Proposition A.2. To avoid notational burden, the result is derived for the cross-sectional case, although the same arguments apply for panel data.

Observe that, letting $S = \begin{bmatrix} 0_{K\times G} & 2\mathrm{I}_K & 0_{K\times G} & -\mathrm{I}_K \end{bmatrix} = \begin{bmatrix} 2 & -1 \end{bmatrix} \otimes \begin{bmatrix} 0_{K\times G} & \mathrm{I}_K \end{bmatrix}$,

$$
\begin{aligned}
\sqrt{n}(\tilde{\beta} - \beta) &= S\sqrt{n} \begin{pmatrix} \hat{\psi}_R - \psi \\ \hat{\psi}_P - \psi \end{pmatrix} \\
&= S \begin{bmatrix} \left(\frac{1}{n}\sum_{i=1}^n \hat{z}_{i,R}\hat{z}_{i,R}'\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{z}_{i,R}(m_i - \hat{m}_{i,R})'\alpha + \left(\frac{1}{n}\sum_{i=1}^n \hat{z}_{i,R}\hat{z}_{i,R}'\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{z}_{i,R}\epsilon_i \\ \left(\frac{1}{n}\sum_{i=1}^n \hat{z}_{i,P}\hat{z}_{i,P}'\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{z}_{i,P}(m_i - \hat{m}_{i,P})'\alpha + \left(\frac{1}{n}\sum_{i=1}^n \hat{z}_{i,P}\hat{z}_{i,P}'\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{z}_{i,P}\epsilon_i \end{bmatrix} \\
&= S \begin{bmatrix} \left(\frac{1}{n}\sum_{i=1}^n \hat{z}_{i,R}\hat{z}_{i,R}'\right)^{-1} & 0 \\ 0 & \left(\frac{1}{n}\sum_{i=1}^n \hat{z}_{i,P}\hat{z}_{i,P}'\right)^{-1} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{z}_{i,R}(m_i - \hat{m}_{i,R})'\alpha \\ \frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{z}_{i,P}(m_i - \hat{m}_{i,P})'\alpha \end{bmatrix} \\
&\quad + S \begin{bmatrix} \left(\frac{1}{n}\sum_{i=1}^n \hat{z}_{i,R}\hat{z}_{i,R}'\right)^{-1} & 0 \\ 0 & \left(\frac{1}{n}\sum_{i=1}^n \hat{z}_{i,P}\hat{z}_{i,P}'\right)^{-1} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{z}_{i,R}\epsilon_i \\ \frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{z}_{i,P}\epsilon_i \end{bmatrix} \\
&=: A + B.
\end{aligned}
$$

Under Assumption 1, assuming that local misclassification rates across classifiers are equivalent,

$$
A \xrightarrow{p} -S\left(\mathrm{I}_2 \otimes \mathbb{E}[z_i z_i']^{-1}\right)\left(\mathbb{1}_{2\times 1} \otimes \begin{bmatrix} \Omega\alpha \\ \Gamma\alpha \end{bmatrix}\right) = -S\left(\mathbb{1}_{2\times 1} \otimes \left\{\mathbb{E}[z_i z_i']^{-1}\begin{bmatrix} \Omega\alpha \\ \Gamma\alpha \end{bmatrix}\right\}\right).
$$

Moreover,

$$
\begin{bmatrix} \frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{z}_{i,R}\epsilon_i \\ \frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{z}_{i,P}\epsilon_i \end{bmatrix} \xrightarrow{d} N\left(\begin{bmatrix} \Xi \\ 0_{(2K+G)\times 1} \end{bmatrix}, \mathbb{1}_{2\times 2} \otimes \mathbb{E}[\epsilon_i^2 z_i z_i']\right).
$$

To see why, observe that

$$
\begin{bmatrix} \frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{z}_{i,R}\epsilon_i \\ \frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{z}_{i,P}\epsilon_i \end{bmatrix} = \mathbb{1}_{2\times 1} \otimes \frac{1}{\sqrt{n}}\sum_{i=1}^n z_i\epsilon_i + \begin{bmatrix} \frac{1}{\sqrt{n}}\sum_{i=1}^n (\hat{z}_{i,R} - z_i)\epsilon_i \\ \frac{1}{\sqrt{n}}\sum_{i=1}^n (\hat{z}_{i,P} - z_i)\epsilon_i \end{bmatrix}.
$$

The first term on the right hand side of the previous equation is the same across classifiers and converges in distribution to $N(0, (\mathbb{1}_{2\times 1}\mathbb{1}_{1\times 2}) \otimes \mathbb{E}[z_i z_i'\epsilon_i^2])$ while the second term converges in probability to $\begin{bmatrix} \Xi' & 0_{1\times(2K+G)} \end{bmatrix}'$ under Assumption 1(c). The result then follows by Slutsky's theorem.

The CMT and Slutsky's theorem then imply that $\sqrt{n}(\tilde{\beta} - \beta) \xrightarrow{d} N(\mu_d, V_d)$, where

$$
\begin{aligned}
\mu_d &= -S\left(\mathbb{1}_{2\times 1} \otimes \left\{\mathbb{E}[z_i z_i']^{-1}\begin{bmatrix} \Omega\alpha \\ \Gamma\alpha \end{bmatrix}\right\}\right) + S\left(\mathrm{I}_2 \otimes \mathbb{E}[z_i z_i']^{-1}\right)\begin{bmatrix} \Xi \\ 0_{(2K+G)\times 1} \end{bmatrix}, \\
V_d &= S\left(\mathrm{I}_2 \otimes \mathbb{E}[z_i z_i']^{-1}\right)\left(\mathbb{1}_{2\times 2} \otimes \mathbb{E}[\epsilon_i^2 z_i z_i']\right)\left(\mathrm{I}_2 \otimes \mathbb{E}[z_i z_i']^{-1}\right) S'.
\end{aligned}
$$

Let us simplify these expressions. Using the properties of Kronecker products and the notation in the proof

of Corollary 1, we can write

$$\begin{aligned}\mu_d &= -\left(\begin{bmatrix}2 & -1\end{bmatrix}\mathbb{1}_{2\times 1}\right)\otimes\left(\begin{bmatrix}0_{K\times G} & \mathrm{I}_K\end{bmatrix}\mathbb{E}[z_i z_i']^{-1}\begin{bmatrix}\Omega\alpha\\ \Gamma\alpha\end{bmatrix}\right)\\ &\quad+\left(\begin{bmatrix}2 & -1\end{bmatrix}\mathrm{I}_2\right)\otimes\left(\begin{bmatrix}0_{K\times G} & \mathrm{I}_K\end{bmatrix}\mathbb{E}[z_i z_i']^{-1}\right)\begin{bmatrix}\Xi\\ 0_{(2K+G)\times 1}\end{bmatrix}\\ &= -(B_{21}\Omega\alpha + B_{22}\Gamma\alpha) + 2B_{21}\Xi\\ &= 0,\end{aligned}$$

where the last equality follows under the assumption that $\rho_{hg} = 1/2$; cf. the proof of Corollary 1.

Similarly, using the properties of Kronecker products, we can rewrite

$$V_d = \left(\begin{bmatrix}2 & -1\end{bmatrix}\mathbb{1}_{2\times 2}\begin{bmatrix}2\\ -1\end{bmatrix}\right)\otimes\left(\begin{bmatrix}0_{K\times G} & \mathrm{I}_K\end{bmatrix}\mathbb{E}[z_i z_i']^{-1}\mathbb{E}[\epsilon_i^2 z_i z_i']\mathbb{E}[z_i z_i']^{-1}\begin{bmatrix}0_{G\times K}\\ \mathrm{I}_K\end{bmatrix}\right) = V_\beta,$$

where V_β is defined in Theorem 3. ■

B Calibrating the Misclassification Bound

To employ the robust inference procedure proposed in this paper, and in particular to define the class of admissible misclassification rates, one needs to choose an upper bound on the worst-case misclassification rate. There exist several options to choose this parameter in practice, but they naturally depend on the type of classifier used in first place. In the following, I list several data-dependent approaches. In any case, the interval length remains $O_p(n^{-1/2})$, so it asymptotically shrinks to zero at the standard parametric rate. This mathematically guarantees that the researcher's confidence intervals remain highly informative asymptotically.

B.1 Random Assignment Benchmark

In the extremely naive scenario where units are assigned uniformly at random into groups, the probability of correct classification is $1/G$. Therefore, one could set

$$\bar{q}_b = \sqrt{n}\left(\frac{G-1}{G}\right).$$

This is essentially the maximum entropy bound. It represents a scenario where the data contains zero information about the group structure. Of course, this choice is likely too loose in practice, yielding conservative inference. If our classifier performs no better than random guessing, the estimator is not just locally biased; it is globally inconsistent, and the group structure is unidentified. A natural concern is that using this bound would yield confidence intervals so wide they are uninformative.

B.2 Posterior Probability

Most classifiers, such as the Finite Mixture Models or the EM-algorithm, output posterior probabilities $\hat{\tau}_{ig} = P(g_i = g|\text{data})$. Misclassification is most likely for units situated at the boundary of clusters. For

instance, if a unit has a 55% probability of being in Group A and 45% in Group B, it contributes heavily to potential error.

One option is to estimate the group-specific expected error rate from the posteriors:

$$\hat{q}_b = \frac{1}{n_b} \sum_{i:\hat{g}_i=b} (1 - \hat{\tau}_{ib}),$$

where n_b is the number of units assigned to group b, and $\hat{\tau}_{ib}$ is the posterior probability that unit i actually belongs to b. Then, calibrate the bound by scaling this up by a safety factor ς to account for finite-sample estimation error of the probabilities themselves:

$$\bar{q}_b = \sqrt{n} (\varsigma \times \hat{q}_b).$$

For instance, a reasonable choice that works well in the simulations is $1 \leqslant \varsigma \leqslant 2$.

B.3 Stability

If our classifier is a black-box and does not provide posteriors (e.g., k-means), one natural option could be to use a stability analysis: If the group structure is robust, re-estimating the groups on subsamples of the data should yield the same assignments. Typically, instabilities would proxy for misclassification risk.

To perform this calibration, we would first randomly split the sample into, say, two halves or bootstrap resample. Then, we would re-estimate groups $\hat{g}^{(s)}$ on the resample, and compute the Hamming distance between the original assignment and the resampled assignment (matching labels to handle permutation invariance). We would then set $\hat{q}$ as the average discordance rate, and finally calibrate $\bar{q}_b = \sqrt{n}\hat{q}$. This approach may be infeasible if the original sample size is not large. Moreover, the availability of general formal results on the use of resampling methods to estimate misclassification error rates in this context appears to be limited (e.g., Monti et al., 2003; Hennig, 2007; Shamir and Tishby, 2010, for some related results). Moreover, resampling methods introduce a notorious computational hurdle: the label switching problem. Mapping group labels across samples seems like an arduous task.

Despite some classifiers only output hard assignments, an alternative is to force the method to output soft assignments. The general principle is to define a unit-specific measure of classification confidence, denoted $\hat{\tau}_{ig} \in [0, 1]$ (the probability unit i belongs to group g), and then aggregate the residual uncertainty $(1 - \hat{\tau}_{i\hat{g}_i})$ to form the bound. This is what the next subsections do.

B.4 Relative Distances

The standard k-means algorithm provides hard assignments, but the underlying geometry offers a rigorous measure of uncertainty based on the relative distance to the nearest competing centroid; see, e.g., Bishop and Nasrabadi (2006, Chapter 9.3.2) or MacKay (2003, Chapter 20.2). A unit can be deemed safe if it is very close to its assigned center $\hat{\alpha}_g$ and far from the next closest center. It can be deemed risky if it lies near the Voronoi boundary. We can attempt to convert distances into pseudo-probabilities using the softmax function. Let $d_{ig} = \|y_i - \hat{\alpha}_g - x_i'\hat{\beta}\|^2$, and define

$$\hat{\tau}_{ig} = \frac{\exp(-d_{ig}/2\hat{\sigma}^2)}{\sum_{k=1}^{G} \exp(-d_{ik}/2\hat{\sigma}^2)},$$

where $\hat{\sigma}^2$ is the sample variance of the residuals. We can then calibrate

$$\bar{q}_b = \sqrt{n}\,(\varsigma \times \hat{q}_b)\,, \quad \hat{q}_b = \frac{1}{n_b} \sum_{i:\hat{g}_i=b} (1 - \hat{\tau}_{ib})\,.$$

Using the group or C-LASSO classifier, the idea is the same. First, calculate the distance of unit i's specific effect $\hat{\alpha}_i$ from the cluster center $\hat{\mu}_g$, and normalize this by the distance to the nearest alternative cluster. To do so, let $\hat{\mu}_g$ be the center of group g. For a unit assigned to g, define the instability index:

$$S_i = \frac{\|\hat{\alpha}_i - \hat{\mu}_g\|}{\min_{k \neq g} \|\hat{\alpha}_i - \hat{\mu}_k\|}.$$

If $S_i \approx 0$, the unit is perfectly classified. If $S_i \approx 1$, the unit is on the boundary. We can approximate the probability of error by S_i, and use the calibration:

$$\bar{q}_b = \sqrt{n}\,(\varsigma \times \hat{q}_b)\,, \quad \hat{q}_b = \frac{1}{n_b} \sum_{i:\hat{g}_i=b} S_i.$$

B.5 Ensemble Agreement

Random forsts are an ensemble of B trees, hence the "vote share" is a direct, consistent estimator of the conditional probability of class membership (Breiman, 2001). For instance, if 90 out of 100 trees classify unit i as Group 1, the uncertainty is 10%. If only 51 trees do, the uncertainty is 49%.

Let V_{ig} be the proportion of trees in the forest that assign unit i to group g. The estimated probability is $\hat{\tau}_{ig} = V_{ig}$. Thus, one may calibrate the bound by

$$\bar{q}_b = \sqrt{n}\,(\varsigma \times \hat{q}_b)\,, \quad \hat{q}_b = \frac{1}{n_b} \sum_{i:\hat{g}_i=b} (1 - V_{ib}).$$

For rigorous practice, one should use out-of-bag votes for unit i (votes from trees where unit i was not used in training) to avoid overfitting the confidence.

Care should be exerted with some deep learning tools, such as neural nets. In that case, it is recommended to use Monte Carlo Dropout to recover a Bayesian measure of uncertainty (Gal and Ghahramani, 2016).

C Further Numerical Simulations

In this section I provide further numerical evidence on the small sample performance of the robust inference method. The designs are targeted to common empirical situations, like very small samples or the use of calibrated misclassification bounds as proposed in Section B. I also provide some discussion on the potential conservativeness of the approach and suggest several finite-sample adjustments.

C.1 Very Small Sample

In this section I continue to explore the finite-sample performance of the robust inference method proposed, and provide additional empirical recommendations for applied researchers.

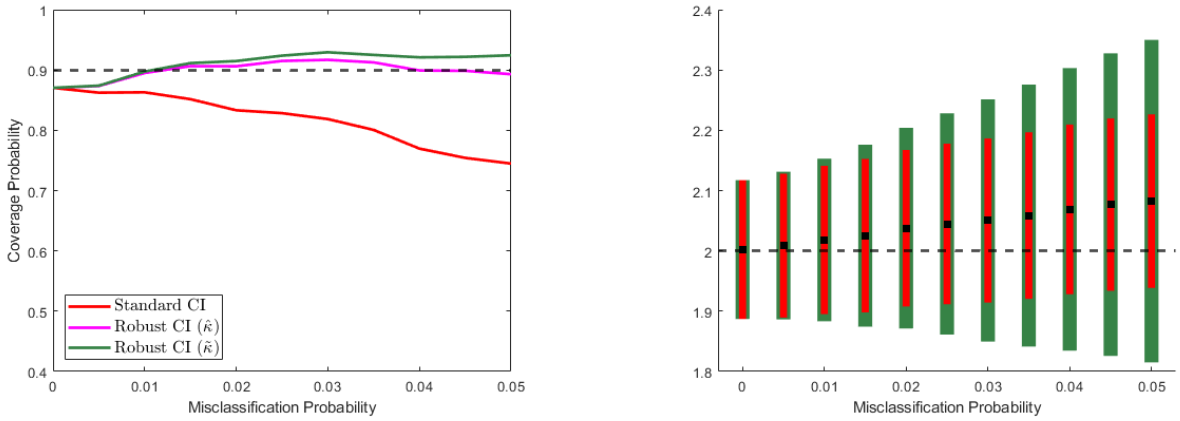
To simplify matters, the design is similar to the cross-sectional model used in Section 5. The main difference is that I now consider a sample of size $n = 50$ with $G = 3$ groups. The finite-sample distribution is approximated by 5,000 Monte Carlo replications.

Similarly to the main text, the true group memberships g_i are drawn uniformly at random. The DGP is given by:

$$y_i = \alpha_{g_i} + x_i\beta + \epsilon_i,$$

where the slope is $\beta = 2$ and the idiosyncratic error is $\epsilon_i \sim N(0, \sigma^2)$. The regressors x_i are generated conditionally on the true group membership: $x_i \sim N(\alpha_{g_i}, 1)$. Crucially, to induce misclassification bias, I specify distinct fixed-effects, hence regressor means across groups, and set $\alpha = (-2, 0, 2)$. This separation ensures that group misclassification generates correlation between the regressor and the effective error term. I set $\sigma = 1$ to control the signal-to-noise ratio, allowing for a clearer comparison of robust versus naive inference approaches. I still model local misclassification where the probability of error scales with sample size. I define the misclassification probability $p = q/\sqrt{n}$, and consider the same cyclic misclassification structure as in Section 5. Like in the main text, I use the oracle calibration for the misclassification bound and set $\bar{q}_b = q$. Given the small sample size, in this section I replace the Normal distribution by a Student's t-distribution to compute the critical values.

Figure C.I compares standard versus robust inference 90% confidence intervals as a function of the misclassification probability p, ranging from 0% to 5%, i.e., for different values of q.



(a) Coverage Probability

(b) Average CI

Figure C.I: Standard versus inflated robust inference for different degrees of misclassification

The results in Figure C.I are in line with those in Figure IIa and are very encouraging for the robust inference methods proposed in this paper. We see that the standard confidence intervals quickly lose coverage control as misclassification increases. In contrast, the robust confidence intervals, under both raw and inflated estimated sensitivity weights, manage to provide empirical coverage very close to the nominal level for all degrees of misclassification. An important distinction from this setup compared to that in Section 5 is that there are $G = 3$ groups now, so the robust inference approach need not be sharp, yet in this case it happens to be.

One important finite-sample threat to the validity of the robust inference approach is the attenuation bias in the group-specific objects used to construct the estimated sensitivity weights $\hat{\kappa}_{ab}$. Figure C.Ia shows that

in this case inflation has little effect, and even the raw estimate does not suffer undercoverage, although it was relevant in Figure IIa. One explanation could be that the number of groups acts as a natural inflator such that no further adjustments are required.

However, more generally and as observed in Figure IIa, the robust confidence interval may suffer slight undercoverage in finite samples, even under oracle calibration, when misclassification is large. The reason is intuitively simple: constructing the robust confidence interval requires evaluating the worst-case bias bound under estimated sensitivity weights. As shown in Corollary 1, the sensitivity of the estimator depends on the group-specific parameters, specifically terms of the form $(\mu_a - \mu_b)(\alpha_a - \alpha_b)$, where $\mu_a = \mathbb{E}[x_i|g_i = a]$. A practical challenge arises because the researcher must estimate these nuisance parameters (μ_g, α_g) using the observed (misclassified) groups. Misclassification causes attenuation bias in these estimates: the observed means $\hat{\mu}_g$ and $\hat{\alpha}_g$ are mixtures of the true underlying parameters, effectively shrinking the observed differences between groups. Calculating the bound using attenuated estimates may lead to an understated risk and undercoverage. To address this, one could apply a finite-sample inflation factor. For a misclassification probability p, the observed differences are attenuated by approximately $(1 - \gamma p)$, where γ depends on the mixing structure (theoretically $\gamma = 2$ for symmetric pairwise swapping). Since the bias depends on the product of regressor and fixed-effects differences, one could inflate the plug-in bias bound by a factor of $(1 - \gamma p)^{-2}$. A value that works remarkably well in the simulations is $\gamma \approx 1.5$. If the expected misclassification were large, this would err on the side of caution and yield potentially conservative, yet valid inference. This is what the green line in Figure C.I represents.

It is worth noting that the robust inference method may sometimes be conservative for a different reason. This is expected behavior when the misclassification bound is too loose: the method constructs the confidence interval to guard against the worst-case adversarial resorting of groups, i.e., swapping groups to maximize bias. The gap between the nominal and the empirical coverage would represent the price of robustness against unknown error structures, which emphasizes the importance of accurately calibrating the bound as suggested in Section B.

One final potential cause of conservativeness is bias cancellation. Roughly speaking, the robust confidence interval adjusts the critical value by the sum of the absolute value of each group's bias contribution. By a triangle inequality argument, it is easy to see that when bias contributions from different groups have opposite signs, the bias effects will cancel out but the bound will remain excessively large. When the bias contributions from all misclassified units point in the same direction, the robust confidence interval will be exact. This argument also hints that conservativeness will likely increase with G.

C.2 Varying Degrees of Group Separation

As alluded to in the introduction, one common cause of misclassification in finite samples is weak group separation. To model this problem and study its effects on inference, I follow the setup in González-Casasús (2025) and consider the following DGP:

$$y_{it} = \tilde{\alpha}_{g_i t} + x_{it}\beta + \epsilon_{it},$$

where $\tilde{\alpha}_{g_i t} = \nu_{g_i}\alpha_{g_i t} + (1 - \nu_{g_i})\zeta_t$. The parameters $\nu_g \in [0, 1]$ control the strength of separation: when $\nu_g = 0$, groups are not separated at all, as they share the same fixed-effects ζ_t; when $\nu_g = 1$, groups are strongly separated. For instance, when there are only two groups, one value of ν is sufficient to control the

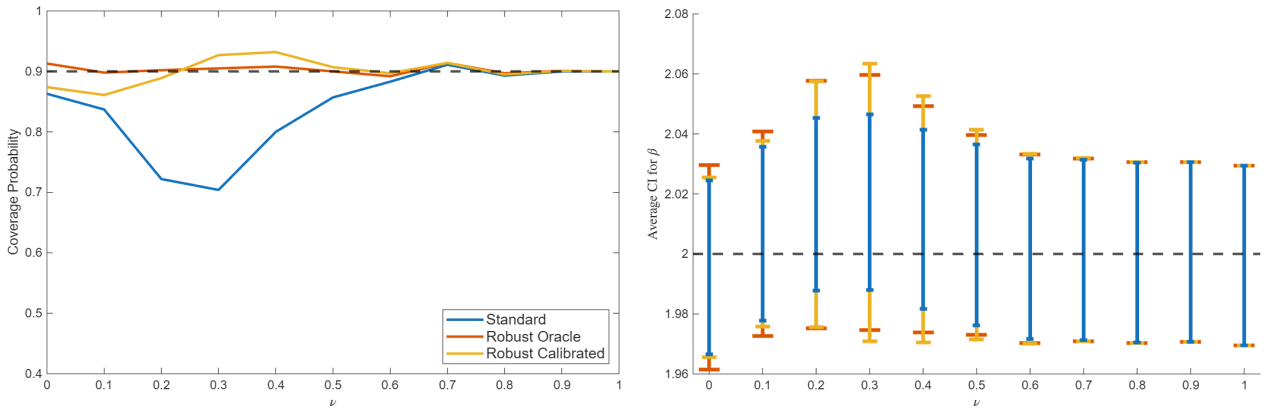
strength of separation:

$$\frac{1}{T}\sum_{t=1}^{T}(\tilde{\alpha}_{1t}-\tilde{\alpha}_{2t})^2=\nu\left(\frac{1}{T}\sum_{t=1}^{T}(\alpha_{1t}-\alpha_{2t})^2\right).$$

This setting is more challenging from the modeler's point of view because one cannot exogenously control misclassification rates, yet it permits creating more realistic endogenous patterns of misclassification due to weak group signals.

I consider a sample of size $n = 200$, $T = 15$ with $G = 2$ groups. The DGP is given by a slope $\beta = 2$ and Gaussian idiosyncratic error $\epsilon_{it} \sim N(0,1)$. The regressors x_{it} are generated conditionally on the true group membership: $x_{it} \sim N(\tilde{\alpha}_{g_it}, 1)$. Crucially, to induce misclassification bias, I specify distinct fixed-effects, hence regressor means across groups, and set $\alpha_t \sim N((-1,1)', \mathrm{I}_2)$. The mean level is set to $\zeta_t \sim N(0,1)$. True groups are generated uniformly at random, and I use 1,000 Monte Carlo repetitions to approximate the finite-sample distribution.

I estimate the parameters and groups using the GFE estimator of Bonhomme and Manresa (2015). I compare the coverage probability of the standard versus the robust inference method. Since this is a form of k-means estimation, I use two different calibrations for the worst-case misclassification bound $\bar{q}_b$: oracle calibration, as in Section 5, and the relative distance calibration suggested in Section B.4. The oracle calibration is paired with an inflation factor of $\gamma = 0.5$, while the relative distance calibration is computed with a safety factor of $\varsigma = 1.5$ and an inflation factor of $\gamma = 0.5$. The informativeness of the resulting confidence intervals appears to be largely insensitive to the choices of (ς, γ), especially for moderate values of ν, such as $\nu \geqslant 0.3$.



(a) Coverage probability

(b) Average confidence interval

Figure C.II: Robust inference on cash flow sensitivity

Figure C.II reveals the behavior of standard versus robust inference for different degrees of group separation ν. Consistent with the findings in González-Casasús (2025), Figure C.IIa shows that standard inference based on the GFE estimator roughly has the right coverage for group separation close to zero, $\nu \approx 0$, and moderately large separation, $\nu \geqslant 0.5$. In the former case, accurate classification is impossible but the consequences are negligible, since there is effectively no heterogeneity. In the latter case, there is large heterogeneity but at the same time accurate classification is easy. However, standard inference breaks down for weak separation, $\nu \in (0.1, 0.5)$, since in that region there is relevant heterogeneity but it is hard to capture. As a consequence, confidence intervals significantly undercover. Figure C.IIb illustrates that the bias due to weak group separation is positive in this case.

Figure C.II also reveals that well-calibrated robust inference is valid regardless of the degree of separation

ν. The findings support the claim in the introduction that the proposed robust inference method provides a constructive solution to the weak separation problem.

D Dataset Construction

The data is obtained from Compustat Annual Fundamentals. I focus on the time period from 2000 until 2015 (inclusive). I restrict attention to the universe of U.S. public firms, excluding Financials (SIC 6000-6999) and Utilities (SIC 4900-4999) as their capital structure and regulation differ fundamentally.

The dataset is constructed similarly to Kaplan and Zingales (1997). In particular, I measure investment or capital expenditures by *capx*. I measure cash flow as the sum of earnings before extraordinary items (*ib*) and depreciation (*dp*). I deflate investment and cash flow by capital which I measure as net property, plant, and equipment (*ppnet*) at the beginning of the fiscal year. I measure average Tobin's Q as the market value of assets divided by the book value of assets (*at*) where the market value of assets equals the book value of assets (Price *prcc_f* $\times$ Shares *csho*) plus total liabiblities (*lt*) less deferred taxes (*txditc*). As do most papers in this literature, I calculate Q at the beginning of a firm's fiscal year.

I drop firm-years with missing data on lagged capital or Q to retain only a balanced panel. Moreover, I winsorize I/K, CF/K, and Q at the 1% and 99% levels to prevent results being driven by few extreme firms.

E Latent Financial Constraints: Robustness Checks

To evaluate how sensitive the estimated parameters are to the choice of model specification in Section 6, I run several variations of the baseline model.

The general goal of this section is to assess how robust the cash flow sensitivity of 0.0356 found in the main text is to different model specifications. First, I investigate the choice of the number of groups. In a context where $n \gg T$, information criteria such as the Bayesian information criterion are inconsistent and can only be guaranteed to provide asymptotically an upper bound on the true number of groups. In this case, such a criterion suggests a number of groups unreasonably large, hence I omit this approach. Following the suggestions in González-Casasús (2025), I instead employ a stability analysis across the number of groups G to see whether the choice of $G = 5$ is reasonable. The idea is that after accounting for the true number of groups, parameter estimates and their confidence bands should stabilize. Figure E.III plots the estimated cash flow sensitivity β_2 in the baseline model for different number of estimated groups G. To do so, I repeatedly run the model fixing a different number of groups each time. We can see that indeed parameter values (solid blue line) and standard errors (dashed black lines) are within a stable region for the value $G = 5$ chosen in the baseline specification. Therefore, this figure provides support to the robustness of the estimated cash flow sensitivity against a misspecified number of groups.

A second possible concern with the baseline specification could be the homogeneity assumption on β_2. As mentioned in the main text, the standard approach in the corporate finance literature is to estimate cash flow sensitivities within group. To that end, I extend the baseline model to include group-specific slopes:

$$I_{it} = \alpha_{g_i t} + \beta_{1,g_i} Q_{it-1} + \beta_{2,g_i} CF_{it} + \epsilon_{it}.$$

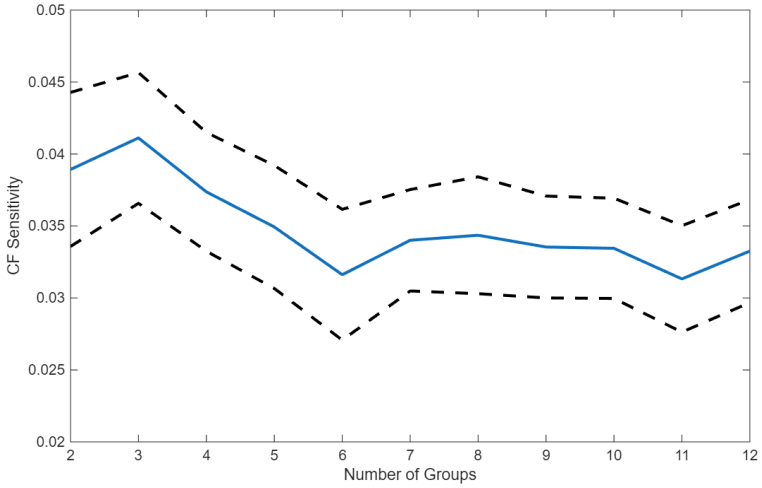


Figure E.III: Stability of cash flow sensitivity

I estimate the model using a variation of the GFE estimator proposed by Bonhomme and Manresa (2015). To do so, and in order to further enhance comparability with earlier estimates in the literature, I use two different samples: the 2000-2015 sample used in the main text, and the 1970-1984 sample commonly used in the literature (e.g., Kaplan and Zingales, 1997; Fazzari et al., 1988).

1970-1984	0.227	0.445	0.351	0.143	0.131
2000-2015	0.101	0.032	0.040	-0.006	0.024

Table I: Estimates of $\beta_{2,g}$, $G = 5$

In Table I we do observe some heterogeneity in the cash flow sensitivity across groups, yet all the estimated values for the 2000-2015 period appear small in magnitude compared to the findings for the 1970-1984 period used in the literature.[15] The average of $\hat{\beta}_2$ across groups for the later sample is 0.0382, which is indeed close to the average sensitivity I identify in the main text. This supports the claim in the main text that cash flow sensitivity appears economically modest in the 2000-2015 period. It is also evident that sensitivities have uniformly declined in recent years.

Finally, I also run several variations of the baseline model to allow for different patterns of unobserved heterogeneity. The common practice in the empirical finance literature is to only include time and/or fixed effects. Allowing for the more general patterns that grouped fixed-effects permit may significantly change the results (Leng et al., 2025). I run the same baseline model with none, individual, time and both fixed-effects.

The results in Table II suggest that the presence of complex forms of unobserved heterogeneity is likely significant in either sample period, hence supports the need to control for it using more robust methods like grouped fixed-effects. Furthermore, we can again observe that cash flow sensitivities seem to have declined in recent years. This may be another explanation to the contrasting results found in the main text.

For completeness, I present all the analogous results to the main text for the 1970-1984 sample period. The GFE estimate is $\hat{\beta}_2 = 0.194$. The standard 95% confidence interval is given by $[0.180, 0.208]$, whereas the

[15] The usual argument of identification of groups up to relabeling implies that the values of each column in Table I need not represent the same group across rows. However, it remains true that the values in the later sampling period are uniformly lower than in the earlier one.

Type of FEs	1970-1984	2000-2015
Grouped	0.194	0.036
Individual	0.266	0.051
Time	0.191	0.045
Ind + Time	0.253	0.046
None	0.272	0.099

Table II: Estimates of β_2

robust one is given by $[0.165, 0.223]$. This is due to a calibrated misclassification bound of

$$\hat{\bar{q}} = (2.74, 5.90, 7.03, 3.34, 5.49),$$

which implies misclassification probabilities of $8\%, 18\%, 21\%, 10\%, 6\%$, respectively. Moreover, the estimated matrix of sensitivity weights is given by

$$\hat{\kappa} = \frac{1}{2} \times 10^{-2} \times \begin{pmatrix} 0 & -0.35 & 3.64 & 1.42 & 0.58 \\ -0.09 & 0 & -1.97 & -0.83 & 0 \\ 1.11 & -2.23 & 0 & 0.26 & 0.36 \\ 1.36 & -2.93 & 0.83 & 0 & 0.27 \\ 0.66 & -0.01 & 1.33 & 0.32 & 0 \end{pmatrix}.$$

The above results imply a maximum bias of 0.017, which represents a relative bias of 9%. Again, this is a non-negligible magnitude and warns that historical sensitivities need to be analyzed with extra care.

F Group-Specific Slopes

In this section I consider a modification of the linear model in Section 2 where both intercepts and slopes vary by group:

$$y_i = \alpha_{g_i} + x_i' \beta_{g_i} + \epsilon_i = w_i' \theta_{g_i} + \epsilon_i,$$

where $w_i = (1, x_i')'$ is the $(K+1)$-dimensional vector of regressors and $\theta_g = (\alpha_g, \beta_g')'$ is the parameter vector for group g. The full parameter vector is now $\psi = (\theta_1', \dots, \theta_G')'$. Therefore, we can write the true and feasible models in compact form, respectively, as

$$y_i = \sum_{k=1}^{G} \mathbb{1}\{g_i = k\} w_i' \theta_k + \epsilon_i \quad \text{and} \quad y_i = \sum_{k=1}^{G} \mathbb{1}\{\hat{g}_i = k\} w_i' \theta_k + \epsilon_i.$$

For simplicity, I only focus on proxy-based classifiers, although the necessary adjustments for residual-based classifiers can be derived following the same arguments as in the main text (i.e., adjusting for the asymptotic bias of the feasible score). I consider the following high-level assumption, which is satisfied under the local misclassification scheme:

Assumption F.1 *(a)* $\mathbb{E}[\|w_i\|^2] < \infty$, $\mathbb{E}[\|\epsilon_i w_i\|^2] < \infty$ *and* $\mathbb{E}[w_i w_i']$ *has full rank.*

(b) As $n \to \infty$, *for any* $k \in [G]$, $\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{\hat{g}_i = k\} w_i w_i' \xrightarrow{p} \mathbb{E}[\mathbb{1}\{g_i = k\} w_i w_i']$ *and, for any* $g \neq k$, $\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbb{1}\{\hat{g}_i = k, g_i = g\} w_i w_i' \xrightarrow{p} \pi_g q_{k|g} \mathbb{E}[w_i w_i' | g_i = g]$.

(c) $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{1}\{\hat{g}_i = a\} w_i \epsilon_i \xrightarrow{d} N(0, \mathbb{E}[\mathbb{1}\{g_i = k\} \epsilon_i^2 w_i w_i'])$.

Theorem F.1 *Under Assumption F.1, as $n \to \infty$, the estimator $\hat{\psi}$ satisfies:*

$$\sqrt{n}(\hat{\psi} - \psi) \xrightarrow{d} N(b, V),$$

where the bias b and variance V are partitioned into group-specific blocks such that the ath block is given by:

$$b_a = \mathbb{E}[w_i w_i' | g_i = a]^{-1} \sum_{g \neq a} \frac{\pi_g}{\pi_a} q_{a|g} \mathbb{E}[w_i w_i' | g_i = g](\theta_g - \theta_a),$$
$$V_a = \pi_a^{-1} \mathbb{E}[w_i w_i' | g_i = a]^{-1} \mathbb{E}[\epsilon_i^2 w_i w_i' | g_i = a] \mathbb{E}[w_i w_i' | g_i = a]^{-1}.$$

Proof of Theorem F.1. Since the parameters θ_a and θ_b are distinct, the optimization problem separates into G independent least squares problems. I focus on the derivation for an arbitrary group a.

The estimator $\hat{\theta}_a$ solves the first-order condition:

$$0 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{1}\{\hat{g}_i = a\} w_i (y_i - w_i' \hat{\theta}_a).$$

Substituting the true model,

$$0 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{1}\{\hat{g}_i = a\} w_i (w_i' \theta_{g_i} + \epsilon_i - w_i' \hat{\theta}_a).$$

Rearranging terms we get

$$\left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\hat{g}_i = a\} w_i w_i' \right) \sqrt{n}(\hat{\theta}_a - \theta_a) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{1}\{\hat{g}_i = a\} w_i (\epsilon_i + w_i'(\theta_{g_i} - \theta_a)).$$

Let us analyze each term in order. By Assumption F.1(b),

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\hat{g}_i = a\} w_i w_i' \xrightarrow{p} \mathbb{E}[\mathbb{1}\{g_i = a\} w_i w_i'] = \pi_a \mathbb{E}[w_i w_i' | g_i = a].$$

Next, decompose

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{1}\{\hat{g}_i = a\} w_i (\epsilon_i + w_i'(\theta_{g_i} - \theta_a)) = \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{1}\{\hat{g}_i = a\} w_i \epsilon_i}_{=:T_1} + \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{1}\{\hat{g}_i = a\} w_i w_i'(\theta_{g_i} - \theta_a)}_{=:T_2}.$$

By Assumption F.1(c),

$$T_1 \xrightarrow{d} N(0, \mathbb{E}[\mathbb{1}\{g_i = k\} \epsilon_i^2 w_i w_i']).$$

Moreover, we can write

$$T_2 = \frac{1}{\sqrt{n}} \sum_{g=1}^{G} \sum_{i: g_i = g} \mathbb{1}\{\hat{g}_i = a\} w_i w_i'(\theta_g - \theta_a) = \frac{1}{\sqrt{n}} \sum_{g \neq a} \sum_{i=1}^{n} \mathbb{1}\{\hat{g}_i = a, g_i = g\} w_i w_i'(\theta_g - \theta_a).$$

By Assumption F.1(b),

$$T_2 \xrightarrow{p} \sum_{g \neq a} \pi_g q_{a|g} \mathbb{E}[w_i w_i' | g_i = g](\theta_g - \theta_a).$$

The result then follows by Assumption F.1(a) and Slutsky's theorem. ■

With the above result we can similarly characterize the bias in each of the elements in $\hat{\psi}$, as the next corollary shows.

Corollary F.1 *Under Assumption F.1, as $n \to \infty$, for any group a:*

$$\begin{aligned} ABias(\hat{\beta}_a) = \text{Var}(x_i | g_i = a)^{-1} \sum_{g \neq a} \frac{\pi_g}{\pi_a} q_{a|g} \big\{ &(\mathbb{E}[x_i | g_i = g] - \mathbb{E}[x_i | g_i = a])(\alpha_g - \alpha_a) \\ &+ (\mathbb{E}[x_i x_i' | g_i = g] - \mathbb{E}[x_i | g_i = a]\mathbb{E}[x_i | g_i = g]')(\beta_g - \beta_a) \big\}, \\ ABias(\hat{\alpha}_a) = \sum_{g \neq a} \frac{\pi_g}{\pi_a} q_{a|g} \left\{ (\alpha_g - \alpha_a) + \mathbb{E}[x_i | g_i = g]'(\beta_g - \beta_a) \right\} &- \mathbb{E}[x_i | g_i = a]' ABias(\hat{\beta}_a). \end{aligned}$$

The interpretation is very similar to the results in Section 2, although with one major difference. Now the slope parameter is subject to two sources of bias: mean shift bias and slope bias; the latter did not exist before because we assumed constant slopes. As before, the fixed-effects bias absorbs the direct composition effect plus the back-propagation of the slope bias evaluated at the group mean. But again, the direct composition effect is extended to include the errors in group-specific slopes.

G Debiased Estimators

In the main text, I characterized the asymptotic bias of the least squares estimator under misclassification and proposed robust confidence intervals to account for the resulting uncertainty. Throughout the paper, I noted that if the misclassification probabilities were known or consistently estimable, one could construct a point estimator that is asymptotically centered at the true parameter. In this appendix, I formally derive such a debiased estimator.

To correct the bias, we must identify the misclassification rates. I introduce the assumption of an auxiliary validation sample, which allows for the consistent estimation of the misclassification structure, and show that subtracting the estimated bias restores consistency.

In many practical settings, identification of misclassification rates is achieved via a small subset of ground truth data (e.g., audited records or re-surveyed participants).

Assumption G.1 *There exists an auxiliary validation subsample $\mathcal{V} \subset [n]$ of size n_V, drawn independently from the same population as the main sample. For all $i \in \mathcal{V}$, we observe the tuple $(y_i, x_i, \hat{g}_i, g_i)$, where g_i is the true group label and $\hat{g}_i$ is the estimated (potentially misclassified) label.*

Assumption G.2 *The validation sample size grows such that $n_V \to \infty$ as $n \to \infty$, and $n_V/n \to c \in [0, 1]$.*

Assumption G.3 *Let Π be the $G \times G$ transition matrix where entries $\pi_{ba} = \mathbb{P}(\hat{g}_i = b \mid g_i = a)$. We assume Π is invertible and that the misclassification mechanism is homogeneous across the sample.*

I propose a two-step plug-in estimator to correct for the asymptotic bias. First, using the validation sample

$\mathcal{V}$, I estimate the transition probabilities via frequency counts:

$$\hat{\pi}_{ba} = \frac{\sum_{i \in \mathcal{V}} \mathbb{1}\{\hat{g}_i = b, g_i = a\}}{\sum_{i \in \mathcal{V}} \mathbb{1}\{g_i = a\}}.$$

Similarly, I estimate the necessary moments of the covariates and group means using the validated subset to obtain consistent estimates of the sensitivity terms $\hat{\kappa}_{ab}$.

Then, I construct the estimated bias term:

$$\widehat{ABias}(\hat{\beta}) = \sum_{a=1}^{G} \sum_{b \neq a} \hat{\kappa}_{ab} \hat{q}_{a|b},$$

where $\hat{q}_{a|b} = \sqrt{n}\hat{\pi}_{ab}$ is the implied count of misclassified units scaled to the full sample size n based on the estimated rates $\hat{\pi}_{ab}$. The debiased estimator is thus defined as:

$$\tilde{\beta} = \hat{\beta} - \widehat{ABias}(\hat{\beta})/\sqrt{n}.$$

Proposition G.1 *Under Assumptions 1 (or 2 for a panel model), G.1-G.3, as $n \to \infty$, $\sqrt{n}(\tilde{\beta} - \beta) \xrightarrow{d} N(0, V_\beta)$.*

Proof of Proposition G.1. Decompose

$$\sqrt{n}(\tilde{\beta} - \beta) = \sqrt{n}(\hat{\beta} - \beta) - \widehat{ABias}(\hat{\beta}).$$

Under Assumption 1 (or 2 for a panel model), the first term on the right-hand side converges in distribution to $N(ABias(\hat{\beta}), V_\beta)$, as Theorems 1 and 2 show.

Proposition 1 (or 2 for a panel model) shows that $\hat{\kappa}_{ab} \xrightarrow{p} \kappa_{ab}$. Moreover, under Assumptions G.1-G.3 and the local misclassification underlying Assumptions 1 and 2, $\sqrt{n}\hat{\pi}_{ab} \xrightarrow{p} q_{a|b}$, what implies by the CMT that

$$\widehat{ABias}(\hat{\beta}) \xrightarrow{p} ABias(\hat{\beta}).$$

The result then follows by Slutsky's theorem. ■

An analogous result can be shown for the grouped fixed-effects. Proposition G.1 restores standard asymptotic normality results by correcting for the asymptotic bias caused by misclassification. An immediate implication of the result is the following.

Corollary G.1 *Under Assumptions 1 (or 2 for a panel model), G.1-G.3, as $n \to \infty$,*

$$\lim_{n \to \infty} \inf_{P \in \mathcal{P}_n} P(c'\beta \in \tilde{\mathcal{C}}_c) = 1 - \alpha, \quad \tilde{\mathcal{C}}_c = \left[c'\tilde{\beta} \pm z_{1-\alpha/2} \sqrt{c'\hat{V}_\beta c / n} \right].$$

The result in Corollary G.1 is possible thanks to the use of auxiliary data to estimate misclassification rates. Availability of validation data may not be always possible. Moreover, if n_V is small relative to n, the uncertainty in $\widehat{ABias}(\hat{\beta})$ may be substantive, which underscores the utility of the robust inference approach when validation data is unavailable or sparse.

H Relative Bias of Classifiers

This section provides a more detailed intuition for the relative classification biases of different methods. Throughout the discussion, I ignore parameter estimation uncertainty and focus on the cross-sectional case for notational simplicity; see Proposition A.1 for a more general and formal justification.

In a residual-based classifier, a unit i that truly belongs to group g is misclassified into group h if its observed outcome $y_i - x_i'\beta$ is closer to the centroid of group h than to the centroid of group g. Let α_g and α_h be the group fixed effects, and assume $\alpha_g < \alpha_h$ without loss of generality. The decision boundary B is the point where the classifier is indifferent between the two groups. For a unit in group g, the outcome is $y_i - x_i'\beta = \alpha_g + \epsilon_i$.

The event of misclassification $\{\hat{g}_i = h | g_i = g\}$ occurs when:

$$y_i - x_i'\beta > B \iff \alpha_g + \epsilon_i > B \iff \epsilon_i > B - \alpha_g.$$

We are first interested in the expected value of the idiosyncratic error ϵ_i given that the unit has crossed the boundary. Letting $c = B - \alpha_g$ represent the distance to the threshold, we seek $\mathbb{E}[\epsilon_i | \epsilon_i > c]$. Using the definition of conditional expectation for a continuous random variable with density $f(\cdot)$ and distribution $F(\cdot)$:

$$\mathbb{E}[\epsilon_i | \epsilon_i > c] = \frac{\int_c^\infty \epsilon f(\epsilon) d\epsilon}{1 - F(c)} = c + \frac{\int_c^\infty (1 - F(\epsilon)) d\epsilon}{1 - F(c)}.$$

The term $\frac{\int_c^\infty (1-F(\epsilon))d\epsilon}{1-F(c)}$ is known as the mean residual life function, denoted in the main text as $m(c)$.

One of the paper's core innovation is the local misclassification assumption. As $n \to \infty$, the probability of misclassification $q_{h|g}$ is assumed to be $O(n^{-1/2})$. For this probability to shrink to zero, the threshold c must be effectively moving into the tail of the distribution. The paper invokes Assumptions 1(d) and 2(d), which regulate the density of the error term near the threshold. Under this assumption, the overshoot becomes negligible relative to the distance c. Intuitively, this means that if a unit is misclassified, it is increasingly likely to be located almost exactly at the boundary B, because the probability of having an even larger ϵ_i (further into the tail) drops off very fast. Therefore, in the limit:

$$\lim_{c \to \infty} \mathbb{E}[\epsilon_i | \epsilon_i > c] = c.$$

Substituting $c = B - \alpha_g$ back into the limit result, and ignoring negligible terms:

$$\mathbb{E}[\epsilon_i | \hat{g}_i = h, g_i = g] = B - \alpha_g.$$

Let us connect this with $\rho_{hg} = \mathbb{E}[\epsilon_i | \hat{g}_i = h, g_i = g]/(\alpha_h - \alpha_g)$ defined in the main text. Because the boundary B must strictly lie between the grouped fixed-effects, $\alpha_g < B < \alpha_h$, it must be that $0 < \rho_{hg} < 1$ with high probability.

When $\hat{g}_i = h \neq g_i = g$, the estimator for β encounters an unobserved gap defined as:

$$(y_i - x_i'\beta) - \alpha_h = (\alpha_g - \alpha_h) + \epsilon_i.$$

Taking expectations, the average gap for a misclassified unit is:

$$(\alpha_g - \alpha_h) + \mathbb{E}[\epsilon_i | \hat{g}_i = h, g_i = g].$$

For a proxy-based classifier, the average gap remains $(\alpha_g - \alpha_h)$ because $\mathbb{E}[\epsilon_i | \hat{g}_i = h, g_i = g] = 0$. However, using our earlier derivations, the average gap for a misclassified unit under residual-based classifiers is given by $(\alpha_g - \alpha_h) + (B - \alpha_g) = B - \alpha_h$.

The relative bias reduction factor implied by these gaps is:

$$1 - \rho_{hg} = \frac{\alpha_h - B}{\alpha_h - \alpha_g},$$

which clearly depends on the type of classifier used through the boundary B.

- GFE/k-means: the objective function of these classifiers minimizes squared residuals, meaning the boundary B is the perpendicular bisector (the midpoint) of the two centroids:

  $$B = \frac{\alpha_g + \alpha_h}{2}.$$

  It then follows that

  $$1 - \rho_{hg} = \frac{\alpha_h - \frac{\alpha_g + \alpha_h}{2}}{\alpha_h - \alpha_g} = \frac{1}{2}.$$

- Finite mixture model: here the classifier follows the maximum a posteriori rule. For two groups g and h with mixing proportions π_g, π_h, the boundary B is defined where the posterior probabilities are equal:

  $$\pi_g f(y_i - x_i'\beta - \alpha_g) = \pi_h f(y_i - x_i'\beta - \alpha_h).$$

  Solving for the boundary in the case of Normal errors $N(0, \sigma^2)$:

  $$B = \frac{\alpha_g + \alpha_h}{2} + \frac{\sigma^2 \ln(\pi_h / \pi_g)}{\alpha_h - \alpha_g}.$$

  Substituting this B:

  $$1 - \rho_{hg} = \frac{\alpha_h - \left(\frac{\alpha_g + \alpha_h}{2} + \frac{\sigma^2 \ln(\pi_h / \pi_g)}{\alpha_h - \alpha_g} \right)}{\alpha_h - \alpha_g} = \frac{1}{2} - \frac{\sigma^2 \ln(\pi_h / \pi_g)}{(\alpha_h - \alpha_g)^2}.$$

  In the local misclassification framework considered in this paper, the relative bias reduction will typically be around $1/2$. However, especially in finite samples, the relative bias reduction may fall outside the unit interval if groups are extremely unbalanced ($\pi_h \gg \pi_g$ or $\pi_h \ll \pi_g$) or the signal-to-noise ratio is very small ($\sigma^2 \ll (\alpha_h - \alpha_g)^2$).

- Rank-based: these classifiers partition the residuals $\hat{u}_i = y_i - x_i'\beta$ into G groups based on empirical quantiles τ_k. The boundary B is the τ_h-th quantile of the residual distribution:

  $$\mathbb{P}(y_i - x_i'\beta \leqslant B) = \sum_{k=1}^{h} \pi_k.$$

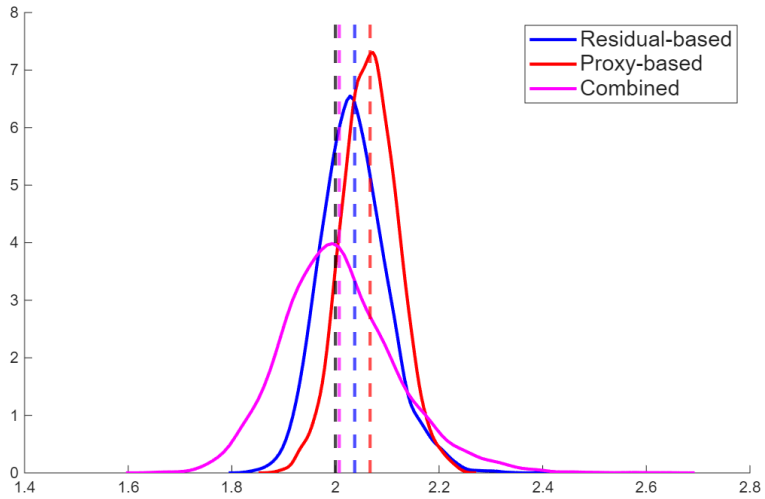


Figure H.IV: Finite-sample distribution of $\hat{\beta}$ under different classifiers

In this case, B, hence ρ_{hg}, can be obtained by inverting the above distribution, although it does not have a simple closed-form solution.

- C-LASSO: this estimator identifies groups by shrinking individual effects toward a set of latent centroids α_k. In the limit, C-LASSO behaves like the GFE/k-means estimator. Because the objective function is based on squared residuals (least squares), the local decision boundaries are the perpendicular bisectors (midpoints) between the centroids:

$$B \approx \frac{\alpha_g + \alpha_h}{2}.$$

Consequently, $1 - \rho_{hg} \approx 1/2$.

The mechanics in the above discussion can also be evaluated via Monte Carlo simulations. In Figure H.IV, we can see the finite-sample distribution of $\hat{\beta}$ under two different classifiers: one based on k-means on a proxy $W_i = \alpha_{g_i} + u_i$, and the other based on the GFE estimator. Exact details on the design are omitted since the specification closely follows the cross-sectional model in Section 5. From the plot it is clear that the center of the finite-sample distribution of $\hat{\beta}$ is closer to the true center (black dashed line) under the residual-based classifier than under the proxy-based classifier. Indeed, the finite-sample ratio of the biases happens to be 0.55, very close to the predicted 0.5. Moreover, the combined estimator $\tilde{\beta}$ successfully removes all the bias despite nonngeligible misclassification, at the expense of some variance inflation.