Universitat Pompeu Fabra Barcelona
Department of Economics and Business

Economics Working Paper Series

Working Paper No. 1956

Covariate-assisted estimation of grouped panel data models

Oriol González-Casasús

November 2026

Covariate-Assisted Estimation of Grouped Panel Data Models

Oriol González-Casasús*

Universitat Pompeu Fabra and Barcelona School of Economics

November 3, 2026

Abstract

Grouped panel data models are usually estimated by classifying units on the residuals of the outcome equation, which ignores variables outside the model informative about the groups. This paper proposes the covariate-assisted estimator, which classifies units by a weighted two-view k-means rule that weights each view by its estimated precision. Auxiliary covariates are redundant when classification is perfect but valuable when it is not. At the precision weight, the squared separation-to-noise ratios of the two views add up and the endogenous bias offset of residual-based classification is preserved, so that misclassification and bias are lower than under either view alone. When the auxiliary covariate is an included regressor, the first-order bias vanishes despite misclassification. In a county panel of minimum wages and restaurant employment, manufacturing employment disciplines a residual-based partition that isolates outlying counties. In the income–democracy panel, the savings rate stabilizes the estimates across the number of groups.

Keywords: Grouped fixed-effects, Misclassification, Auxiliary Covariates, k-means.
JEL Classification: C13, C23, C38.

*Department of Economics and Business, Universitat Pompeu Fabra. Email: oriol.gonzalez@upf.edu. This paper is a major revision of Chapter 1 of my PhD dissertation at the University of Pennsylvania. I am grateful for thoughtful comments and feedback from Xu Cheng and Frank Schorfheide. I also thank Karun Adusumilli, Frank Diebold, Jacob Dorn, Wayne Yuan Gao, Elena Manresa, Yulong Wang, and seminar and conference participants at several venues for helpful suggestions. I acknowledge financial support from the Spanish Agencia Estatal de Investigación, through the Severo Ochoa Programme for Centres of Excellence in R&D (Barcelona School of Economics CEX2024-001476-S).

1 Introduction

Panel data models with grouped fixed-effects (GFE) have become a standard tool for modeling unobserved heterogeneity. In the linear grouped panel data model of Bonhomme and Manresa (2015), units belong to one of a small number of latent groups, each with its own time path of fixed-effects, and memberships are estimated jointly with the parameters by a k-means classification of the residuals. The approach retains the flexibility of fixed-effects while remaining parsimonious, and it has been applied widely (e.g., Kline and Walters, 2016, Grunewald, Klasen, Martínez-Zarzoso and Muris, 2017, Sanogo, 2019, Mendolia, Stavrunova and Yerokhin, 2021, Gómez-Puig, Sosvilla-Rivero and Martínez-Zarzoso, 2022). Because the groups are estimated from the residuals of the outcome equation, the procedure ignores any variable outside the model that carries information about the latent types, even though such variables are often readily available. A firm's input choices reflect its productivity type, a country's savings behavior reflects long-run determinants of its political development, and a household's observable characteristics reflect the latent class that governs its decisions. This paper asks how such auxiliary covariates should be used, and shows that using them in the right way improves both classification and estimation of the model parameters.

Within the framework in which grouped estimators are usually analyzed, this question has only a trivial answer. That framework rests on oracle conditions under which the estimated partition is asymptotically equivalent to the true one, so every classifier delivers the limit distribution of the oracle estimator. The choice of classifier then has no first-order effect on the estimates, and auxiliary covariates can be discarded at no cost however strong their group signal. In the short panels typical of applications, however, classification is rarely perfect and different classifiers produce visibly different estimates, which oracle conditions cannot explain. This paper instead builds on the local misclassification framework of González-Casasús (2026*a*), in which the bias induced by misclassification is of the same order as sampling uncertainty and depends on how the units are classified. This is the property that any ranking of classifiers requires, and under it the value of auxiliary covariates becomes a well-posed question. The answer is nonetheless not obvious. Residual-based classifiers select the misclassified units on their outcome shocks, which offsets part of the bias these units induce, whereas proxy-based classifiers offset none. Yet proxies may misclassify fewer units, because in short panels the residuals are often noisier relative to the separation of the groups than the auxiliary covariates are. A good classifier must therefore exploit the accuracy of the proxy without giving up the offset of the residuals.

The answer proposed here is the covariate-assisted (CAS) estimator. It classifies units by a weighted two-view k-means rule in which each unit contributes two vectors, the residual trajectory of the outcome equation and the vector of auxiliary covariates, which are assumed to share the latent group structure. For instance, a savings rate and a country's propensity to democratize may be shaped by the same latent type through cultural traits. Importantly, the CAS estimator recovers the groups endogenously, with no membership equation and no restriction linking the groups to the covariates of the outcome equation, which is one of the main appeals of residual-based classifiers. The view weight is the main object of the analysis and is estimated as the precision weight under a Gaussian quasi-likelihood. Thus, the procedure has no tuning parameters, and it nests residual-based and proxy-based classification. Because the shared partition is an overidentifying restriction, the two views also provide a diagnostic for the validity of the proxy, a quasi-likelihood ratio statistic whose null distribution is calibrated by a bootstrap under the fitted model.

The first contribution is the estimator and the geometry of its classification error. In a Gaussian design, the probability that a unit is misclassified between two groups is a decreasing function of the separation-to-noise ratio of the weighted distance, and that ratio is maximized at the precision weight, where its square equals the sum of the squared ratios of the residual-based and the proxy-based rules. The two sources of information add up in quadrature, whatever their relative strength, and the optimal weight does not depend on how strong either signal is, which is why it can be estimated without knowledge of the groups. The offset of the outcome shocks of misclassified units is exactly one half at the precision weight, as under residual-based classification, below one half when the proxy is overweighted and above it when the residuals are. The first-order misclassification bias at the precision weight is therefore smaller than that of the residual-based classifier and less than half that of the proxy-based one. The gain is in classification rather than a mechanical bias cancellation. This separates the estimator from the combination of parameter estimates that González-Casasús (2026*a*) proposes as a bias correction. That combination is first-order unbiased only when the two single-view classifiers misclassify at the same rate, a knife-edge condition, and otherwise it can have a bias larger than the one it removes, with a variance inflated by the disagreement between the classifiers. When the proxy is an included regressor, which carries information about the partition precisely when the latent heterogeneity is a fixed effect and misclassification is costly, the misclassified units are also selected on their covariates, and at the precision weight the two offsets cancel the first-order bias altogether. The second contribution concerns imperfect proxies. The same geometry describes a proxy whose group structure is coarser or finer than that of the outcome equation. For a proxy that is misaligned for a fraction of the units, it gives the fraction the estimator tolerates before losing its advantage and the weights that restore it, while the overidentification inherent in two views yields a diagnostic of the shared partition. The third contribution is the oracle case. An exact representation of the CAS estimator as the GFE estimator of a stacked panel with more periods, and with a separation bounded below by the proxy's, shows that the proxy brings the classifier closer to the oracle, in particular when the groups of the outcome equation are weakly separated, the case in which the oracle property is known to be fragile.

Simulations confirm the theory in short panels. With moderately separated groups, GFE misclassifies a non-negligible fraction of the units and its standard confidence interval significantly undercovers. The CAS estimator reduces misclassification and bias by a large factor and raises the coverage closer to the nominal level, with realized offsets above one half as the theory predicts. The validity diagnostic has the right size and detects misaligned proxies, and the bias-aware intervals of González-Casasús (2026*a*) are shorter under CAS. The first of two empirical applications estimates the employment effect of the minimum wage from a panel of about two thousand U.S. counties over eight years, using the employment trajectory of manufacturing, a sector the minimum wage does not reach, as the proxy for the local demand trajectories that generate the groups. The residual-based partition of this short panel fits noise as the number of groups grows, and its bias-aware intervals become uninformative, whereas the two-view rule keeps the elasticity at about -0.05 with intervals that account for classification error. The second application revisits the relationship between income and democracy, with seven five-year periods, using the national savings rate as an auxiliary covariate on the grounds that savings behavior and political development share long-run determinants via cultural patterns. The GFE estimates of the persistence of democracy move between 0.25 and 0.60 with the number of groups, whereas the estimates under the savings-rate partition are stable, around 0.64 for the persistence and 0.10 for the income coefficient, with a cumulative income effect of 0.28 under that proxy. Two alternative proxies, the median age of the population and lagged income itself, are also used as robustness checks.

Related Literature. The paper belongs to the literature on linear grouped panel data models initiated by Bonhomme and Manresa (2015), surveyed in Bonhomme and Denis (2025), and it differs from the existing ways of bringing covariates into the classification. The two-step procedures of Bonhomme, Lamadon and Manresa (2022) and Cheng, Schorfheide and Shao (2025) classify units on moments of the data in a first step and estimate the model on the groups in a second, which uses the covariates but discards the information in the residuals and requires the moments to identify the types. The CAS estimator keeps the residual view, adds the proxy view with a weight that the data determine, and nests both single-view rules. Chetverikov and Manresa (2022) assume that the regressors have a factor structure related to the grouped fixed-effects, which allows a spectral estimator that avoids classification altogether, and this paper shows that, within the classification-based approach, the same structure removes the first-order misclassification bias. Su, Shi and Phillips (2016) classify units on heterogeneous slopes, Aguilar and Boot (2022) allow the error variance to depend on the group, Lumsdaine, Okui and Wang (2023) consider changes in the group structure, and Lewis, Melcangi, Pilossoph and Toner-Rodgers (2023) propose a fuzzy clustering approximation to GFE. Auxiliary information in the form of prior restrictions on the memberships is common in Bayesian treatments (Kim and Wang, 2019, Zhang, 2020, 2022, Liu, 2023, Huang, 2023), whereas the CAS estimator imposes no model for the memberships. In the clustering literature, the classification step is a weighted multi-view k-means problem. Binkiewicz, Vogelstein and Rohe (2017) and Kitamura and Laage (2024) propose covariate-assisted spectral clustering procedures for networks, and Löffler, Zhang and Zhou (2021) study the optimality of spectral clustering in Gaussian mixtures, where the misclassification rate is governed by the same separation-to-noise ratio as here. The contribution relative to that literature is the characterization of the weight that is optimal for a downstream regression parameter rather than for the recovery of the partition. The weighting of views by inverse noise variance is the classical precision weighting of shrinkage and model combination (Green and Strawderman, 1991, Judge and Mittelhammer, 2004, Mittelhammer and Judge, 2005), applied to classification scores rather than to parameter estimates. The measurement-error perspective on misclassification goes back to Aigner et al. (1973), and the distinctive feature of estimated groups is that the error is driven by the same shocks that enter the outcome equation.

Outline. Section 2 presents the model, the two sources of group information and the estimator. Section 3 characterizes classification error and bias under imperfect classification, derives the optimality properties of the precision weight, and compares the estimator with the combination of parameter estimates. Section 4 discusses the oracle case through the stacked representation of the estimator. Section 5 reports the simulations, Section 6 contains the two empirical applications, and Section 7 concludes. Proofs are in the Appendix, and further results are in the Supplemental Appendix.

Notation. For $a \in \mathbb{N}$, $[a] = \{1, \dots, a\}$. $\|\cdot\|$ is the Euclidean norm. Φ and ϕ denote the standard normal distribution and density functions, $\bar{\Phi} = 1 - \Phi$, and $\lambda(c) = \phi(c)/\bar{\Phi}(c)$ is the inverse Mills ratio.

2 Model and Estimator

2.1 Two sources of group information

The model of interest is the linear grouped panel data model

$$y_{it} = x_{it}'\beta + \alpha_{g_i t} + v_{it}, \qquad i \in [N],\ t \in [T], \tag{1}$$

where $g_i \in [G]$ is the latent group of unit i, α_{gt} is the fixed-effect of group g in period t, $x_{it} \in \mathbb{R}^K$ is a vector of covariates and v_{it} an idiosyncratic error. In vector form, $y_i = x_i\beta + \alpha_{g_i} + v_i$, with x_i the $T \times K$ matrix with rows x_{it}' and $\alpha_g = (\alpha_{g1}, \ldots, \alpha_{gT})'$ the time profile of group g. No restriction is placed on the relation between g_i and x_i, so the group effects are genuine fixed-effects. The main focus of this paper is the common slope parameter β, although the group profiles and the partition itself are of interest too.

The additional ingredient is a vector of auxiliary covariates, or proxies, $z_i \in \mathbb{R}^L$ for each unit, with the structure

$$z_i = \zeta_{g_i} + u_i, \qquad i \in [N], \tag{2}$$

where $\zeta_g \in \mathbb{R}^L$ is the center of group g in the proxy space and u_i is proxy noise. The proxies can be a panel variable observed in every period ($L = T$), a set of time-invariant characteristics, several such variables stacked, or an included regressor. The only restriction that links (1) and (2) is that the partition $\{g_i\}_{i\in[N]}$ is the same in both, so that units that share a time path of unobserved heterogeneity in the outcome equation share a center in the proxy space. The centers ζ_g are otherwise unrestricted and need not be related to the profiles α_g in location, scale or shape. This is the sense in which the proxies share the latent group structure, and it is weaker than assuming that they measure the profiles α_g. For example, in the second application of Section 6, the outcome is a democracy index and the proxy is the national savings rate. Both are assumed to be shaped by a country's latent type, but a savings rate is not a measurement of a propensity to democratize.

Throughout, the proxy noise u_i is assumed to be independent of the outcome error v_i given (x_i, g_i), and the proxies are either excluded from the outcome equation or already included in it, so that their use in the classification step raises no omitted variable concern. When z_i is an included regressor, the independence of u_i and v_i is the strict exogeneity of that regressor, and this case is treated separately in Section 3.4.

2.2 The CAS estimator

Given a weight $\delta \in [0, 1]$ on the residual view, consider the weighted least squares objective

$$Q_\delta(\beta, \alpha, \zeta, g) = \sum_{i=1}^{N} \left\{ \delta \left\| y_i - x_i\beta - \alpha_{g_i} \right\|^2 + (1-\delta) \left\| z_i - \zeta_{g_i} \right\|^2 \right\}, \tag{3}$$

with $\beta \in \mathcal{B}$, $\alpha = (\alpha_1, \ldots, \alpha_G) \in \mathcal{A}^{T\times G}$, $\zeta = (\zeta_1, \ldots, \zeta_G) \in \mathbb{R}^{L\times G}$ and $g \in [G]^N$. The objective is that of the GFE estimator of Bonhomme and Manresa (2015) augmented with a second equation that has no covariates and its own group centers, and its role is to motivate the classification rule. For given (β, α, ζ), the assignment that minimizes (3) is

$$\hat{g}_i(\delta) = \underset{h\in[G]}{\arg\min}\ \delta \left\| y_i - x_i\beta - \alpha_h \right\|^2 + (1-\delta) \left\| z_i - \zeta_h \right\|^2, \tag{4}$$

a weighted two-view k-means rule in which each unit is assigned to the group whose pair of centers is closest in the weighted distance. The rule compares two squared distances per group, one in the space of residual trajectories and one in the proxy space, each measured from the center of that group in that space.

A covariate-assisted estimator is the least squares estimator of the outcome equation on the groups that this rule produces for a given δ. Given a partition $\hat{g}(\delta)$, let $\check{y}_{it}$ and $\check{x}_{it}$ denote the outcome and the covariates

after removing $(\hat{g}_i, t)$-cell means. The estimator of β is

$$\hat{\beta} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \check{x}_{it} \check{x}_{it}' \right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \check{x}_{it} \check{y}_{it}, \tag{5}$$

the least squares estimator of Bonhomme and Manresa (2015) with the partition $\hat{g}(\delta)$ treated as data, and the profiles $\hat{\alpha}_{gt}$ are the means of $y_{it} - x_{it}'\hat{\beta}$ in the cells (g, t). The proxy centers $\hat{\zeta}_g$, the means of z_i in the estimated groups, describe the groups in the proxy space but play no role in the estimation of β beyond the classification step. The definition covers the whole range of weights. For $\delta \in (0, 1]$, the joint minimizer of (3) has this form, since at the minimum the partition is the assignment (4) evaluated at the least squares parameters and the parameters are the least squares solution on that partition. With $\delta = 1$, the estimator is the GFE estimator, whose classification is residual-based. With $\delta = 0$, the objective does not involve β at all, the partition is the k-means partition of the proxies, and the estimator classifies units on the proxies alone before running least squares on the estimated groups, which is a proxy-based classifier in the terminology of González-Casasús (2026*a*). Between the two extremes, both sources of group information enter the partition, and in every case the estimator of β is the least squares estimator (5) on that partition.

While one could view δ as a tuning parameter that encodes a prior on the importance of the auxiliary data, there is an optimal value for it in a specific class of models, and that value is estimable. Let σ_v^2 and σ_u^2 denote the idiosyncratic variances of the two views, and consider the Gaussian quasi-likelihood of the two equations with homoskedastic errors. Up to constants, its negative logarithm is

$$\mathcal{L}(\beta, \alpha, \zeta, g, \sigma_v^2, \sigma_u^2) = NT \log \sigma_v^2 + \frac{1}{\sigma_v^2} \sum_{i=1}^{N} \|y_i - x_i\beta - \alpha_{g_i}\|^2 + NL \log \sigma_u^2 + \frac{1}{\sigma_u^2} \sum_{i=1}^{N} \|z_i - \zeta_{g_i}\|^2 . \tag{6}$$

For given variances, minimizing (6) over the remaining arguments is the same as minimizing Q_δ with

$$\delta^* = \frac{\sigma_u^2}{\sigma_v^2 + \sigma_u^2}, \tag{7}$$

the precision weight, under which each view is weighted by the inverse of its noise variance. For given $(\beta, \alpha, \zeta, g)$, the minimizing variances are the within-group mean squared residuals of each view, $\hat{\sigma}_v^2 = (NT)^{-1} \sum_i \|y_i - x_i\beta - \alpha_{g_i}\|^2$ and $\hat{\sigma}_u^2 = (NL)^{-1} \sum_i \|z_i - \zeta_{g_i}\|^2$. The CAS estimator is the least squares estimator (5) on the partition of the joint minimizer of (6), the quasi-maximum likelihood (QML) solution, with $\hat{\delta} = \hat{\sigma}_u^2/(\hat{\sigma}_v^2 + \hat{\sigma}_u^2)$ as its implied weight.[1]

Since (6) at given variances is proportional to Q_δ at $\delta = \sigma_u^2/(\sigma_v^2 + \sigma_u^2)$, the QML solution minimizes $Q_{\hat{\delta}}$ at its own implied weight. The CAS estimator is therefore the covariate-assisted estimator with weight $\hat{\delta}$, and $\hat{\delta}$ is a fixed point of the map that takes a weight to the precision weight implied by the minimizer of Q_δ, the one at which (6) is smallest if there are several, and nothing below requires it to be unique. What the quasi-likelihood adds to the fixed-weight objective is the choice of the weight, and it makes that choice without any normalization of the views. Concentrated over the parameters and the variances for a given partition, (6) becomes $NT \log \hat{\sigma}_v^2(g) + NL \log \hat{\sigma}_u^2(g)$, which is invariant to the scale of each view, whereas

[1] The minimizer exists because there are finitely many partitions, provided that no partition fits a view exactly, which holds when the proxy is continuously distributed and G is smaller than the number of distinct proxy vectors, and similarly for the residuals. With discrete proxies or ties, the implementation used below floors each variance at a negligible fraction of the total variance of its view.

Q_δ concentrated over the parameters is $\delta NT\hat{\sigma}_v^2(g) + (1-\delta)NL\hat{\sigma}_u^2(g)$, which is not. Section 3 shows that the precision weight is also the weight that maximizes the accuracy of the classification step, so the quasi-likelihood criterion and the classification objective agree.

The estimator is computed by iterating between the assignment and the updating of the parameters and the variances. In the algorithm below, (w_y, w_z) denote the weights that multiply the two squared distances in the assignment rule,

$$\hat{g}_i = \underset{h\in[G]}{\arg\min}\ w_y \left\|y_i - x_i\beta - \alpha_h\right\|^2 + w_z \left\|z_i - \zeta_h\right\|^2, \tag{8}$$

so that $(w_y, w_z) = (\delta, 1-\delta)$ reproduces (4) and $(w_y, w_z) = (1/\sigma_v^2, 1/\sigma_u^2)$ reproduces the assignment implied by (6).

Algorithm 1. *Estimation Procedure*

1. *Compute a preliminary $\beta^{(0)}$ (e.g., by pooled least squares or equal to zero), draw G units at random and use their residual trajectories $y_i - x_i\beta^{(0)}$ and proxy vectors z_i as initial centers $(\alpha^{(0)}, \zeta^{(0)})$. Set the initial weights to the inverse total variances of the two views, $w_y = 1/\mathrm{var}(y_{it} - x_{it}'\beta^{(0)})$ and $w_z = 1/\mathrm{var}(z_{il})$, and obtain $g^{(0)}$ by (8). Set $k = 0$.*
2. *Given $g^{(k)}$, update $\beta^{(k+1)}$ by least squares on $(g^{(k)}, t)$-demeaned data, $\alpha^{(k+1)}$ and $\zeta^{(k+1)}$ as cell means, and $\hat{\sigma}_v^{2(k+1)}$, $\hat{\sigma}_u^{2(k+1)}$ as the within mean squared residuals of each view. Set $w_y = 1/\hat{\sigma}_v^{2(k+1)}$ and $w_z = 1/\hat{\sigma}_u^{2(k+1)}$.*
3. *Given $(\beta^{(k+1)}, \alpha^{(k+1)}, \zeta^{(k+1)})$ and (w_y, w_z), update $g^{(k+1)}$ by (8).*
4. *Stop when $g^{(k+1)} = g^{(k)}$. Otherwise set $k = k+1$ and return to Step 2.*

Each step weakly decreases (6), because the assignment, the parameters and the variances are exact minimizers given the other arguments, and there are finitely many partitions, so the algorithm converges in a finite number of iterations. As with any k-means type procedure, it converges to a local minimum. In practice, it is run from many random starts, keeping the solution with the smallest value of (6), and the best solution can be refined by the variable neighborhood search of Bonhomme and Manresa (2015), which perturbs the memberships of a few units at a time and reiterates, with (6) as the objective. For a fixed weight δ, the weights are held at $(w_y, w_z) = (\delta, 1-\delta)$ throughout, the variances are not updated, and the solution kept is the one with the smallest value of Q_δ. The computational cost per start is that of the GFE algorithm with $T + L$ periods.

Several features of the estimator are worth noting. First, the assignment (4) with the precision weight is invariant to the location and the scale of each view. Rescaling z_i by a constant rescales $\hat{\sigma}_u^2$ by its square and leaves the weighted distance unchanged, and group-specific shifts in either view are absorbed by the centers. The group-wise standardization that a direct comparison of the two views would require is therefore unnecessary. Second, with J proxy blocks $z_i^{(1)}, \ldots, z_i^{(J)}$ of different nature or scale, (6) has one variance per block and each block is weighted by its own precision. A block with a large noise variance relative to its scale contributes little, and a block that is more precisely measured contributes more. Blocks with heteroskedastic or correlated noise can be whitened by a within-group covariance estimate, in which case the squared distances become Mahalanobis distances (Remark 4).

Third, when the proxy is a noisy measurement of the profiles themselves, $\zeta_g = \alpha_g$ and $L = T$, the assignment (4) with the true centers coincides with k-means on the convex combination $\delta(y_i - x_i\beta) + (1-\delta)z_i$, and δ^* is the weight that minimizes the mean squared error of this combination as an estimator of the common profile. The estimator classifies units on the residual-based measurement shrunk towards the proxy-based one, which is the covariate-assisted shrinkage the acronym CAS refers to. Section 3 shows that the same weight is optimal for classification in general, where it maximizes the separation-to-noise ratio of the assignment rule.

Fourth, when T is large relative to G, each view may be replaced by its projection on its leading G principal components before clustering, as in Chetverikov and Manresa (2022). The geometry of Section 3 then applies to the projected views, and the results below are stated for the raw views, which is the version that nests the GFE estimator.

2.3 Misclassification and the parameter of interest

The accuracy of the classification step is measured throughout by the misclassification rate

$$\ell(\hat{g}, g) = \min_{\gamma \in \Gamma} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{\gamma(\hat{g}_i) \neq g_i\}}, \tag{9}$$

where Γ is the set of permutations of $[G]$, which accounts for the arbitrary labeling of the groups. In the oracle case discussed in Section 4, $\ell(\hat{g}, g)$ vanishes faster than sampling error, yet it is often non-zero in finite samples, as the simulations of Section 5 report. The misclassification rate is the first of the two channels through which the classifier affects the estimator of β. The second channel is the composition of the misclassified units. The least squares step treats $\mathbb{1}_{\{\hat{g}_i = h\}}$ as data, so a misclassified unit is attributed the profile of its destination group, and the cost of the error for β depends on how far the unit's own outcome is from that profile and on how its covariates differ from those of the destination group. Two classifiers with the same misclassification rate can therefore induce different biases, and Section 3 quantifies both channels for the two-view rule.

2.4 A diagnostic for the validity of the proxy

The shared partition is the only restriction that links (1) and (2), and it is an overidentifying restriction, since each view identifies a partition on its own. This suggests a diagnostic. Let $\hat{g}_1$ and $\hat{g}_0$ denote the single-view partitions, obtained by minimizing (3) at $\delta = 1$ and at $\delta = 0$, and let $\hat{\sigma}_v^2(g)$ and $\hat{\sigma}_u^2(g)$ be the within-group variances of the two views under a partition g, as defined after (7). Concentrated over the parameters and the variances, the quasi-likelihood (6) is $NT \log \hat{\sigma}_v^2(g) + NL \log \hat{\sigma}_u^2(g)$ under a shared partition, which the CAS partition $\hat{g}$ minimizes, and $NT \log \hat{\sigma}_v^2(g_y) + NL \log \hat{\sigma}_u^2(g_z)$ when each view is allowed its own partition, which $(\hat{g}_1, \hat{g}_0)$ minimizes. The quasi-likelihood ratio statistic

$$\mathrm{LR} = NT \log \frac{\hat{\sigma}_v^2(\hat{g})}{\hat{\sigma}_v^2(\hat{g}_1)} + NL \log \frac{\hat{\sigma}_u^2(\hat{g})}{\hat{\sigma}_u^2(\hat{g}_0)} \tag{10}$$

is the loss of fit that the restriction imposes, and it is nonnegative up to the local optima of the three problems. Under a valid proxy, the three partitions estimate the same object and differ only through misclassification, so LR is governed by the misclassification probabilities of the single-view rules. Under an invalid proxy, the units whose group in the proxy equation differs from their group in the outcome equa-

tion cannot be fitted well by both views at once, and LR grows with their number. The null distribution of LR is not standard, because it depends on how many units the single-view rules misclassify and on how much idiosyncratic noise their partitions fit. I calibrate it by a parametric bootstrap under the fitted two-view model, which imposes validity. Draw $v_i^* \sim N(0, \hat{\sigma}_v^2 I_T)$ and $u_i^* \sim N(0, \hat{\sigma}_u^2 I_L)$, or resample the within-group residuals of the fitted model, set $y_i^* = x_i\hat{\beta} + \hat{\alpha}_{\hat{g}_i} + v_i^*$ and $z_i^* = \hat{\zeta}_{\hat{g}_i} + u_i^*$, recompute the three partitions and the statistic, and report the share of bootstrap draws in which the statistic exceeds LR as the p-value. The bootstrap reproduces both the misclassification of the single-view rules and their noise fitting under the maintained model, and Section 5.3 shows that it has the right size and power against the misaligned designs of Supplemental Appendix S.5. A proxy without group structure, whose centers coincide, is not an alternative for this diagnostic, since it does not contradict the shared partition. This is as it should be, because such a proxy is harmless at the population level (Remark 5), although in finite samples the k-means fit of its noise can misclassify more units than the residual view alone, as the simulations of Section 5.3 show.

3 Classification Error and Bias

When classification is perfect, auxiliary data cannot improve the estimation. The partition is the only object that the classification step delivers, and once it is recovered the estimator is the oracle. Group memberships are discrete, so there is no efficiency consideration attached to how they are estimated, and no gain is left once they are right. This is the oracle paradigm, which can be expected to hold when the time dimension or the separation-to-noise ratio is large, and in which auxiliary data are redundant and inconsequential. Otherwise, classification is imperfect, the estimator inherits a bias from the misclassified units, and there are gains from incorporating auxiliary data into the classification problem. This is the regime that motivates the CAS estimator, and it is the regime that the local misclassification framework of González-Casasús (2026*a*) describes.

This section characterizes the classification error of the two-view rule and the bias it induces in $\hat{\beta}$ when the time dimension is small and classification errors are of the same magnitude as sampling uncertainty, which is the regime in which the choice of classifier matters. Section 3.1 states the first-order description of the estimator under imperfect classification that serves as the starting point. Section 3.2 derives the misclassification probabilities and the shock offsets of the two-view rule in a Gaussian design. Section 3.3 combines the two to establish the optimality properties of the precision weight, Section 3.4 treats the case in which the proxy is an included regressor, and Section 3.5 compares the estimator with the combination of parameter estimates obtained under different classifiers.

3.1 Imperfect classification and the bias of the common parameters

The starting point is a first-order description of the estimator of β under imperfect classification, which González-Casasús (2026*a*) derives from primitive conditions and I take here as a high-level assumption to streamline the analysis. Let $\pi_g = \mathbb{P}(g_i = g)$, $\mu_g = \mathbb{E}[x_i | g_i = g] \in \mathbb{R}^{T\times K}$ and $S = \sum_g \pi_g \mathbb{E}[(x_i - \mu_g)'(x_i - \mu_g)|g_i = g]$, the average within-group covariance of the covariates. A classifier $\hat{g}$ is described by three sets of constants, one for every ordered pair of groups $a \neq b$. The misclassification probability $p_{a|b} = \mathbb{P}(\hat{g}_i = a \mid g_i = b)$ is the probability that a unit of group b is assigned to group a. The shock offset $\bar{v}_{ab} \in \mathbb{R}^T$ is the average outcome error of the units misclassified from b to a, and the covariate shift

$s_{ab} \in \mathbb{R}^{T\times K}$ is the average deviation of their covariates from the group mean μ_b.

Assumption 1 (First-order description under imperfect classification)**.** *For the classifier $\hat{g}$ with constants $(p_{a|b}, \bar{v}_{ab}, s_{ab})_{a\neq b}$, the least squares estimator* (5) *on the estimated groups satisfies*

$$\mathbb{E}[\hat{\beta}] = \beta + \mathrm{Bias}(\hat{\beta}), \qquad \mathrm{Var}(\hat{\beta}) = V, \tag{11}$$

where V is the variance of the oracle estimator that uses the true groups and

$$\mathrm{Bias}(\hat{\beta}) = \sum_{a=1}^{G}\sum_{b\neq a}(\kappa_{ab} - \kappa^s_{ab})\, p_{a|b}, \qquad \kappa_{ab} = S^{-1}\pi_b(\mu_a - \mu_b)'\gamma_{ab}, \quad \kappa^s_{ab} = S^{-1}\pi_b s'_{ab}(\alpha_a - \alpha_b), \tag{12}$$

with $\gamma_{ab} = \alpha_a - \alpha_b - \bar{v}_{ab}$ the unabsorbed gap of a unit misclassified from b to a.

The assumption is a stylized description of the sampling distribution of $\hat{\beta}$, to be read as the first-order approximation that becomes exact along the local misclassification limit of González-Casasús (2026*a*), where $\hat{\beta}$ is moreover normally distributed. Only the first two moments are used below, so the assumption is stated in those terms, and normality is invoked only when a statement concerns the coverage of confidence intervals, where it follows from the local limit. The local approximation concerns the estimated memberships, whatever rule produced them, so the assumption applies to the two-view rule once its probabilities, offsets and shifts are computed, which is the task of Section 3.2. The arguments below exploit the structure of (11), namely that classification error shifts the distribution of $\hat{\beta}$ without changing its dispersion, by an amount that is linear in the misclassification probabilities with coefficients that depend on the compositional contrast of the covariates and on the part of the profile gap that least squares has to absorb.

A classifier has proportional endogenous misclassification with offset ρ_{ab} if $\bar{v}_{ab} = \rho_{ab}(\alpha_a - \alpha_b)$, in which case $\gamma_{ab} = (1 - \rho_{ab})(\alpha_a - \alpha_b)$. Proxy-based classifiers have $\rho_{ab} = 0$ and, under proxy noise independent of the covariates, $s_{ab} = 0$. Residual-based k-means classifiers have $\rho_{ab} = 1/2$ for the pairs that carry misclassification mass. The misclassification probabilities $p_{a|b}$ are not identified from the data, but they are what a better classifier reduces, and the bias-aware confidence intervals of González-Casasús (2026*a*) are built from bounds on them.

One qualification concerns the covariate shift. In general, s_{ab} is not identified, which is why the inference procedures of González-Casasús (2026*a*) are confined to classifiers that do not select units on the included regressors. Classifiers that use an included regressor are within the scope of Assumption 1 but outside the scope of those inference procedures. Section 3.4 shows that, for the two-view rule, the shift is pinned down by the geometry, so that the bias is known in that case too.

3.2 The geometry of the two-view rule

I now compute the misclassification probabilities and the offsets of the two-view rule in a Gaussian design. To isolate the geometry, the computation conditions on the population centers, as in the motivating analysis of classification error of Bonhomme and Manresa (2015). The estimated centers of the k-means solution shift the decision boundaries by an amount of the order of the estimation error, which is negligible relative to the distance between centers in the local misclassification regime, as Proposition 1 below and Supplemental Appendix S.1 relate to the primitive conditions of González-Casasús (2026*a*).

Assumption 2 (Gaussian two-view design). *Conditional on* (x_i, g_i), $v_i \sim N(0, \sigma_v^2 I_T)$ *and* $u_i \sim N(0, \sigma_u^2 I_L)$ *are independent, with* $\sigma_v^2, \sigma_u^2 > 0$.

For an ordered pair of groups (g, h) with $h \neq g$, let $\Delta_{hg} = \alpha_h - \alpha_g$, $\nabla_{hg} = \zeta_h - \zeta_g$, $a_{hg} = \|\Delta_{hg}\|^2$ and $b_{hg} = \|\nabla_{hg}\|^2$, and define the pairwise event that a unit of group g is closer, in the weighted distance (4) with population centers, to the centers of group h than to its own,

$$E_{hg}(\delta) = \left\{ \delta \|v_i - \Delta_{hg}\|^2 + (1-\delta) \|u_i - \nabla_{hg}\|^2 < \delta \|v_i\|^2 + (1-\delta)\|u_i\|^2 \right\}. \tag{13}$$

With $G = 2$, $E_{hg}(\delta)$ is the misclassification event. With $G \geq 3$, it contains the event that the unit is assigned to h, since the latter requires h to beat every other group, so its probability bounds the misclassification probability $p_{h|g}$ of Section 3.1 from above, with equality when $G = 2$. I call $\bar{p}_{h|g}(\delta) = \mathbb{P}(E_{hg}(\delta) \mid g_i = g)$ the pairwise bound, or the pairwise misclassification probability when $G = 2$. Define

$$\tau_{hg}(\delta) = \delta a_{hg} + (1-\delta) b_{hg}, \qquad V_{hg}(\delta) = \delta^2 a_{hg} \sigma_v^2 + (1-\delta)^2 b_{hg} \sigma_u^2, \qquad \mathrm{SNR}_{hg}(\delta) = \frac{\tau_{hg}(\delta)}{\sqrt{V_{hg}(\delta)}}. \tag{14}$$

Here, $\tau_{hg}(\delta)$ is the weighted squared distance between the two pairs of centers, $V_{hg}(\delta)$ is the variance of the weighted noise projected on the contrasts, and $\mathrm{SNR}_{hg}(\delta)$ is their ratio. At $\delta = 1$, $\mathrm{SNR}_{hg}(1) = \sqrt{a_{hg}/\sigma_v^2}$ is the separation-to-noise ratio of the residual view, and at $\delta = 0$, $\mathrm{SNR}_{hg}(0) = \sqrt{b_{hg}/\sigma_u^2}$ is that of the proxy view.

Theorem 1 (Geometry of the two-view rule). *Let Assumption 2 hold, fix* $h \neq g$ *with* $a_{hg}, b_{hg} > 0$, *and write* $c_{hg}(\delta) = \mathrm{SNR}_{hg}(\delta)/2$, *the separation-to-noise ratio of the midpoint between the two pairs of centers. For every* $\delta \in [0, 1]$, *the following hold.*

(a) (Probability) $\bar{p}_{h|g}(\delta) = \mathbb{P}\left(E_{hg}(\delta) \mid g_i = g\right) = \bar{\Phi}\left(c_{hg}(\delta)\right)$, *and* $p_{h|g}(\delta) = \bar{p}_{h|g}(\delta)$ *when* $G = 2$.

(b) (Additivity) $\mathrm{SNR}_{hg}(\delta)^2 \leq a_{hg}/\sigma_v^2 + b_{hg}/\sigma_u^2 = \mathrm{SNR}_{hg}(1)^2 + \mathrm{SNR}_{hg}(0)^2$, *with equality if and only if* $\delta = \delta^* = \sigma_u^2/(\sigma_v^2 + \sigma_u^2)$. *In particular,* δ^* *depends neither on the pair* (g, h) *nor on the separations* a_{hg}, b_{hg}.

(c) (Offsets) $\mathbb{E}\left[v_i \mid E_{hg}(\delta), g_i = g\right] = \rho_{hg}(\delta) \Delta_{hg}$ *and* $\mathbb{E}\left[u_i \mid E_{hg}(\delta), g_i = g\right] = \rho_{hg}^z(\delta) \nabla_{hg}$, *with*

$$\rho_{hg}(\delta) = \frac{\delta \sigma_v^2 \, \lambda(c_{hg}(\delta))}{\sqrt{V_{hg}(\delta)}}, \qquad \rho_{hg}^z(\delta) = \frac{(1-\delta)\sigma_u^2 \, \lambda(c_{hg}(\delta))}{\sqrt{V_{hg}(\delta)}}. \tag{15}$$

Both are bounded below by their local misclassification values

$$\bar{\rho}_{hg}(\delta) = \frac{\delta \sigma_v^2 \, \tau_{hg}(\delta)}{2 V_{hg}(\delta)} \quad \text{and} \quad \bar{\rho}_{hg}^z(\delta) = \frac{(1-\delta)\sigma_u^2 \, \tau_{hg}(\delta)}{2 V_{hg}(\delta)},$$

strictly except at the endpoint at which the offset vanishes, where both are zero, and they approach them in the local misclassification limit $c_{hg}(\delta) \to \infty$. *The local misclassification values satisfy*

$$\bar{\rho}_{hg}(\delta^*) = \bar{\rho}_{hg}^z(\delta^*) = \frac{1}{2}, \qquad \bar{\rho}_{hg}(1) = \frac{1}{2}, \qquad \bar{\rho}_{hg}(0) = 0, \tag{16}$$

and $\bar{\rho}_{hg}(\delta) < 1/2$ *for* $\delta \in (0, \delta^*)$ *while* $\bar{\rho}_{hg}(\delta) > 1/2$ *for* $\delta \in (\delta^*, 1)$.

Part (a) says that the weighted rule misclassifies a unit when a scalar Gaussian discriminant, the weighted sum of the projections of the two noises on the respective contrasts, exceeds half the weighted squared distance between the pairs of centers. The probability is governed by the separation-to-noise ratio $\mathrm{SNR}_{hg}(\delta)$, and part (b) is the central fact. At the precision weight, the squared ratios of the two views add up, so the two-view rule is strictly more accurate than either single-view rule whenever both views carry some group signal, and the gain is largest when the two views are of comparable strength. Since $\bar{\Phi}$ is strictly decreasing, δ^* minimizes the misclassification probability of every pair simultaneously. The weight does not depend on how strong the group signal is in each view, because the squared distances in (4) already scale with the signal. It only corrects for the different noise levels, which is why it is estimable without knowledge of the groups.

Part (c) describes the units that are misclassified. Conditional on misclassification, the discriminant lies at the threshold plus an overshoot, and Gaussian regression of each noise on the discriminant attributes to the outcome shock a share of the threshold proportional to its contribution to the variance of the discriminant. At δ^*, the two contributions are equal and each shock absorbs exactly one half of its own gap, as the residual-based k-means rule does at $\delta = 1$. Overweighting the proxy ($\delta < \delta^*$) attributes a larger share of the misclassification to the proxy noise, and the outcome shock absorbs less than one half of the gap. Overweighting the residuals ($\delta > \delta^*$) attributes a larger share to the outcome shock, which then absorbs more than one half. The inverse Mills ratio in (15) is the overshoot of the discriminant beyond the threshold. It makes the finite-separation offsets larger than their local misclassification values, as in the one-view case, and it vanishes relative to the threshold in the local misclassification limit.

Remark 1 (Beyond Gaussian errors)**.** *Gaussianity gives closed forms for the probabilities and the offsets, yet part (b) does not depend on it. For any errors that are homoskedastic and uncorrelated within each view,* $\mathrm{SNR}_{hg}(\delta)$ *is the ratio of the decision threshold to the standard deviation of the discriminant* $D_i = \delta\Delta_{hg}'v_i + (1-\delta)\nabla_{hg}'u_i$*, and the Cauchy–Schwarz argument shows that* δ^* *maximizes it. For sub-Gaussian errors, the misclassification probability is bounded by* $\exp(-c\,\mathrm{SNR}_{hg}(\delta)^2)$*, so the precision weight maximizes the exponent of the bound. Parts (a) and (c) can be weakened along the lines of González-Casasús (2026a), whose microfoundation allows spherically symmetric errors with light tails, but for the two-view rule that class reduces to the Gaussian case, because the stacked error at the precision weight has independent blocks and a vector with independent spherical blocks is spherical only when it is Gaussian (Kagan, Linnik and Rao, 1973, Theorem 5.3.3). Supplemental Appendix S.1 gives the primitive conditions, and Supplemental Appendix S.5 reports simulations with t-distributed, heteroskedastic and serially correlated errors.*

The analysis above conditions on the population centers. The next proposition shows that, at the precision weight, the k-means solution with estimated centers falls under the conditions in Supplemental Appendix S.2 of González-Casasús (2026*a*), which derive Assumption 1 for k-means classifiers with estimated centers.

Proposition 1 (Estimated centers)**.** *Let Assumption 2 hold and define the stacked residual vectors* $\tilde{r}_i = (\sqrt{\delta^*}(y_i - x_i\beta)', \sqrt{1-\delta^*}z_i')' \in \mathbb{R}^{T+L}$*. The assignment* (4) *with weight* δ^* *is the* k*-means assignment of* $\tilde{r}_i$ *to the stacked centers* $\theta_h = (\sqrt{\delta^*}\alpha_h', \sqrt{1-\delta^*}\zeta_h')'$*, and the stacked error* $\tilde{v}_i = (\sqrt{\delta^*}v_i', \sqrt{1-\delta^*}u_i')'$ *is spherical,* $\tilde{v}_i \sim N(0, \sigma^2 I_{T+L})$ *with* $\sigma^2 = \sigma_v^2\sigma_u^2/(\sigma_v^2+\sigma_u^2)$*. Consequently, the stacked error satisfies the spherical light-tailed error condition of Assumption S.2.1(a) in González-Casasús (2026a), and under the remaining conditions*

of that assumption for the stacked model, Theorem S.2.1 there derives Assumption 1, with the one-half offset, for the k-means rule with estimated centers in the stacked problem. Its offset $\bar{v}_{hg} = \frac{1}{2}(\theta_h - \theta_g) + o(1)$ projects on the outcome block to $\frac{1}{2}\sqrt{\delta^*}\Delta_{hg}$, which is $\rho_{hg} = 1/2$ in the original scale. For $\delta \neq \delta^*$, the stacked error is not spherical, and the offsets are those of Theorem 1(c).

Proposition 1 is the formal link between the two-view rule and the local misclassification framework. Proposition S.1 in Supplemental Appendix S.1 transposes the drifting design of González-Casasús (2026*a*) to the two-view rule, under which the rule with estimated centers and a preliminary weight of a polynomial rate satisfies Assumption 1, and shows that the updated weight has that rate. Corollary S.1 there verifies the rate condition for the fixed points of Algorithm 1 in a shrinking neighborhood of the truth. The precision-weighted rule is the GFE classifier applied to a stacked panel with spherical errors, so the properties established for the GFE classifier, including the behavior of its estimated centers, carry over to the stacked problem, with the qualification about the identification condition noted in Section 4. The stacked representation also explains the role of the precision weight from a different angle. It is the weight that equalizes the noise variances across the two blocks of the stacked system, which is what the one-half rule requires.

3.3 Optimality of the precision weight

Combining Theorem 1 with the bias formula (12) yields the first-order bias of a covariate-assisted estimator as a function of the weight δ. For an external proxy whose noise is independent of the covariates, the covariate shift s_{ab} vanishes for every δ, as it does for both the residual-based and the proxy-based classifiers. The misclassification probabilities are those of Theorem 1(a), and the offsets are the local misclassification values of Theorem 1(c), which are the constants of the local limit that justifies Assumption 1. The following corollary states the result for two groups, where every object has a closed form, and Remark 3 discusses the general case.[2]

Corollary 1 (First-order bias). *Let $G = 2$ and let Assumption 1 hold for the classifier $\hat{g}(\delta)$ with probabilities $p(\delta) = p_{2|1}(\delta) = p_{1|2}(\delta) = \bar{\Phi}(c(\delta))$, offsets $\bar{\rho}(\delta) = \bar{\rho}_{21}(\delta) = \bar{\rho}_{12}(\delta)$ and $s_{ab} = 0$. Then*

$$\operatorname{Bias}(\hat{\beta}(\delta)) = (1 - \bar{\rho}(\delta))\, p(\delta)\, \varkappa, \qquad \varkappa = S^{-1} \sum_{t=1}^{T} (\mu_{2t} - \mu_{1t})\, (\alpha_{2t} - \alpha_{1t})\,, \tag{17}$$

and the variance of $\hat{\beta}(\delta)$ is the oracle variance V for every δ. In particular, whenever $\varkappa \neq 0$ and $a_{21}, b_{21} > 0$,

$$\|\operatorname{Bias}(\hat{\beta}(\delta^*))\| = \tfrac{1}{2} p(\delta^*) \|\varkappa\| < \tfrac{1}{2} p(1) \|\varkappa\| = \|\operatorname{Bias}(\hat{\beta}(1))\|, \quad \|\operatorname{Bias}(\hat{\beta}(\delta^*))\| < \tfrac{1}{2} \|\operatorname{Bias}(\hat{\beta}(0))\|. \tag{18}$$

The corollary separates the two channels through which the classifier affects the bias. The probability $p(\delta)$ is minimized at δ^* by Theorem 1(b), and the offset at δ^* equals the one-half offset of residual-based classification by Theorem 1(c). The precision-weighted estimator thus dominates the GFE estimator in first-order bias, at the same first-order variance, hence in mean squared error and in the coverage of standard confidence intervals to the order of Assumption 1.[3] It dominates proxy-based classification by more than

[2] For a weight other than δ^*, the primitive conditions of Supplemental Appendix S.1 are only sketched, so statements at a fixed $\delta \neq \delta^*$ are made under Assumption 1 with the offsets of Theorem 1(c).

[3] In finite samples, the dispersion of the estimators may also depend on the classifier, through the estimation of the profiles and

a factor of two, because the latter both misclassifies more units and absorbs none of the gap. The same ordering applies to the bias-aware confidence intervals of the local misclassification framework, which are built from bounds on the probabilities and from the offset, since a lower probability and the same offset produce a shorter interval at the same coverage. The corollary is the formal content of the claim that a researcher who is interested only in β is still better off using the auxiliary covariates, provided they are used in the two-view rule rather than on their own. The exact offsets (15) can be substituted for $\bar{\rho}(\delta)$ in (17) as a finite-sample refinement that accounts for the overshoot, and Section 5.2 shows that the refinement tracks the simulations.

Remark 2 (Bias-aware inference)**.** *The bias-aware confidence intervals of González-Casasús (2026a) apply to the CAS estimator without modification. They are built from the sensitivities κ_{ab} of* (12)*, computed at the estimated groups with lower bounds on the offset and upper bounds on the outgoing misclassification probabilities of each group, which that paper calibrates for k-means rules from the soft assignments of the units to the estimated centers. Since the precision-weighted rule is the k-means rule of the stacked problem (Proposition 1), the same calibration applies with the stacked distances, and the resulting intervals are shorter than under GFE whenever the calibrated bounds are smaller, which Corollary 1 predicts and Section 5.3 and Supplemental Appendix S.7 confirm.*

Minimizing classification error and minimizing bias are, however, different objectives. The product $(1 - \bar{\rho}(\delta))p(\delta)$ in (17) is not minimized at δ^*. By Theorem 1(c), $1 - \bar{\rho}(\delta)$ keeps decreasing for δ slightly above δ^*, where $p(\delta)$ is flat, so the bias-minimizing weight δ_b satisfies $\delta_b > \delta^*$ and overweights the residuals. When the proxy is much stronger than the residuals, $\bar{\rho}(\delta)$ exceeds one for intermediate weights, which with $\sigma_v^2 = \sigma_u^2$ happens if and only if $b/a > 8$, that is, $\mathrm{SNR}(0) > 2\sqrt{2}\,\mathrm{SNR}(1)$. There is then a weight at which the first-order bias changes sign and vanishes. This is a real feature of the two-view rule. Units that are misclassified despite a heavy weight on a weak residual view are those whose outcome shocks are so large that they overshoot the gap. The bias-minimizing weight is computable from the Gaussian geometry as $\delta_b = \arg\min_\delta |1 - \bar{\rho}(\delta)|\,\bar{\Phi}(c(\delta))$ with $(a, b, \sigma_v^2, \sigma_u^2)$ replaced by estimates, but it relies on the Gaussian design, on the linearity of the bias in the probabilities and on the rarity of misclassification, which are exactly the conditions under which the gain over δ^* is small. The simulations of Section 5 show a flat bias profile between δ^* and δ_b and a root mean squared error that is essentially minimized at δ^*. I therefore recommend δ^* as the default and δ_b as a sensitivity check.

Remark 3 (More than two groups)**.** *With $G \geq 3$, a unit of group g is assigned to h when its stacked score falls in the Voronoi cell of θ_h, which is a subset of the half-space whose probability is the pairwise probability $\bar{p}_{h|g}(\delta)$ of Theorem 1(a). Theorem 1 therefore orders the pairwise probabilities, which bound the misclassification probabilities from above, and not the misclassification probabilities themselves. The two coincide to first order when the misclassification mass of the cell concentrates near the midpoint of $[\theta_g, \theta_h]$, which happens in the local misclassification limit when no third center lies in the ball with that diameter. This ball condition holds for the pairs of centers that are nearest neighbors, which carry the misclassification mass in the limit, and it is a condition on the stacked centers, so it has to be checked at each weight. For a pair that satisfies it at the three weights, the flow has a smaller misclassification probability at δ^* than at $\delta = 1$ or $\delta = 0$ to first order, with the same offset as at $\delta = 1$. Since the flow-specific terms of* (12) *can have different signs, the ordering*

the slopes on misclassified groups, a second-order effect that Assumption 1 leaves out and that may not be small as the simulations of Section 5 reflect. It works in the same direction as the bias, so the dominance in mean squared error and coverage is a statement about the first-order bias that the second-order dispersion reinforces rather than qualifies.

(18) *holds flow by flow and for the worst-case bias over the admissible probabilities, rather than for the signed sum. Proposition S.2 in Supplemental Appendix S.2 states the result in terms of bounds.*

Remark 4 (Heteroskedastic or correlated noise)**.** *If $v_i \sim N(0, \Sigma_v)$ and $u_i \sim N(0, \Sigma_u)$ with general covariance matrices, replacing each view by its whitened version, $\Sigma_v^{-1/2}(y_i - x_i\beta)$ and $\Sigma_u^{-1/2} z_i$, makes the stacked error spherical with unit variance, so Theorem 1 and Proposition 1 apply with $\delta^* = 1/2$, $a_{hg} = \Delta_{hg}' \Sigma_v^{-1} \Delta_{hg}$ and $b_{hg} = \nabla_{hg}' \Sigma_u^{-1} \nabla_{hg}$. With T fixed and N large, Σ_v and Σ_u are consistently estimable from within-group residuals, and the resulting rule is a generalized least squares k-means. Without whitening, a single scalar weight cannot be optimal for all pairs when the noise variance differs across directions, and the offsets deviate from one half in the way described for the one-view case in Section S.2.1 of González-Casasús (2026a).*

Remark 5 (Finer or coarser group structure in the proxy)**.** *The geometry also describes what happens when the proxy does not share the partition exactly. If the proxy is coarser, so that two model groups g and h share a center, then $b_{hg} = 0$ and the proxy view contributes nothing to the discrimination of that pair, $\mathrm{SNR}_{hg}(\delta) = \mathrm{SNR}_{hg}(1)$ for every δ, while it keeps contributing to every other pair. At the population level, a coarser proxy is therefore harmless and still useful. If the proxy is finer, so that a model group is split into subgroups with distinct centers, the two-view rule with G equal to the number of model groups estimates for that group the average of its subcenters, and the dispersion of the subcenters around the average inflates the within-group proxy variance, which lowers the weight of the proxy through $\hat{\sigma}_u^2$. The two-view rule then combines a valid, if noisier, proxy view with the residual view. Supplemental Appendix S.5 reports simulations for both cases, together with the case of a proxy without any group structure, which is the extreme of coarseness and the one in which the finite-sample behavior of k-means on the additional dimensions can hurt. A third possible departure is a proxy that is misaligned for a fraction of the units, which Proposition S.3 in Supplemental Appendix S.3 treats with the same geometry. A misaligned unit is pulled towards the wrong group by the proxy view and towards its own by the residual view, so the fraction of such units that the rule tolerates before losing its advantage over GFE grows with the weight on the residuals, from a few percent at the precision weight in the baseline design to one half, the point at which the proxy stops carrying group information, as the weight approaches one.*

3.4 The proxy as an included regressor

The covariates have been allowed throughout to be correlated with the groups. This correlation is what makes the group effects fixed-effects rather than random effects, and it is what makes classification error costly for β, since the coefficient $\varkappa$ in Corollary 1 is the covariance between the group contrasts of the covariates and those of the profiles. A regressor whose group means differ therefore carries information about the partition precisely in the situations in which misclassification is costly. It is natural to ask whether this information can be used in the classification step, as in the production function example in which productivity types are reflected in input choices (cf. Cheng et al., 2025), or in the setting of Chetverikov and Manresa (2022), in which the regressors have a factor structure aligned with the grouped effects. Using an included regressor to classify units has a consequence that an external proxy does not have. The regressor enters the outcome equation, so the units that the rule misclassifies are selected on their covariates and the covariate shift s_{hg} in (12) is nonzero. As noted in Section 3.1, such classifiers are covered by the bias formula but not by the inference procedures of the local misclassification framework, because the shift is not identified in general. In the two-view rule, the shift is pinned down by the geometry, and it works in favor of the estimator.

Consider the scalar case $K = 1$ and write the regressor as $x_i = \mu_{g_i} + e_i$, with $\mu_g \in \mathbb{R}^T$ the group-specific mean path and e_i a deviation without group structure, and take $z_i = x_i$, so that $\zeta_g = \mu_g$ and $u_i = e_i$. The decomposition is the grouped conditional mean structure that Corollary 1 already uses through μ_g and S, with the additional requirement, needed for the Gaussian geometry, that the deviations be homoskedastic and independent of the group. The exogeneity required for (1) is that e_i be independent of v_i.

Theorem 2 (Regressor-assisted classification)**.** *Let $G = 2$, $K = 1$, $z_i = x_i$ with $e_i \sim N(0, \sigma_u^2 I_T)$ independent of $v_i \sim N(0, \sigma_v^2 I_T)$ conditional on g_i, and let Assumption 1 hold for $\hat{g}(\delta)$ with the probabilities $p(\delta)$ of Theorem 1(a) and the local misclassification offsets of Theorem 1(c), $\bar{v}_{hg} = \bar{\rho}(\delta)(\alpha_h - \alpha_g)$ and $s_{hg} = \bar{\rho}^z(\delta)(\mu_h - \mu_g)$. Then*

$$\mathrm{Bias}(\hat{\beta}(\delta)) = (1 - \bar{\rho}(\delta) - \bar{\rho}^z(\delta))\, p(\delta)\, \varkappa, \qquad \varkappa = S^{-1} \sum_{t=1}^{T} (\mu_{2t} - \mu_{1t})(\alpha_{2t} - \alpha_{1t}), \tag{19}$$

and in particular $\mathrm{Bias}(\hat{\beta}(\delta^*)) = 0$, so that at the precision weight $\hat{\beta}(\delta^*)$ has the mean and the variance of the oracle estimator. For the single-view rules, $\mathrm{Bias}(\hat{\beta}(1)) = \frac{1}{2} p(1) \varkappa$ and $\mathrm{Bias}(\hat{\beta}(0)) = \frac{1}{2} p(0) \varkappa$.

The theorem has two parts. First, classifying on a regressor alone is not worse than residual-based classification at equal misclassification probabilities, contrary to what the comparison of proxy-based and residual-based classifiers would suggest for an external proxy. The selection of misclassified units on their covariates halves the compositional shift that drives the bias, exactly as the selection on outcome shocks halves the unabsorbed gap. Second, the two halves add up at the precision weight. A unit misclassified by the two-view rule is one whose outcome shock points towards the destination profile and whose regressor noise points towards the destination mean. At δ^*, each absorbs one half of its gap, and the attribution of the wrong profile is exactly compensated by the composition of the misclassified units in terms of the regressor. The first-order bias vanishes and the estimator has the mean and the variance of the oracle, hence its distribution along the local limit, without any debiasing and without knowledge of the misclassification probabilities. The result is first-order and holds in the local misclassification regime, and it is a statement within Assumption 1.[4] When $K > 1$ and only one regressor is used as a proxy, Proposition S.4 in Supplemental Appendix S.4 shows that the compositional channel of the proxy regressor is cancelled while the channels of the remaining regressors keep the factor $1 - \bar{\rho}(\delta^*) = 1/2$ of Corollary 1, net of the part of their contrast that the proxy regressor's noise explains. The two-view rule then removes the component of the bias generated by the regressor that carries the group signal and halves the rest. The whole first-order bias vanishes when the group-mean contrasts of the other regressors load on those of the proxy regressor with the same coefficient γ with which their deviations load on its noise, that is, when a common factor drives both the means and the deviations of the regressors with the same loadings. A shared group structure alone, with different loadings in the means and in the deviations, does not deliver this cancellation.

Remark 6 (Fixed versus random effects)**.** *When the proxy is an included regressor, the diagnostic of Section 2.4 acquires a second reading. Write $x_i = \mu_{g_i^x} + e_i$, where g_i^x is the partition of the conditional mean of the regressor, which need not coincide with the partition g_i of the outcome equation, and e_i is independent of (g_i, g_i^x). Then $\mathbb{E}[x_i \mid g_i = g] = \sum_k \mu_k \mathbb{P}(g_i^x = k \mid g_i = g)$, so the regressor is mean independent of the latent group if the conditional distribution of g_i^x given g_i does not depend on g_i, and only then unless the μ_k*

[4] The primitive conditions of Supplemental Appendix S.1 assume proxy noise independent of the covariates, which an included regressor violates, so the theorem rests on the high-level assumption in a way that Corollary 1 does not.

are affinely dependent. Mean independence is the condition under which the group effects are random effects and pooled least squares is consistent for β. Three cases follow. If $g_i^x = g_i$, which is the maintained assumption of this section, the group effects are fixed-effects whenever the μ_g differ, and the regressor is a valid proxy. If g_i^x is independent of g_i, the group effects are random effects, the omitted group effect is mean independent of the regressor, and classification error is harmless for β under any rule that does not select on the outcome shocks, although the regressor is not a valid proxy. In the intermediate case, in which the two partitions are dependent but distinct, the effects are fixed-effects and the regressor is an invalid proxy. The diagnostic tests the first case against the other two, so a non-rejection supports the fixed-effects specification together with the validity of the regressor as a proxy, whereas a rejection does not by itself establish random effects. Discriminating between the last two cases is a test of the independence of g_i^x and g_i, based on the cross-tabulation of the single-view partitions, where misclassification attenuates the dependence so that only rejections of independence are robust, or on a comparison of the GFE and the pooled least squares estimators in the spirit of Hausman (1978). A regressor without group structure is trivially mean independent of the group, and the diagnostic has nothing to test. The income and democracy application of Section 6.2 uses lagged income as a proxy in this way.

3.5 Comparison with the combination of parameter estimates

The bias offset of residual-based classification suggests a different way of exploiting a proxy. Instead of combining the two sources of information in the classification step, one may estimate β twice, once under the residual-based classifier and once under the proxy-based classifier, and combine the two estimates so that their first-order biases cancel. González-Casasús (2026*a*) shows that the combination

$$\tilde{\beta} = 2\hat{\beta}(1) - \hat{\beta}(0) \tag{20}$$

is first-order unbiased with the oracle variance when the two classifiers satisfy Assumption 1 with the same misclassification probabilities, the residual-based classifier has offset exactly one half, and the proxy-based classifier has offset zero and no covariate shift. The next proposition evaluates the combination in the Gaussian design of this section and compares it with the CAS estimator.

Proposition 2 (Combination of parameter estimates)**.** *Let $G = 2$ and let the conditions of Corollary 1 hold at $\delta = 1$ and $\delta = 0$. Then*

$$\mathrm{Bias}(\tilde{\beta}) = (p(1) - p(0))\,\varkappa, \tag{21}$$

so that $\tilde{\beta}$ is first-order unbiased if and only if $p(1) = p(0)$, and $\|\mathrm{Bias}(\tilde{\beta})\| > \|\mathrm{Bias}(\hat{\beta}(1))\|$ whenever $p(0) < p(1)/2$ or $p(0) > 3p(1)/2$. If instead the proxy is an included regressor as in Theorem 2, then $\mathrm{Bias}(\tilde{\beta}) = (p(1) - \frac{1}{2}p(0))\varkappa$, which equals the bias of the GFE estimator when $p(1) = p(0)$. In both cases, the variance of $\tilde{\beta}$ is V to the order of Assumption 1.

The proposition identifies three respects in which the CAS estimator improves on the combination. The first is that the combination removes the bias only when the two classifiers are equally accurate, a knife-edge condition that cannot be checked, because the probabilities are not identified, and that fails in general. Away from it, the combination extrapolates and its bias can exceed the bias it was meant to correct. The CAS estimator has no such condition. Its bias is smaller than that of the GFE estimator for every configuration of the probabilities, by the factor $p(\delta^*)/p(1) < 1$ of Corollary 1, which approaches one only when the proxy carries almost no signal. It is also less than one half of that of proxy-based classification. The second is that

the combination relies on the one-half rule as an exact identity. The offset of a k-means classifier at finite separation exceeds one half by the overshoot in (15), and the offset of a regressor-based proxy classifier is one half rather than zero, so in both cases the combination over- or under-corrects. The CAS estimator uses the offsets only as a description of its bias, not as an ingredient of its construction, and when the proxy is a regressor it attains the oracle distribution precisely because both offsets are one half. The third is variance. The two classifiers misclassify different units, so the two estimates are not perfectly correlated in finite samples and the combination, which has weights (2,-1), inflates the variance; see footnote 3.

4 The Oracle Case and the Stacked Representation

In Assumption 1, the oracle case is $p_{a|b} = 0$ for all pairs, in which $\hat{\beta}$ has the mean and the variance of the oracle whatever the classifier. It is the case that the large-(N,T) theory of Bonhomme and Manresa (2015) describes, in which the time dimension grows, the groups are strictly separated, and the probability that any unit is misclassified becomes negligible relative to sampling error. González-Casasús (2026*b*) shows that under weak separation it is reached only when the cross-section grows slowly relative to the time dimension. Nothing is left to gain once the partition is recovered, so in this case the proxy cannot improve on the oracle and the weight is irrelevant. What a proxy can do is to bring the classifier closer to the oracle case, by making the misclassification probabilities negligible in situations in which those of the residual-based classifier are not. This section makes the point through an exact representation of the CAS estimator as a GFE estimator of a longer panel, whose separation between groups is a weighted average of the separations of the two views.

4.1 The stacked representation

Fix $\delta \in (0,1]$ and define, for each unit, the stacked outcome, covariates and profiles

$$\tilde{y}_i = \begin{pmatrix} \sqrt{\delta}\, y_i \\ \sqrt{1-\delta}\, z_i \end{pmatrix} \in \mathbb{R}^{\tilde{T}}, \qquad \tilde{x}_i = \begin{pmatrix} \sqrt{\delta}\, x_i \\ 0_{L\times K} \end{pmatrix} \in \mathbb{R}^{\tilde{T}\times K}, \qquad \theta_g = \begin{pmatrix} \sqrt{\delta}\, \alpha_g \\ \sqrt{1-\delta}\, \zeta_g \end{pmatrix} \in \mathbb{R}^{\tilde{T}}, \tag{22}$$

with $\tilde{T} = T + L$, and the stacked error $\tilde{v}_i = (\sqrt{\delta}v_i', \sqrt{1-\delta}u_i')'$. Then $\tilde{y}_i = \tilde{x}_i\beta + \theta_{g_i} + \tilde{v}_i$ is a grouped panel data model with $\tilde{T}$ periods in which the last L periods have no covariates, and

$$Q_\delta(\beta, \alpha, \zeta, g) = \sum_{i=1}^{N} \|\tilde{y}_i - \tilde{x}_i\beta - \theta_{g_i}\|^2 \tag{23}$$

is the GFE objective of Bonhomme and Manresa (2015) for the stacked panel. The separation between the groups of the stacked panel is

$$\frac{1}{\tilde{T}}\|\theta_g - \theta_h\|^2 = \frac{\delta\|\alpha_g - \alpha_h\|^2 + (1-\delta)\|\zeta_g - \zeta_h\|^2}{T+L}, \tag{24}$$

a weighted average of the separations of the two views.

Proposition 3 (Stacked representation). *Fix $\delta \in (0,1]$. The following hold.*

(a) (Equivalence) *A covariate-assisted estimator with weight δ is the GFE estimator of the stacked panel $(\tilde{y}_i, \tilde{x}_i)_{i\in[N]}$ with $\tilde{T}$ periods, and the oracle estimator of the stacked panel, the least squares estimator*

on the true groups, is the oracle estimator of the original panel.

(b) (Separation) *For all $g \neq h$,*

$$\frac{1}{\tilde{T}} \|\theta_g - \theta_h\|^2 \geq \max\left\{ \frac{\delta T}{T+L} \cdot \frac{\|\alpha_g - \alpha_h\|^2}{T},\ \frac{(1-\delta)L}{T+L} \cdot \frac{\|\zeta_g - \zeta_h\|^2}{L} \right\},$$

so that the separation per period of the stacked panel is bounded below by a fraction of the separation per period of either view.

(c) (Misclassification) *Under Assumption 2, the stacked error at the precision weight is spherical with variance $\sigma^2 = \sigma_v^2\sigma_u^2/(\sigma_v^2 + \sigma_u^2)$, and the pairwise misclassification probability of Theorem 1(a) for the pair (g, h), which equals the misclassification probability when $G = 2$ and bounds it from above otherwise, satisfies*

$$\bar{p}_{h|g}(\delta^*) = \bar{\Phi}\left(\frac{\|\theta_g - \theta_h\|}{2\sigma}\right) \leq \min\left\{\bar{p}_{h|g}(1), \bar{p}_{h|g}(0)\right\},$$
$$\bar{p}_{h|g}(\delta^*) \leq \frac{1}{2}\exp\left(-\frac{\|\theta_g - \theta_h\|^2}{8\sigma^2}\right) = \frac{1}{2}\exp\left(-\frac{a_{hg}/\sigma_v^2 + b_{hg}/\sigma_u^2}{8}\right),$$

where θ_g is evaluated at δ^*.

Part (a) is the formal content of the stacked representation. Every property of the GFE estimator that depends on the structure of the panel rather than on the values of its periods holds for the CAS estimator with the stacked panel in place of the original one. The regularity conditions of Bonhomme and Manresa (2015) for the stacked panel reduce to those for the original panel. Their identification condition requires the covariates to have enough within-cell variation for every partition, and since the covariates vanish in the proxy block, the condition for the stacked panel is the condition for the original panel scaled by $\delta T/\tilde{T}$. Their dependence conditions concern the error series within each unit, and the stacked error is the concatenation of two independent series, so the conditions apply to each block separately. The proviso is that their large-T arguments require both blocks to grow, that is, L of the same order as T, which is the case of a proxy observed in every period and not that of a time-invariant proxy. The separation condition is the one that differs substantively, and parts (b) and (c) describe how.

Part (b) says that the separation per period of the stacked panel is at least a fraction of the separation per period of the better separated view, so that a strictly separated proxy keeps the stacked groups separated whatever the separation of the profiles in the outcome equation. The fraction attached to the proxy view is $(1-\delta)L/(T+L)$, which stays away from zero when L/T does, so the safeguard described next requires the proxy to be observed over a number of periods comparable to T, and it is lost for a time-invariant proxy when T grows. The weak separation concern of González-Casasús (2026*b*) is the situation in which $T^{-1}\|\alpha_g - \alpha_h\|^2$ is small. The misclassification probabilities of the residual-based classifier are then negligible only when the expected number of misclassified units, which grows with N and shrinks with the separation and with the length of the panel, is small, a condition that forces the cross-section to be small relative to the time dimension, the more so the weaker the separation. In the stacked panel, the separation is bounded below by a fraction of the proxy's, so the CAS classifier may be in the strictly separated case of Bonhomme and Manresa (2015) even when the residual-based classifier is in the weakly separated case,

and it can reach the oracle case under the conditions of Bonhomme and Manresa (2015) rather than under the restrictive conditions that weak separation imposes. This is the sense in which the proxy is a safeguard against the weak separation concern, and it requires $\delta < 1$. The profiles of the outcome equation may even coincide across groups without affecting the stacked separation. In that case, misclassification is harmless for β, since $\varkappa = 0$ in Corollary 1, and the proxy delivers the partition itself.

Part (c) expresses, in the Gaussian design, the familiar fact that the misclassification probability of a k-means classifier decreases exponentially in the number of periods times the separation per period, which is the mechanism behind the oracle property in Bonhomme and Manresa (2015). At the precision weight, the exponent is the sum of the exponents of the two views. A proxy observed in every period doubles the number of observations per unit that enter the classification step, from T to $2T$, while it adds nothing to the information about β, since the covariates vanish in the proxy block. When the two views have the same separation per period relative to their noise, the exponent doubles exactly as it would in a panel twice as long. The proxy lengthens the panel for the purpose of classification alone, which is where a short panel falls short. The first inequality in (c) adds that the two-view classifier is at least as accurate as the better of the two single-view classifiers, so that it reaches the oracle case whenever either view alone would.

Two further consequences of the representation are used below. First, it transfers the robustness of the GFE estimator to an overspecified number of groups. González-Casasús (2026*b*) shows that, under conditions that restrict the cross-sectional dependence of the errors and the interplay between error-fitting refinements of the true partition and the covariates, the GFE estimator with an overspecified number of groups is asymptotically equivalent to the estimator with the true number of groups, and by part (a) the result applies to the covariate-assisted estimator with a fixed weight through the stacked panel. The CAS estimator re-estimates the weight at each G, so the transfer is exact for a fixed weight and holds for the estimated one as long as the weight is stable across G, which it is in the applications. In practice, G can be chosen generously, and the stability of the estimates across G is a useful diagnostic, as the applications of Section 6 illustrate. Second, in the oracle case, the within-group variances of the two views are the error variances, so the estimated weight is the precision weight δ^*. The weight no longer matters for β once the partition is recovered, but it remains well defined, and it is only when classification is imperfect that it plays the role described in Section 3.

Remark 7 (Extensions)**.** *The two-view rule modifies only the classification step of the GFE algorithm, so it extends to the models to which the GFE approach has been extended, with the proxy equation unchanged, and the stacked representation carries over. Two cases are relevant for the applications of Section 6. With a lagged outcome among the regressors, the GFE estimator remains consistent under the conditions of Bonhomme and Manresa (2015) and part (a) applies verbatim. In the analysis of Section 3, the lagged outcome is a predetermined regressor whose covariate shift under residual-based classification is a known function of the offset and of the persistence parameter (González-Casasús, 2026a, Remark 1). With both unit-specific and grouped effects, the model in deviations from unit means has the structure of* (1) *and is estimated as in Bonhomme and Manresa (2015), with the residual view replaced by the demeaned residuals. Demeaning projects the errors on the within space, where they remain spherical, and the profiles in deviations lie in that space, so the geometry of Section 3 applies with $T - 1$ effective periods. When the proxy is demeaned as well, as in the application, both within-group variances carry the same factor $(T-1)/T$ and the precision weight is unchanged. A proxy with group-specific levels need not be demeaned, which is an advantage of the two-view formulation, because the demeaning of the outcome does not remove them.*

5 Monte Carlo Evidence

This section examines the finite-sample performance of the CAS estimator in short panels, with three objectives. It verifies the misclassification rates and offsets predicted by Theorem 1, it compares the bias, dispersion and coverage of the estimator with those of the GFE estimator, of proxy-based classification and of the combination of parameter estimates of Section 3.5, and it illustrates Theorem 2. Supplemental Appendix S.5 reports the behavior of the estimator under different forms of proxy misspecification. All designs use $N = 100$, $T = 7$, $K = 1$, $\beta = 1$, $G = 2$ and 2,000 replications unless stated otherwise. Every estimator is computed with Algorithm 1 from 100 random starts followed by 50 steps of a variable neighborhood search in the spirit of Bonhomme and Manresa (2015). Standard errors are the large-T clustered standard errors of Bonhomme and Manresa (2015) that treat the estimated groups as known, which are the default choice of practitioners. The combined estimator $\tilde{\beta} = 2\hat{\beta}(1) - \hat{\beta}(0)$ uses the standard error of the GFE estimator, as the first-order theory prescribes.

5.1 Baseline design: external proxy with unrelated levels

The outcome equation is (1) with profiles that diverge linearly over time, $\alpha_{1t} = -1.7\,t/T$ and $\alpha_{2t} = 1.7\,t/T$, scaled by a parameter $\xi \geq 0$ that controls the separation of the groups in the outcome equation, so that $y_{it} = x_{it}\beta + \xi\alpha_{g_it} + v_{it}$ with $v_{it} \sim N(0, 2^2)$. The regressor is correlated with the group profile, $x_{it} \sim N(\alpha_{g_it}, 1.5^2)$, so that the covariate contrast and the profile contrast are synchronized and the misclassification bias is positive. The proxy is a seven-dimensional vector, so $L = T$, whose centers are a level shift of an arbitrary common shape, $\zeta_1 = 0.3 + 0.2\sin(l)$ and $\zeta_2 = \zeta_1 + 1.085$ for $l \in [7]$, with $u_{il} \sim N(0, 1)$. Its levels and shape bear no relation to the profiles α_g. Memberships are drawn uniformly at random. The parameters are chosen so that at $\xi = 1$ the residual view and the proxy view have the same separation-to-noise ratio, $\mathrm{SNR}(1) = \mathrm{SNR}(0) = 2.87$, which corresponds to a misclassification probability of 7.5% per unit with known centers. The precision weight is $\delta^* = 1/5$, and the two-view rule has $\mathrm{SNR}(\delta^*) = 4.06$, which corresponds to 2.1% misclassification. The values $\xi \in \{0, 0.5, 1, 1.5\}$ span no separation, very weak separation (24% misclassification with known centers), the baseline, and strong separation (1.6%).

Table 1 reports the results and Figure 1 the sampling distributions of $\hat{\beta}$. Three patterns stand out. First, the misclassification rates follow the theory. The proxy and the two-view rules are within a percentage point of their known-center benchmarks at $\xi = 1$, whereas the GFE rule is further from its benchmark (10.7% against 7.5%) because, with $T = 7$, the estimation error of the centers is relatively larger for the view that also has to estimate β. The additivity of Theorem 1(b) is visible across ξ. As the outcome separation weakens, the GFE misclassification rate deteriorates rapidly (38% at $\xi = 0.5$) while the two-view rate degrades gracefully (6.5%), because the proxy view keeps contributing its share of the signal, and the estimated weight equals δ^* in every design. Second, the realized offsets follow Theorem 1(c) with a qualification. They are zero for the proxy rule and, at $\xi \geq 1$, above one half for the GFE and the two-view rules, but below the exact values (15), which are 0.66 and 0.59 in this design. The gap is due to the estimated centers. The same draws classified with the population centers and the true slope give offsets and misclassification rates within 0.01 of the theoretical values, whereas estimated profiles absorb part of the shocks of the units misclassified into their group, which pulls the realized offset towards one half. The first-order bias formula with the local offset of one half is accordingly closer to the realized bias than the formula with the exact offset, as Section 5.2 documents. Third, the bias of $\hat{\beta}$ follows from the two channels. At $\xi = 1$, the GFE estimator has a bias larger than its standard deviation and its nominal 95% interval covers

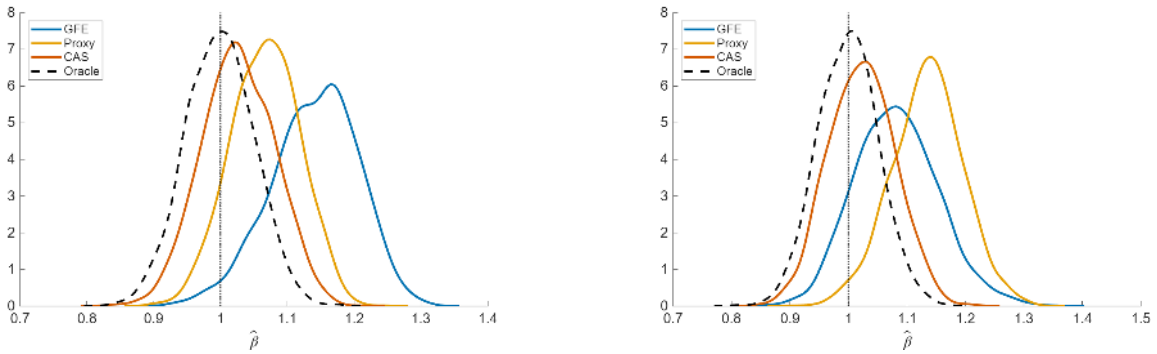
Table 1: Baseline design: external proxy with unrelated levels

	Misclassification			$\hat{\beta}$				
	MC	Theory	$\hat{\rho}$	Bias	SD	RMSE	Coverage	$\hat{\delta}$
				$\xi = 0$				
Oracle	–	–	–	-0.001	0.049	0.049	0.95	
GFE	0.460	0.500	–	0.001	0.043	0.043	0.92	
Proxy	0.083	0.076	–	0.000	0.046	0.046	0.95	
CAS	0.088	0.076	–	0.000	0.050	0.050	0.92	0.20
Combined	–	–	–	0.002	0.057	0.057	0.81	
				$\xi = 0.5$				
Oracle	–	–	–	-0.001	0.051	0.051	0.95	
GFE	0.377	0.236	0.42	0.141	0.065	0.155	0.15	
Proxy	0.083	0.076	-0.01	0.067	0.052	0.085	0.72	
CAS	0.065	0.054	0.55	0.026	0.054	0.060	0.86	0.20
Combined	–	–	–	0.214	0.105	0.238	0.10	
				$\xi = 1$				
Oracle	–	–	–	-0.001	0.051	0.051	0.94	
GFE	0.107	0.075	0.58	0.083	0.071	0.109	0.54	
Proxy	0.083	0.076	0.00	0.135	0.061	0.148	0.37	
CAS	0.026	0.021	0.55	0.023	0.056	0.061	0.89	0.20
Combined	–	–	–	0.030	0.117	0.121	0.56	
				$\xi = 1.5$				
Oracle	–	–	–	0.000	0.050	0.050	0.94	
GFE	0.019	0.016	0.56	0.027	0.059	0.064	0.88	
Proxy	0.082	0.076	0.00	0.201	0.074	0.215	0.20	
CAS	0.006	0.005	0.54	0.009	0.053	0.054	0.93	0.20
Combined	–	–	–	-0.148	0.102	0.180	0.29	

Note: $N = 100$, $T = 7$, $G = 2$, 2,000 replications. "Theory" is the misclassification probability $\bar{\Phi}(\mathrm{SNR}(\delta)/2)$ of Theorem 1(a) with population centers, at $\delta = 1$ (GFE), $\delta = 0$ (Proxy) and $\delta = \delta^* = 0.2$ (CAS). $\hat{\rho}$ is the realized offset, the average over misclassified units of $\Delta' v_i / \|\Delta\|^2$ with Δ the gap of the outcome profiles in the direction of the destination group. With population centers and the true slope, the realized offsets on the same draws are 0.91, 0.65 and 0.58 for the residual-based rule and 0.63, 0.58 and 0.57 for the precision-weighted rule at $\xi = 0.5$, 1 and 1.5, and the misclassification rates equal the "Theory" column to two digits. "Combined" is $2\hat{\beta}_{\mathrm{GFE}} - \hat{\beta}_{\mathrm{Proxy}}$. Coverage refers to the nominal 95% interval based on the large-T clustered standard errors, computed for the combined estimator with the standard error of the GFE estimator. $\hat{\delta}$ is the average QML weight.

in about half of the replications, the proxy rule has a larger bias and lower coverage despite misclassifying fewer units, because it absorbs none of the gap, and the CAS estimator has a bias of 0.023, a root mean squared error about half that of GFE, and coverage of 89%. The dispersion of the estimators also depends on the classifier in finite samples, and is smallest for the rule that misclassifies least. At $\xi = 0$, the groups do not differ in the outcome equation, misclassification is harmless and all estimators are unbiased, as the bias formula predicts.

Figure 1 shows the same facts as sampling distributions. At $\xi = 1$, the CAS density sits almost on top of the oracle's, whereas the GFE density is shifted by more than one standard deviation and the proxy density further still. The proxy density is narrower than the GFE density but centered farther from the truth, which is the trade-off between the rate and the offset of Corollary 1 made visible. At $\xi = 0.5$, the GFE density moves far to the right and develops a shoulder, because the partition is recovered in some replications and



(a) Very weak separation: $\xi = 0.5$ (b) Baseline: $\xi = 1$

Figure 1: Sampling distributions of $\hat{\beta}$ in the baseline design

Note: Kernel density estimates over 2,000 replications of the GFE estimator (blue), the proxy-based estimator (orange), the CAS estimator (vermillion) and the oracle that uses the true groups (black, dashed). The vertical line is the true value $\beta = 1$.

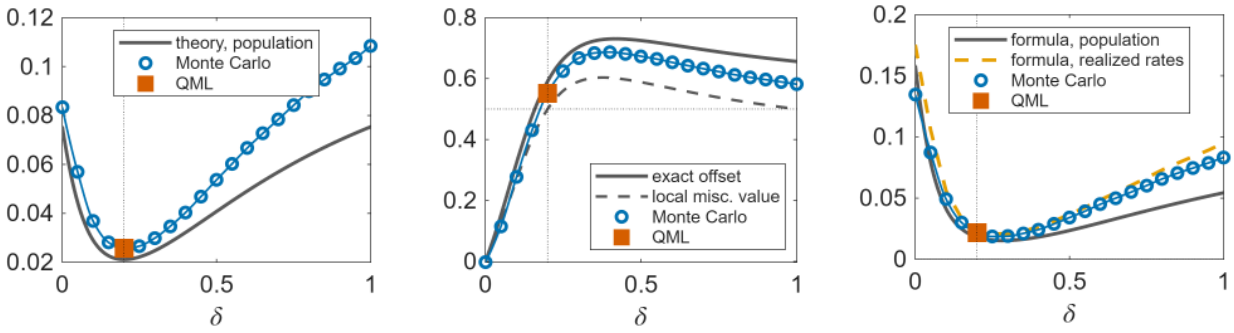
largely lost in others, while the CAS density stays close to the oracle's. Thus, the two-view rule inherits the stability of the proxy view and the offset of the residual view. The combined estimator, in contrast, behaves as Proposition 2 predicts. Its bias is small at $\xi = 1$, where the two single-view rules happen to have similar rates, and it is 0.214 at $\xi = 0.5$ and -0.148 at $\xi = 1.5$, where the GFE rule is respectively much less and much more accurate than the proxy rule, so that the extrapolation overshoots in one direction or the other, with a standard deviation about twice that of the CAS estimator and coverage far below the nominal level.

5.2 The role of the weight

Figure 2 isolates the role of the weight in the baseline design with $\xi = 1$ by computing the CAS estimator for a grid of fixed weights $\delta \in \{0, 0.05, \ldots, 1\}$, together with the QML estimator, and overlays the predictions of Theorem 1 and Corollary 1 with population centers. The misclassification rate is U-shaped with its minimum at the precision weight, where the QML estimate lands, and both endpoints are substantially worse than the interior minimum. The realized offset has the shape of the exact formula (15) but lies below it, between the local value of one half and the exact value, for the reason given in Section 5.1. The bias combines the two channels. It falls steeply from 0.13 at $\delta = 0$ to 0.021 at δ^*, is nearly flat between δ^* and 0.3, where the bias-minimizing weight δ_b lies (0.26 with the local offsets and 0.28 with the exact ones), and increases to 0.083 at $\delta = 1$, and the root mean squared error at δ^* is within 0.002 of its minimum over the grid. The first-order formula (17) with population quantities and the exact offsets reproduces the shape of the bias profile and its level near δ^*, and underestimates the bias for large weights, where the estimated-center rule misclassifies more than the known-center rule and the exact offset overstates the realized one. Evaluated at the realized rates and offsets (dashed line), the formula is within about fifteen percent of the realized bias for every $\delta \geq 0.1$, so the bias formula of the local misclassification framework describes the two-view rule as well as the one-view rules.

5.3 Further designs

Supplemental Appendix S.5 reports five further sets of designs, whose messages are the following. First, when the regressor itself is the proxy, $z_i = x_i$ (Table S.1), both realized offsets of the two-view rule are close to one half and, as Theorem 2 predicts, the bias disappears. The estimator has the bias, dispersion and



(a) Misclassification rate (b) Offset $\hat{\rho}(\delta)$ (c) Bias of $\hat{\beta}$

Figure 2: Misclassification, offset and bias as functions of the weight δ

Note: Baseline design with $\xi = 1$, 2,000 replications. Blue circles are Monte Carlo averages for fixed weights, joined by a shape-preserving interpolant, and the filled square is the QML estimator (average weight 0.20). Gray solid lines are the predictions of Theorem 1 and Corollary 1 with population centers. The gray dashed line in panel (b) is the local misclassification value $\bar{\rho}(\delta)$, and the orange dashed line in panel (c) is the first-order bias formula (17) evaluated at the realized misclassification rates and offsets. The vertical dotted line marks $\delta^* = 0.2$.

coverage of the oracle to within a few thousandths at every ξ, and the cancellation holds beyond the local misclassification regime, with offsets that exceed one half by the same overshoot. The combined estimator is badly biased at $\xi \leq 1$ in this design, which is the second case of Proposition 2, because the regressor-based classifier has an offset of one half rather than zero and the combination removes the wrong amount of bias.

Second, departures of the proxy from the shared partition behave as Remark 5 and Proposition S.3 describe (Tables S.2 and S.3). A coarser proxy, with one center for two model groups, is useful for the pairs it separates and harmless for the pair it does not, and a finer proxy, with two centers per model group, costs some accuracy but leaves the CAS estimator far better than GFE whether the model is estimated with the model groups or with the proxy groups. A proxy misaligned for a fraction π of the units leaves the CAS estimator preferable to GFE at $\pi = 0.05$ and worse from $\pi = 0.15$ on, since the precision weight trusts the proxy in proportion to its precision and not to its validity, in line with the tolerance that Proposition S.3 computes for the precision weight. A proxy without group structure raises the misclassification rate of the two-view rule slightly above that of GFE, because the k-means solution fits noise in the additional dimensions, which the estimated weight does not detect since it measures noise and not signal. Estimating the model with three groups instead of two affects the CAS estimator little, whereas the GFE estimator deteriorates, in line with the stability across G that the proxy affords in the income and democracy application below.

Third, the diagnostic of Section 2.4 has the right size in the baseline regime, with rejection rates of 5.1% and 10.5% at the 5% and 10% levels under a valid proxy, and power of 45%, 93% and 99% against misalignments of 5%, 15% and 30% of the units, so it detects with high probability the misalignments that are harmful (Table S.4). It rejects a proxy without group structure in 18% of the replications, a false rejection of the shared partition but a correct signal that the proxy carries no information about the groups, since the statistic then comes almost entirely from the proxy term, and it over-rejects mildly, 7.1% at the 5% level and 16.8% at the 10% level, when the GFE rule misclassifies more than a third of the units.

Fourth, the bias-aware intervals of González-Casasús (2026*a*) with the calibrated bounds of Remark 2 (Table S.5) are closer to the nominal level and shorter under CAS than under GFE, by a third in length at $\xi = 1$,

because the calibrated bounds are smaller under the two-view rule. The recommended safety factor for calibrated bounds and inflation mechanism for plug-in sensitivities in González-Casasús (2026*a*) enhances the coverage. Both corrections are local devices that work when misclassification is moderate, and the two-view rule keeps the classifier in that regime.

Fifth, departures from the Gaussian design (Table S.6) leave the ordering of the three rules unchanged. With $t(3)$ errors in both views, the two-view rule misclassifies more units than with Gaussian errors but still half as many as GFE, and the estimated weight is unchanged because the quasi-likelihood uses only second moments. With heteroskedastic and serially correlated errors, all three Euclidean classifiers deteriorate, since a single scalar weight cannot be optimal for every direction of the error (Remark 4), and whitening each view by its within-group covariance matrix, estimated from a first CAS fit, restores the geometry of Theorem 1 with $\delta^* = 1/2$, brings the coverage of the standard CAS interval closer to the nominal, and lowers the root mean squared error below that of the least squares oracle of the Gaussian design, because the whitened estimator is a generalized least squares estimator of β. The whitened rule is therefore the recommended version when the within-group residuals of a first fit are not approximately spherical.

6 Empirical Applications

Two applications illustrate the estimator in settings of different nature. The first is the employment effect of the minimum wage across United States counties, with eight years and about two thousand counties, in which the proxy is the employment trajectory of a sector that the minimum wage does not reach and the regressor of interest is state legislation, so that its covariate shift is zero by construction. It shows the inferential side of the framework, with a residual-based partition that isolates outlying counties as the number of groups grows and bias-aware intervals that remain informative under the two-view rule. The second is the relationship between income and democracy across countries, the leading grouped panel data application, with seven periods and ninety countries, in which the proxy, the savings rate, is linked to political development through cultural traits and disagrees with the democracy residuals about the partition. It shows what the two-view rule changes when the proxy is informative and what the diagnostic and alternative proxies can and cannot settle. In both cases, the proxy is a readily available series from the same data source, which is the situation that motivates the paper. Each subsection presents the setting, the proxy, the main estimates and the checks that Supplemental Appendices S.6 and S.7 report in full.

6.1 Minimum wage and county employment

The first application concerns the employment effect of the minimum wage, estimated from county-level employment in restaurants, the sector most exposed to it. The debate between Dube, Lester and Reich (2010) and Neumark, Salas and Wascher (2014) turns on how to control for heterogeneous local employment trends, which two-way fixed-effects ignore and county-specific trends absorb at the cost of the identifying variation. Grouped time effects are the natural middle ground, since local labor markets follow a small number of regional trajectories, and the specification is

$$\log emp_{it} = \beta_1 \log mw_{it} + \beta_2 \log pop_{it} + \alpha_{g_i t} + \eta_i + v_{it}, \tag{25}$$

with county effects η_i and grouped time effects, estimated on county-demeaned data as in Remark 7. Employment is the annual average employment in food services and drinking places (NAICS 722) from the

Table 2: Minimum wage and restaurant employment: estimates by number of groups

						Bias-aware 90% interval	
G	GFE	CAS	$\hat{\delta}$	Diff.	p-value	GFE	CAS
1	-0.032						
	(0.049)						
2	-0.095	-0.036	0.82	0.14	0.990	[-0.143, -0.047]	[-0.078, 0.006]
	(0.021)	(0.025)					
3	-0.050	-0.050	0.80	0.40	0.120	[-0.080, -0.020]	[-0.092, -0.009]
	(0.013)	(0.023)					
4	-0.064	-0.068	0.76	0.36	0.195	[-0.132, 0.005]	[-0.120, -0.016]
	(0.010)	(0.025)					
5	-0.020	-0.043	0.79	0.44	0.085	[-0.485, 0.444]	[-0.078, -0.007]
	(0.011)	(0.015)					
6	-0.050	-0.032	0.77	0.49	0.030	[-0.678, 0.578]	[-0.079, 0.015]
	(0.006)	(0.018)					
7	-0.064	-0.021	0.75	0.55	0.075	[-0.718, 0.590]	[-0.067, 0.024]
	(0.006)	(0.017)					
8	-0.073	-0.045	0.76	0.53	0.040	[-0.643, 0.496]	[-0.142, 0.053]
	(0.007)	(0.012)					
9	-0.069	-0.032	0.74	0.57	0.120	[-0.625, 0.486]	[-0.213, 0.149]
	(0.007)	(0.014)					
10	-0.073	-0.080	0.73	0.62	0.035	[-0.704, 0.559]	[-0.256, 0.095]
	(0.008)	(0.019)					

Note: Elasticity $\hat{\beta}_1$ of county restaurant employment with respect to the minimum wage, $N = 1{,}976$ counties, $T = 8$ (2012–2019), with county effects and grouped time effects; $G = 1$ is the two-way fixed-effects estimator. Standard errors clustered by state in parentheses. "Diff." is the share of counties classified differently by CAS and GFE. The p-value is that of the validity diagnostic with $B = 199$ residual bootstrap draws. The bias-aware intervals are the symmetric 90% intervals of González-Casasús (2026*a*) with calibrated bounds and safety factor two. All estimators use 200 random starts and 200 variable neighborhood search steps.

Quarterly Census of Employment and Wages, the minimum wage is the annual average of the binding state or federal minimum from Vaghul and Zipperer (2016), and population is the Census county estimate. The panel covers 2012 to 2019, $T = 8$, and the $N = 1{,}976$ counties with disclosed employment in every year, in the fifty states and the District of Columbia. Over the window, 27 states, covering 40% of the counties, raised their nominal minimum wage. Standard errors are clustered by state, the level at which the regressor varies.

The proxy is the county's log employment in manufacturing (NAICS 31–33), from the same source, demeaned in the same way. The grouped component of manufacturing employment is taken to be the regional demand trajectory that generates the groups, which is the premise of the debate about local trends, while its county-specific deviations, plant openings and closures, are specific to the sector. A county-specific demand shock common to both sectors would violate the independence of the two noises, and this is the assumption that the diagnostic examines. Manufacturing wages lie far above the minimum for almost all workers in the sector, so the minimum wage has no direct bearing on it, and state legislation is independent of plant-level shocks, so the covariate shift of Section 3.1 is zero for the minimum wage, and it is small for the population control, whose within-group deviations have a correlation of about 0.1 with the manufacturing residuals.

Table 2 reports the elasticity for $G = 1$ to 10. Every specification gives a small negative elasticity, within the range of the literature, and what the two-view rule changes is the credibility of the estimate rather than its sign. The GFE estimates move with G in a way that has no economic interpretation, and their

state-clustered standard errors shrink to 0.006–0.008 for $G \geq 6$, which reflects the residual variance that a partition of two thousand counties on eight residuals can fit rather than precision. From $G = 5$ on, the residual-based rule isolates small groups of counties with outlying employment trajectories, four counties with about a hundred restaurant jobs or fewer at $G = 5$ and groups of two to four counties at every larger G. Their profile contrast is an order of magnitude larger than that of the other groups and correlated with their minimum-wage history, so every classification flow into or out of such a group carries a large sensitivity κ_{ab} in (12), and the GFE bias-aware intervals explode, $[-0.49, 0.44]$ at $G = 5$, although the calibrated misclassification bounds of the GFE partition are no larger than those of the two-view rule. The two-view rule reclassifies 14% of the counties at $G = 2$ and 36% to 62% from $G = 4$ on, and it forms no such group before $G = 7$, because the manufacturing trajectories of the outlying counties are unremarkable. Its estimates move over a narrower range, with standard errors of 0.012 to 0.025 that reflect the uncertainty of the partition, and its bias-aware intervals exclude zero for $G = 3$ to 5 and remain informative up to $G = 7$, after which small groups appear in this partition too. The estimated weight of 0.73 to 0.82 says that the manufacturing view is the noisier one per unit of variance, and the diagnostic does not reject the shared partition at the 5% level for $G \leq 5$. The economic conclusion, an elasticity of about -0.05 with an interval that excludes zero at $G = 3$ to 5 but not at larger G, is then a statement whose uncertainty accounts for classification error.

Three points qualify this reading, and Supplemental Appendix S.6 documents them. First, the information criteria of Table S.10 do not single out a group structure, since with two thousand counties the penalty is small relative to the reduction of the residual variance that an additional group buys on eight periods, so the relevant comparison is across the range of G rather than at one value. Second, at $G = 2$, each bias-aware interval excludes the other estimator's point estimate although the partitions differ for 14% of the counties only, because each bias set is local to its own partition, a point developed for the other application below. Third, the point estimates suggest that the effect is concentrated in the smaller counties, where the minimum wage is more likely to bind, with two-way fixed-effects elasticities of -0.004 in the 1,246 larger counties of the sample and -0.096 in the 730 smaller ones, a difference within their standard errors, and of -0.023 on the 2,158 counties with a complete restaurant series.

6.2 Income and democracy

The second application revisits the relationship between a country's income and its level of democracy, a leading example of an empirical conclusion that depends on how unobserved heterogeneity is modeled. The positive cross-country association between income per capita and democracy disappears once country fixed-effects are included (Acemoglu, Johnson, Robinson and Yared, 2008). Bonhomme and Manresa (2015) argue that the historical events that shape political and economic development are shared by groups of countries rather than country-specific, and estimate the grouped specification

$$democracy_{it} = \beta_1 democracy_{i,t-1} + \beta_2 income_{i,t-1} + \alpha_{g_i t} + v_{it}, \tag{26}$$

on the balanced five-year panel of Acemoglu et al. (2008), 1970–2000, with $N = 90$ countries and $T = 7$ periods. The specification has been extensively revisited from the grouped panel data perspective (e.g., Kim and Wang, 2019, Zhang, 2022, Aguilar and Boot, 2022, Mugnier, 2022, Lewis et al., 2023, Rivero, 2023), and the analyses in González-Casasús (2026*b*) and González-Casasús (2026*a*) document that, with seven periods, the estimated groups are far from stable and the standard intervals are unreliable. Classification is

Table 3: Income and democracy: estimates by number of groups

	Lagged democracy $\hat{\beta}_1$			Income $\hat{\beta}_2$			Cumulative $\hat{\beta}_2/(1-\hat{\beta}_1)$				
G	GFE	Proxy	CAS	GFE	Proxy	CAS	GFE	Proxy	CAS	$\hat{\delta}$	Diff.
2	0.601	0.663	0.661	0.061	0.086	0.090	0.152	0.256	0.265	0.15	0.36
	(0.041)	(0.048)	(0.049)	(0.011)	(0.015)	(0.016)	(0.021)	(0.025)	(0.027)		
3	0.406	0.639	0.640	0.089	0.098	0.098	0.151	0.272	0.272	0.12	0.56
	(0.051)	(0.048)	(0.048)	(0.011)	(0.014)	(0.014)	(0.013)	(0.019)	(0.019)		
4	0.302	0.632	0.635	0.082	0.100	0.102	0.118	0.271	0.279	0.11	0.59
	(0.053)	(0.049)	(0.044)	(0.009)	(0.015)	(0.013)	(0.011)	(0.024)	(0.021)		
5	0.255	0.628	0.636	0.079	0.103	0.103	0.106	0.278	0.284	0.11	0.62
	(0.049)	(0.050)	(0.044)	(0.009)	(0.016)	(0.013)	(0.009)	(0.023)	(0.022)		
6	0.465	0.620	0.631	0.064	0.109	0.107	0.119	0.287	0.290	0.10	0.64
	(0.042)	(0.047)	(0.045)	(0.007)	(0.014)	(0.013)	(0.011)	(0.022)	(0.021)		
7	0.403	0.629	0.648	0.065	0.101	0.101	0.108	0.272	0.287	0.11	0.63
	(0.042)	(0.048)	(0.043)	(0.008)	(0.015)	(0.013)	(0.010)	(0.024)	(0.023)		
8	0.333	0.627	0.630	0.070	0.102	0.096	0.104	0.275	0.258	0.11	0.59
	(0.042)	(0.049)	(0.047)	(0.008)	(0.015)	(0.012)	(0.010)	(0.024)	(0.024)		
9	0.335	0.605	0.647	0.070	0.101	0.096	0.105	0.255	0.271	0.10	0.66
	(0.042)	(0.051)	(0.045)	(0.007)	(0.015)	(0.012)	(0.009)	(0.023)	(0.024)		
10	0.348	0.628	0.700	0.068	0.101	0.088	0.104	0.272	0.293	0.11	0.68
	(0.043)	(0.046)	(0.038)	(0.007)	(0.014)	(0.011)	(0.009)	(0.025)	(0.025)		

Note: Balanced panel of Acemoglu et al. (2008), $N = 90$, $T = 7$. GFE is the estimator of Bonhomme and Manresa (2015) ($\delta = 1$), Proxy classifies on the savings rate only ($\delta = 0$), and CAS is the CAS estimator with the QML weight $\hat{\delta}$. All estimators are computed from 1,000 random starts followed by 500 steps of the variable neighborhood search of Bonhomme and Manresa (2015). Large-T clustered standard errors in parentheses. "Diff." is the share of countries classified differently by CAS and GFE, after relabeling.

thus the bottleneck of this application, which makes it a natural testing ground for auxiliary information.

The proxy is the national savings rate, available in the same data set for the same periods.[5] The argument for a shared group structure runs through cultural traits. The political culture school holds that values and beliefs shape political institutions (Inglehart and Welzel, 2018), and savings behavior is strongly influenced by cultural traits (Carroll, Rhee and Rhee, 1999, Costa-Font, Giuliano and Ozcan, 2018), so countries that follow similar paths of political development are expected to display similar savings patterns. Nothing is assumed about the levels or the time paths of savings relative to the democracy profiles, since the two-view rule only uses the partition. One caveat applies from the outset. In Acemoglu et al. (2008), the savings rate is one of the instruments for income, so it is arguably tied to the regressor of interest, and a partition aligned with income classes absorbs part of the income variation, as the estimates with lagged income itself as the proxy show below.

Table 3 reports the estimates of the persistence of democracy β_1, of the income coefficient β_2 and of the implied cumulative income effect $\beta_2/(1-\beta_1)$ for $G = 2, \ldots, 10$, for the GFE estimator, for classification on the savings rate alone, and for the CAS estimator. The GFE estimates reproduce those of Bonhomme and Manresa (2015) for every G and vary widely with the number of groups, with a persistence estimate between 0.25 and 0.60, which is what one expects when the classification is driven by seven noisy residuals per country and the symptom that the stability analysis of González-Casasús (2026*b*) recommends examining. The CAS estimates are stable, with a persistence of 0.63 to 0.66 for $G \leq 9$, an income coefficient of 0.09 to 0.11 and a cumulative effect of 0.26 to 0.29, more than twice the GFE value and of the same

[5] Five missing observations are imputed with the sample mean, and the partition is unchanged under a within-country imputation.

magnitude as the pooled estimate of Acemoglu et al. (2008); only at $G = 10$, with nine countries per group, does the two-view partition start to fit noise. The estimated weight is 0.10 to 0.15, because the within-group variance of the savings rate is about eight times smaller than that of the democracy residuals, and classification on the savings rate alone gives nearly the same estimates. The two-view rule reallocates 7% of the countries relative to the proxy-only partition at $G = 4$, against 59% relative to the GFE partition, so in this application it is, for practical purposes, classification on the savings rate. What it adds is the weight, which makes the dominance of the proxy a data-driven fact rather than a choice, together with the diagnostic and the sensitivity analysis below, and not the bias reduction of Corollary 1, which presumes a residual view with a comparable weight. The information criteria of Table S.10 point in the same direction: the BIC criterion of Bonhomme and Manresa (2015) decreases over the whole range for the GFE fit, whereas its two-view analogue is minimized at four or five groups, which is the range on which the detailed results are reported.

Two points qualify this reading. First, stability across G is what one expects when the classification is accurate, since additional groups then split true groups without changing the estimates, but it is also what a precisely measured proxy produces whether or not it is valid, so it is a necessary rather than a sufficient sign, and the larger residual variance of the outcome equation under the CAS partition (0.034 against 0.023 at $G = 4$) is a reminder that the GFE partition minimizes it by construction, in part by fitting idiosyncratic shocks. Second, the sensitivity analysis over the weight (Figure S.2) shows a sharp transition rather than a gradual one, with the partition and the estimates switching from the proxy regime to the GFE regime between $\delta = 0.2$ and $\delta = 0.5$ and confidence bands that do not overlap across regimes, which makes the disagreement between the views transparent.

The disagreement raises the question of Section 2.4, namely whether the savings rate and the democracy residuals share one partition at all, and Supplemental Appendix S.7 reports the checks. The diagnostic, with the within-group residuals resampled by country, does not reject the shared partition at the 5% level for $G \leq 5$, is at the margin at $G = 6$ and rejects it from $G = 7$ on, where both single-view partitions fit noise, but it is sensitive to the bootstrap design in a panel of seven periods and its power is limited at this sample size. Two alternative proxies from the same data set, the median age of the population, a marker of the demographic transition (Cincotta, 2009, Weber, 2013, Dyson, 2013), and lagged income itself, an included regressor in the sense of Section 3.4, give the same persistence estimate and smaller income coefficients, so that the cumulative effect ranges from 0.04 to 0.28 across proxies, and the diagnostic rejects none of the three at $G = 4$ although their partitions differ pairwise for about half of the countries. The persistence of democracy is thus the finding that survives the choice of proxy, and the size of the income effect is not identified by this sample without a stand on which proxy is valid. The bias-aware intervals of González-Casasús (2026*a*) reinforce the first conclusion, since the CAS intervals for β_1, which are close to the standard ones because the calibrated outgoing probabilities of the two-view rule are small, lie above the GFE intervals for every G except 2 and 6, with the caveat that the calibration presumes the validity of the proxy. The whitened rule of Remark 4 gives the same partition at $G = 4$ up to 3% of the countries.

7 Conclusion

This paper studies how auxiliary covariates that share the latent group structure of a grouped panel data model can be used to improve its estimation. I propose the CAS estimator, a weighted two-view k-means

rule with the weight determined by a Gaussian quasi-likelihood. At the precision weight, the squared separation-to-noise ratios of the two views add up, so fewer units are misclassified, while the endogenous bias offset of residual-based classification is preserved. The estimator therefore has a strictly smaller first-order misclassification bias than the GFE estimator, less than half that of proxy-based classification, and when the proxy is an included regressor the first-order bias vanishes. In the oracle case, the estimator is the GFE estimator of a stacked panel in which the proxy lengthens the time dimension for the purpose of classification, so that the separation condition under which the partition is recovered is weaker than for the GFE estimator.

The main limitation is that the validity of the proxy, the coincidence of its partition with that of the outcome equation, is a maintained assumption that the estimated weight does not test, since the weight measures precision and not validity. The results show that a modest misalignment is tolerated and a large one is not. The quasi-likelihood ratio diagnostic exploits the overidentifying nature of the shared partition to detect misalignment, with a bootstrap calibration that the simulations support, and the bias-aware intervals of González-Casasús (2026*a*) report the uncertainty that remains. The two applications show the trade-offs involved. Two directions seem worth pursuing. One is a distribution theory for the diagnostic that dispenses with the bootstrap, which requires the joint behavior of three k-means partitions of the same data. The other is the choice among several proxies, which the quasi-likelihood handles with one weight per proxy block and whose joint validity the diagnostic can assess block by block, a question that the income and democracy application makes concrete.

References

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

Aguilar, J. and Boot, T. (2022), 'Grouped heterogeneity in linear panel data models with heterogeneous error variances', *Available at SSRN 4031841* .

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

Binkiewicz, N., Vogelstein, J. T. and Rohe, K. (2017), 'Covariate-assisted spectral clustering', *Biometrika* **104**(2), 361–377.

Bonhomme, S. and Denis, A. (2025), Fixed effects and beyond: Bias reduction, groups, shrinkage and factors in panel data, Working Paper 2526, Banco de España.

Bonhomme, S., Lamadon, T. and Manresa, E. (2022), 'Discretizing unobserved heterogeneity', *Econometrica* **90**(2), 625–643.

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

Carroll, C. D., Rhee, B.-K. and Rhee, C. (1999), 'Does cultural origin affect saving behavior? evidence from immigrants', *Economic Development and Cultural Change* **48**(1), 33–50.

Cheng, X., Schorfheide, F. and Shao, P. (2025), Clustering for multi-dimensional heterogeneity with an application to production function estimation, PIER Working Paper 25-014, Penn Institute for Economic Research, University of Pennsylvania.

Chetverikov, D. and Manresa, E. (2022), 'Spectral and post-spectral estimators for grouped panel data models', *arXiv preprint arXiv:2212.13324* .

Cincotta, R. P. (2009), 'Half a chance: Youth bulges and transitions to liberal democracy', *Environmental Change and Security Program Report* **13**, 10–18.

Costa-Font, J., Giuliano, P. and Ozcan, B. (2018), 'The cultural origin of saving behavior', *PloS one* **13**(9), e0202290.

Dube, A., Lester, T. W. and Reich, M. (2010), 'Minimum wage effects across state borders: Estimates using contiguous counties', *Review of Economics and Statistics* **92**(4), 945–964.

Dyson, T. (2013), 'On demographic and democratic transitions', *Population and Development Review* **38**, 83–102.

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. (2026*a*), Beyond the oracle: Inference with estimated groups, Working Paper 1601, Barcelona School of Economics.

González-Casasús, O. (2026*b*), Grouped fixed-effects estimation is more robust than you think, Working Paper 1599, Barcelona School of Economics.

Green, E. J. and Strawderman, W. E. (1991), 'A james-stein type estimator for combining unbiased and possibly biased estimators', *Journal of the American Statistical Association* **86**(416), 1001–1006.

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.

Hausman, J. A. (1978), 'Specification tests in econometrics', *Econometrica* **46**(6), 1251–1271.

Huang, J. (2023), 'Quasi-bayesian inference for grouped panels', *Working Paper* .

Inglehart, R. and Welzel, C. (2018), Political culture and democracy, *in* 'New directions in comparative politics', Routledge, pp. 141–164.

Judge, G. G. and Mittelhammer, R. C. (2004), 'A semiparametric basis for combining estimation problems under quadratic loss', *Journal of the American Statistical Association* **99**(466), 479–487.

Kagan, A. M., Linnik, Y. V. and Rao, C. R. (1973), *Characterization Problems in Mathematical Statistics*, Wiley, New York.

Kim, J. and Wang, L. (2019), 'Hidden group patterns in democracy developments: Bayesian inference for grouped heterogeneity', *Journal of Applied Econometrics* **34**(6), 1016–1028.

Kitamura, Y. and Laage, L. (2024), 'Estimating stochastic block models in the presence of covariates', *arXiv preprint arXiv:2402.16322* .

Kline, P. and Walters, C. R. (2016), 'Evaluating public programs with close substitutes: The case of head start', *The Quarterly Journal of Economics* **131**(4), 1795–1848.

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

Liu, L. (2023), 'Density forecasts in panel data models: A semiparametric bayesian perspective', *Journal of Business & Economic Statistics* **41**(2), 349–363.

Löffler, M., Zhang, A. Y. and Zhou, H. H. (2021), 'Optimality of spectral clustering in the gaussian mixture model', *The Annals of Statistics* **49**(5), 2506–2530.

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.

Mendolia, S., Stavrunova, O. and Yerokhin, O. (2021), 'Determinants of the community mobility during the covid-19 epidemic: The role of government regulations and information', *Journal of Economic Behavior & Organization* **184**, 199–231.

Mittelhammer, R. C. and Judge, G. G. (2005), 'Combining estimators to improve structural model estimation and inference under quadratic loss', *Journal of econometrics* **128**(1), 1–29.

Mugnier, M. (2022), 'A simple and computationally trivial estimator for grouped fixed effects models', *arXiv preprint arXiv:2203.08879* .

Neumark, D., Salas, J. M. I. and Wascher, W. (2014), 'Revisiting the minimum wage–employment debate: Throwing out the baby with the bathwater?', *ILR Review* **67**(3_suppl), 608–648.

Rivero, J. A. (2023), 'Unobserved grouped heteroskedasticity and fixed effects', *arXiv preprint arXiv:2310.14068* .

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.

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

Vaghul, K. and Zipperer, B. (2016), Historical state and sub-state minimum wage data, Working paper, Washington Center for Equitable Growth. Data release 1.4.0, through December 2022.

Weber, H. (2013), 'Demography and democracy: The impact of youth cohort size on democratic stability in the world', *Democratization* **20**(2), 335–357.

Zhang, B. (2020), 'Forecasting with bayesian grouped random effects in panel data', *arXiv preprint arXiv:2007.02435* .

Zhang, B. (2022), 'Incorporating prior knowledge of latent group structure in panel data models', *arXiv preprint arXiv:2211.16714* .

A Proofs for Section 3

Proof of Theorem 1. Fix $h \neq g$ and drop the pair subscripts. Expanding the squared norms in (13),

$$\delta\left(\|v_i-\Delta\|^2-\|v_i\|^2\right)+(1-\delta)\left(\|u_i-\nabla\|^2-\|u_i\|^2\right)=\delta\left(a-2\Delta' v_i\right)+(1-\delta)\left(b-2\nabla' u_i\right),$$

so that $E_{hg}(\delta)=\{D_i>\tau(\delta)/2\}$ with the discriminant $D_i=\delta\Delta' v_i+(1-\delta)\nabla' u_i$. Under Assumption 2, conditional on $g_i=g$, $\Delta' v_i\sim N(0,a\sigma_v^2)$ and $\nabla' u_i\sim N(0,b\sigma_u^2)$ are independent, so $D_i\sim N(0,V(\delta))$ with $V(\delta)=\delta^2 a\sigma_v^2+(1-\delta)^2 b\sigma_u^2$.

(a) $\mathbb{P}(D_i>\tau(\delta)/2)=\bar{\Phi}(\tau(\delta)/(2\sqrt{V(\delta)}))=\bar{\Phi}(c(\delta))$.

(b) Write $\tau(\delta)=\delta a+(1-\delta)b=\left(\delta\sqrt{a\sigma_v^2}\right)\sqrt{a/\sigma_v^2}+\left((1-\delta)\sqrt{b\sigma_u^2}\right)\sqrt{b/\sigma_u^2}$. By the Cauchy–Schwarz inequality,

$$\tau(\delta)^2\leq\left(\delta^2 a\sigma_v^2+(1-\delta)^2 b\sigma_u^2\right)\left(\frac{a}{\sigma_v^2}+\frac{b}{\sigma_u^2}\right)=V(\delta)\left(\mathrm{SNR}(1)^2+\mathrm{SNR}(0)^2\right),$$

which is the inequality. Equality holds if and only if the vectors $(\delta\sqrt{a\sigma_v^2},(1-\delta)\sqrt{b\sigma_u^2})$ and $(\sqrt{a/\sigma_v^2},\sqrt{b/\sigma_u^2})$ are proportional, that is, if and only if

$$\delta\sqrt{a\sigma_v^2}/\sqrt{a/\sigma_v^2}=(1-\delta)\sqrt{b\sigma_u^2}/\sqrt{b/\sigma_u^2},$$

which reduces to $\delta\sigma_v^2=(1-\delta)\sigma_u^2$ and hence to $\delta=\sigma_u^2/(\sigma_v^2+\sigma_u^2)$. The condition does not involve a, b or the pair.

(c) The vector $(v_i',D_i)'$ is Gaussian with $\mathrm{Cov}(v_i,D_i)=\delta\,\mathrm{Cov}(v_i,\Delta' v_i)=\delta\sigma_v^2\Delta$, since u_i is independent of v_i. Hence $\mathbb{E}[v_i|D_i]=\delta\sigma_v^2\Delta D_i/V(\delta)$ and, since $E_{hg}(\delta)$ is an event in D_i,

$$\mathbb{E}\left[v_i\mid E_{hg}(\delta)\right]=\frac{\delta\sigma_v^2\Delta}{V(\delta)}\mathbb{E}\left[D_i\mid D_i>\tau(\delta)/2\right]=\frac{\delta\sigma_v^2\Delta}{V(\delta)}\sqrt{V(\delta)}\,\lambda(c(\delta))=\frac{\delta\sigma_v^2\,\lambda(c(\delta))}{\sqrt{V(\delta)}}\Delta,$$

using $\mathbb{E}[Z|Z>c]=\lambda(c)$ for $Z\sim N(0,1)$. The same computation with $\mathrm{Cov}(u_i,D_i)=(1-\delta)\sigma_u^2\nabla$ gives the second offset. The Mills ratio satisfies $\lambda(c)>c$ for all c and $\lambda(c)/c\to 1$ as $c\to\infty$, so $\rho(\delta)>\delta\sigma_v^2\,c(\delta)/\sqrt{V(\delta)}=\delta\sigma_v^2\tau(\delta)/(2V(\delta))=\bar{\rho}(\delta)$ with asymptotic equivalence, and likewise for ρ^z. For the local misclassification values, $\bar{\rho}(1)=\sigma_v^2 a/(2a\sigma_v^2)=1/2$ and $\bar{\rho}(0)=0$. At δ^*, writing $x=\delta^*\sigma_v^2=(1-\delta^*)\sigma_u^2=\sigma_v^2\sigma_u^2/(\sigma_v^2+\sigma_u^2)$, one has $\tau(\delta^*)=x(a/\sigma_v^2+b/\sigma_u^2)$ and $V(\delta^*)=x^2(a/\sigma_v^2+b/\sigma_u^2)$, so $\bar{\rho}(\delta^*)=x\cdot x(a/\sigma_v^2+b/\sigma_u^2)/\left(2x^2(a/\sigma_v^2+b/\sigma_u^2)\right)=1/2$, and symmetrically $\bar{\rho}^z(\delta^*)=1/2$. Finally,

$$\begin{aligned}\bar{\rho}(\delta)-\frac{1}{2}&=\frac{\delta\sigma_v^2\tau(\delta)-V(\delta)}{2V(\delta)}=\frac{\delta\sigma_v^2\left(\delta a+(1-\delta)b\right)-\delta^2 a\sigma_v^2-(1-\delta)^2 b\sigma_u^2}{2V(\delta)}\\&=\frac{(1-\delta)b\left(\delta\sigma_v^2-(1-\delta)\sigma_u^2\right)}{2V(\delta)},\end{aligned}$$

which is negative for $\delta<\delta^*$ and positive for $\delta\in(\delta^*,1)$ since $b>0$. □

Proof of Proposition 1. For any h, $\|\tilde{r}_i-\theta_h\|^2=\delta^*\|y_i-x_i\beta-\alpha_h\|^2+(1-\delta^*)\|z_i-\zeta_h\|^2$, so the minimizer over h is (4). The stacked error has covariance $\mathrm{diag}(\delta^*\sigma_v^2 I_T,(1-\delta^*)\sigma_u^2 I_L)$, which equals $\sigma^2 I_{T+L}$ with

$\sigma^2 = \delta^*\sigma_v^2 = (1-\delta^*)\sigma_u^2 = \sigma_v^2\sigma_u^2/(\sigma_v^2+\sigma_u^2)$, whereas for $\delta \neq \delta^*$ the two blocks have different variances. The stacked model $\tilde{y}_i = \tilde{x}_i\beta + \theta_{g_i} + \tilde{v}_i$, with $\tilde{x}_i = (\sqrt{\delta^*}x_i', 0)'$, is a grouped linear model with $T+L$ periods, Gaussian spherical errors independent of the covariates and k-means classification, so its errors satisfy Assumption S.2.1(a) of González-Casasús (2026*a*) (the isotropic Gaussian case). Under parts (b) and (c) of that assumption for the stacked model, the conclusions of Theorem S.2.1 there are the local rates and the offset $\bar{v}_{hg} = \frac{1}{2}(\theta_h - \theta_g) + o(1)$ along their drifting design. The first T coordinates of $\tilde{v}_i$ are $\sqrt{\delta^*}v_i$ and the first T coordinates of $\theta_h - \theta_g$ are $\sqrt{\delta^*}\Delta_{hg}$, so $\mathbb{E}[v_i|\hat{g}_i = h, g_i = g] \to \frac{1}{2}\Delta_{hg}$. □

Proof of Corollary 1. With $G = 2$, (12) has two terms, for the flows $1 \to 2$ and $2 \to 1$. With $s_{ab} = 0$ and $\gamma_{ab} = (1-\bar{\rho}(\delta))(\alpha_a - \alpha_b)$,

$$\text{Bias}(\hat{\beta}(\delta)) = S^{-1}\pi_1(\mu_2-\mu_1)'(1-\bar{\rho}(\delta))(\alpha_2-\alpha_1)p_{2|1} + S^{-1}\pi_2(\mu_1-\mu_2)'(1-\bar{\rho}(\delta))(\alpha_1-\alpha_2)p_{1|2}.$$

By Theorem 1(a) with $a_{21} = a_{12}$ and $b_{21} = b_{12}$, the two misclassification probabilities coincide, so $p_{2|1} = p_{1|2} = p(\delta)$, and the two offsets coincide as well. Since $(\mu_1-\mu_2)'(\alpha_1-\alpha_2) = (\mu_2-\mu_1)'(\alpha_2-\alpha_1)$ and $\pi_1+\pi_2 = 1$, the sum equals $(1-\bar{\rho}(\delta))p(\delta)\varkappa$. The variance is V for every δ by Assumption 1. For (18), note that $\bar{\rho}(\delta^*) = \bar{\rho}(1) = 1/2$ and $\bar{\rho}(0) = 0$ by (16), and that $p(\delta^*) < p(1)$ and $p(\delta^*) < p(0)$ because $\bar{\Phi}$ is strictly decreasing and $c(\delta^*) > \max\{c(1), c(0)\}$ by Theorem 1(b) when $a_{21}, b_{21} > 0$. □

Proof of Theorem 2. With $z_i = x_i = \mu_{g_i} + e_i$, the proxy view has centers $\zeta_g = \mu_g$ and noise $u_i = e_i$, so Theorem 1 applies with $\nabla_{hg} = \mu_h - \mu_g$ and σ_u^2 the variance of e_i. By Theorem 1(c), the units misclassified from g to h have $\mathbb{E}[v_i|E_{hg}(\delta)] = \rho(\delta)\Delta_{hg}$ and $\mathbb{E}[x_i - \mu_g|E_{hg}(\delta)] = \mathbb{E}[e_i|E_{hg}(\delta)] = \rho^z(\delta)(\mu_h - \mu_g)$. Their local misclassification values are $\bar{\rho}(\delta)\Delta_{hg}$ and $\bar{\rho}^z(\delta)(\mu_h - \mu_g)$, which are the offsets and the covariate shifts in the statement. Substituting $\gamma_{ab} = (1-\bar{\rho}(\delta))(\alpha_a - \alpha_b)$ and $s_{ab} = \bar{\rho}^z(\delta)(\mu_a - \mu_b)$ in (12),

$$\kappa_{ab} - \kappa_{ab}^s = S^{-1}\pi_b\left[(1-\bar{\rho}(\delta)) - \bar{\rho}^z(\delta)\right](\mu_a-\mu_b)'(\alpha_a-\alpha_b),$$

and summing over the two flows as in the proof of Corollary 1 gives (19). At δ^*, $\bar{\rho}(\delta^*) = \bar{\rho}^z(\delta^*) = 1/2$ by (16), so the bias vanishes and (11) gives the mean and the variance of the oracle. At $\delta = 1$, $\bar{\rho}(1) = 1/2$ and $\bar{\rho}^z(1) = 0$, and at $\delta = 0$, $\bar{\rho}(0) = 0$ and $\bar{\rho}^z(0) = \sigma_u^2(b/2)/(b\sigma_u^2) = 1/2$. □

Proof of Proposition 2. Under the conditions of Corollary 1 at $\delta = 1$ and $\delta = 0$, (17) gives $\text{Bias}(\hat{\beta}(1)) = \frac{1}{2}p(1)\varkappa$ and $\text{Bias}(\hat{\beta}(0)) = p(0)\varkappa$. Assumption 1 applies to each estimator and, to the same order, both are linear in the same oracle score (González-Casasús, 2026*a*, Theorem 1), so $\tilde{\beta} - \beta = 2(\hat{\beta}(1) - \beta) - (\hat{\beta}(0) - \beta)$ has mean $2\cdot\frac{1}{2}p(1)\varkappa - p(0)\varkappa = (p(1)-p(0))\varkappa$ and variance $(2-1)^2V = V$. The bias exceeds $\frac{1}{2}p(1)\|\varkappa\|$ in norm if and only if $|p(1)-p(0)| > p(1)/2$, which is $p(0) < p(1)/2$ or $p(0) > 3p(1)/2$. When the proxy is an included regressor, Theorem 2 gives $\text{Bias}(\hat{\beta}(0)) = \frac{1}{2}p(0)\varkappa$, so the mean becomes $(p(1) - \frac{1}{2}p(0))\varkappa$, which equals $\frac{1}{2}p(1)\varkappa$ when $p(1) = p(0)$. The variance statement follows from the same linear representation. □

B Proof for Section 4

Proof of Proposition 3. (a) By (23), Q_δ is the GFE objective of Bonhomme and Manresa (2015) for the stacked panel with profiles θ_g, so the joint minimizers of the two objectives coincide, with $\hat{\theta}_g = (\sqrt{\delta}\hat{\alpha}_g', \sqrt{1-\delta}\hat{\zeta}_g')'$, and in particular their partitions coincide. It remains to show that, for any partition g, the least squares

estimator of β on the stacked panel equals (5) on the original panel. Let $\ddot{\tilde{x}}_{it}$ and $\ddot{\tilde{y}}_{it}$ be the stacked variables after removing (g_i, t)-cell means over $t \in [\tilde{T}]$. In the outcome block ($t \le T$), the cell means of $\tilde{x}_{it} = \sqrt{\delta} x_{it}$ are $\sqrt{\delta}$ times those of x_{it}, so $\ddot{\tilde{x}}_{it} = \sqrt{\delta}\ddot{x}_{it}$ and $\ddot{\tilde{y}}_{it} = \sqrt{\delta}\ddot{y}_{it}$, and in the proxy block ($t > T$) the covariates are identically zero, so $\ddot{\tilde{x}}_{it} = 0$. Hence

$$\left(\sum_i \sum_{t \le \tilde{T}} \ddot{\tilde{x}}_{it}\ddot{\tilde{x}}_{it}'\right)^{-1} \sum_i \sum_{t \le \tilde{T}} \ddot{\tilde{x}}_{it}\ddot{\tilde{y}}_{it} = \left(\delta \sum_i \sum_{t \le T} \ddot{x}_{it}\ddot{x}_{it}'\right)^{-1} \delta \sum_i \sum_{t \le T} \ddot{x}_{it}\ddot{y}_{it},$$

which is (5) for the partition g. Applied to the estimated partition, this identifies the CAS estimator with the GFE estimator of the stacked panel, and applied to the true partition it identifies the two oracle estimators.

(b) By (24), $\tilde{T}^{-1}\|\theta_g - \theta_h\|^2 = \frac{T}{T+L}\,\delta\, T^{-1}\|\alpha_g - \alpha_h\|^2 + \frac{L}{T+L}(1-\delta)L^{-1}\|\zeta_g - \zeta_h\|^2$, which is the equality, and both terms are nonnegative, which gives the bound by their maximum.

(c) By the proof of Proposition 1, the stacked error at δ^* has covariance $\sigma^2 I_{\tilde{T}}$ with $\sigma^2 = \delta^*\sigma_v^2 = (1-\delta^*)\sigma_u^2 = \sigma_v^2\sigma_u^2/(\sigma_v^2 + \sigma_u^2)$. Moreover $\|\theta_g - \theta_h\|^2 = \delta^* a_{hg} + (1-\delta^*)b_{hg}$, so $\|\theta_g - \theta_h\|^2/\sigma^2 = a_{hg}/\sigma_v^2 + b_{hg}/\sigma_u^2 = \mathrm{SNR}_{hg}(\delta^*)^2$ by Theorem 1(b), and Theorem 1(a) gives $\bar{p}_{h|g}(\delta^*) = \bar{\Phi}(\mathrm{SNR}_{hg}(\delta^*)/2) = \bar{\Phi}(\|\theta_g - \theta_h\|/(2\sigma))$. The first inequality follows from Theorem 1(b) and the monotonicity of $\bar{\Phi}$, since $\mathrm{SNR}_{hg}(\delta^*) \ge \max\{\mathrm{SNR}_{hg}(1), \mathrm{SNR}_{hg}(0)\}$. The second is the Gaussian tail bound $\bar{\Phi}(c) \le \frac{1}{2}\exp(-c^2/2)$ for $c \ge 0$, applied at $c = \|\theta_g - \theta_h\|/(2\sigma)$. □

Supplemental Appendix to "Covariate-Assisted Estimation of Grouped Panel Data Models"

Oriol González-Casasús

This supplemental appendix contains results that complement the paper. Section S.1 gives primitive conditions for Assumption 1 under the two-view rule. Section S.2 extends Theorem 1 and Corollary 1 to more than two groups, Section S.3 to a proxy that is misaligned for a fraction of the units, and Section S.4 extends Theorem 2 to several regressors. Section S.5 reports the simulation results that Sections 5.1 to 5.3 of the paper discuss, and Sections S.6 and S.7 report further results for the two applications of Section 6.

S.1 Primitive Conditions for Assumption 1

This section gives primitive conditions under which the two-view rule with estimated centers and a preliminary weight satisfies Assumption 1, with the probabilities and offsets of Theorem 1. The $G \geq 3$ version of Corollary 1 announced in Remark 3 and the extension of Theorem 2 to several regressors are in Sections S.2 and S.4.

Assumption 1 is derived in González-Casasús (2026*a*) from conditions on the estimated memberships, and Supplemental Appendix S.2 there shows that the k-means rule of the GFE estimator, with centers and slopes estimated from the same sample, generates memberships that satisfy those conditions along a drifting design in which the groups separate slowly. The following assumption transposes that design to the two-view rule. Throughout, $n = N$, the stacked objects are those of Proposition 1, $\theta_g^0 = (\sqrt{\delta^*}\alpha_g^{0\prime}, \sqrt{1-\delta^*}\zeta_g^{0\prime})'$, $\sigma^2 = \sigma_v^2\sigma_u^2/(\sigma_v^2+\sigma_u^2)$, and $d^* = \min_{h\neq g}\|\theta_h^0 - \theta_g^0\|$.

Assumption S.1 (Two-view k-means design). *(a)* (Errors) *Assumption 2 holds, (v_i, u_i) is independent of (x_i, g_i), and the distribution of (v_i, u_i) does not depend on n.*

(b) (Drifting separation) *(x_i, g_i, v_i, u_i) are i.i.d. across i, the distribution of (x_i, g_i) and the slope β do not depend on n, and the profiles and centers are $\alpha_{n,g} = s_n\alpha_g^0$ and $\zeta_{n,g} = s_n\zeta_g^0$, where $\theta_1^0, \ldots, \theta_G^0$ are distinct and $s_n \to \infty$ is defined by $\sqrt{n}\,\bar{\Phi}(s_n d^*/(2\sigma)) = q_0$ for a given $q_0 \in (0, \infty)$.*

(c) (Classification) *$\hat{g}_i = \arg\min_{h\in[G]} \tilde{\delta}\|y_i - x_i\tilde{\beta} - \tilde{\alpha}_h\|^2 + (1-\tilde{\delta})\|z_i - \tilde{\zeta}_h\|^2$, where the preliminary parameters $\tilde{\psi} = (\tilde{\beta}', \tilde{\alpha}', \tilde{\zeta}')'$ and the preliminary weight $\tilde{\delta}$ may depend on the whole sample and satisfy $\|\tilde{\psi} - \psi_n\| = O_p(n^{-1/4-\eta})$ and $|\tilde{\delta} - \delta^*| = O_p(n^{-1/4-\eta})$ for some $\eta > 0$, with $\psi_n = (\beta', \alpha_n', \zeta_n')'$.*

Part (a) is the Gaussian design of Section 3.2 with exogenous proxy noise. The Gaussian requirement is not incidental here. The microfoundation of González-Casasús (2026*a*) allows any spherically symmetric error with light tails, but the two-view rule at the precision weight is a Euclidean rule on the stacked error $(\sqrt{\delta^*}v_i', \sqrt{1-\delta^*}u_i')'$, and a vector with independent spherical blocks is itself spherical only when it is Gaussian (Kagan et al., 1973, Theorem 5.3.3), so the stacked error falls in their class only in the Gaussian case. Part (b) is their drifting design, with the stacked centers in place of the profiles, so that the pairs of groups at minimal stacked distance are confused with probability of order $n^{-1/2}$. Part (c) is their high-level condition on the rate of the preliminary parameters that the rule uses, extended to the weight, which is one more parameter of the rule. It is a condition on the availability of preliminary estimates with a polynomial

rate. A two-step version of the estimator, with a preliminary weight and preliminary parameters of the required rate, is covered as stated. For the fixed points of Algorithm 1, in which the parameters and the weight are computed from the partition they define, part (c) of the proposition establishes that the QML weight computed from any partition with the properties in (a) has the required rate, and Corollary S.1 below transposes the argument of Corollary S.2.2 of González-Casasús (2026*a*) to show that every fixed point of the algorithm in a shrinking neighborhood of the truth satisfies part (c), so that the conclusions hold for the local solutions of the algorithm near the truth. The global minimizer of (6) is not covered, as the global minimizer of the GFE objective is not covered beyond two groups in that paper.

Proposition S.1 (Assumption 1 for the two-view rule)**.** *Let Assumption S.1 hold together with the sampling conditions of Assumption 1(b)–(c) of González-Casasús (2026a) for the outcome equation, and let $\mathcal{N}$ be the set of pairs (g,h) with $h \neq g$ and $\|\theta_h^0 - \theta_g^0\| = d^*$. Then, with $n_{hg} = \sum_i \mathbb{1}_{\{\hat{g}_i=h, g_i=g\}}$,*

(a) (Probabilities and exogeneity) *For every $h \neq g$, $n^{-1/2} n_{hg} \xrightarrow{p} \pi_g q_{h|g}$ with $q_{h|g} = q_0 \mathbb{1}_{\{(g,h)\in\mathcal{N}\}}$, and*

$$n^{-1/2} \sum_i \mathbb{1}_{\{\hat{g}_i=h, g_i=g\}} (x_i - \mu_g) \xrightarrow{p} 0;$$

(b) (One-half offsets) *For every $(g,h) \in \mathcal{N}$,*

$$n^{-1/2} \sum_i \mathbb{1}_{\{\hat{g}_i=h, g_i=g\}} v_i - \pi_g q_{h|g} \rho_n (\alpha_{n,h} - \alpha_{n,g}) \xrightarrow{p} 0,$$
$$n^{-1/2} \sum_i \mathbb{1}_{\{\hat{g}_i=h, g_i=g\}} u_i - \pi_g q_{h|g} \rho_n (\zeta_{n,h} - \zeta_{n,g}) \xrightarrow{p} 0,$$

with $\rho_n = \lambda(c_n)/(2c_n) \to \frac{1}{2}$, where $c_n = s_n d^*/(2\sigma)$; in particular the centering can be taken to be $\frac{1}{2}\pi_g q_{h|g}(\alpha_{n,h} - \alpha_{n,g})$;

(c) (Updated weight) *The within-group variances of Algorithm 1 evaluated at $\hat{g}$ satisfy*

$$\hat{\sigma}_v^2 - \sigma_v^2 = O_p(n^{-1/2} s_n^2), \qquad \hat{\sigma}_u^2 - \sigma_u^2 = O_p(n^{-1/2} s_n^2),$$

so that the updated weight satisfies $\hat{\delta} - \delta^* = O_p(n^{-1/2} s_n^2) = o_p(n^{-1/4-\eta})$ for every $\eta < 1/4$, which is the rate required of $\tilde{\delta}$ in Assumption S.1(c).

Consequently, the two-view rule satisfies Assumption 1 along the design, with $p_{h|g} = q_{h|g}/\sqrt{n}$, offsets $\rho_{hg} = 1/2$ and $s_{hg} = 0$, and Theorem 1 and Corollary 1 of González-Casasús (2026a) give

$$\sqrt{n}(\hat{\beta} - \beta) - \sum_{(g,h)\in\mathcal{N}} \kappa_{n,hg}\, q_{h|g} \xrightarrow{d} N(0, V_\beta), \qquad \kappa_{n,hg} = \tfrac{1}{2} S^{-1} \pi_g (\mu_h - \mu_g)'(\alpha_{n,h} - \alpha_{n,g}) = s_n \kappa_{hg}^0,$$

with $\kappa_{hg}^0 = \frac{1}{2} S^{-1} \pi_g (\mu_h - \mu_g)'(\alpha_h^0 - \alpha_g^0)$, and the centering sequence is the first-order bias of Assumption 1 evaluated at the drifting profiles. It grows like $s_n \asymp \sqrt{\log n}$, so it is of smaller order than $\sqrt{n}$ and the estimator is consistent, while the bias is not negligible relative to the standard error at any sample size along the design.

Proof of Proposition S.1. Write the model in stacked form, $\tilde{y}_i = \tilde{x}_i \beta + \theta_{n,g_i} + \tilde{v}_i$ with

$$\tilde{y}_i = (\sqrt{\delta^*} y_i', \sqrt{1-\delta^*} z_i')', \qquad \tilde{x}_i = (\sqrt{\delta^*} x_i', 0)', \qquad \theta_{n,g} = s_n \theta_g^0, \qquad \tilde{v}_i = (\sqrt{\delta^*} v_i', \sqrt{1-\delta^*} u_i')'.$$

Under Assumption S.1(a), $\tilde{v}_i \sim N(0, \sigma^2 I_{T+L})$ is independent of $(\tilde{x}_i, g_i)$, which is Assumption S.2.1(a) of González-Casasús (2026*a*) with tail index equal to two (the hazard rate of the Gaussian is bounded by a polynomial and its radial part has a Weibull tail with that index). Assumption S.1(b) is their Assumption S.2.1(b) for the stacked panel, with $\sqrt{n}\,\mathbb{P}(\tilde{v}_{i1} \geq s_n d^*/2) = \sqrt{n}\,\bar{\Phi}(s_n d^*/(2\sigma)) = q_0$. For the classification condition, the two-view rule with $\tilde{\delta} = \delta^*$ is the Euclidean k-means rule on the stacked data with parameters $(\tilde{\beta}, \tilde{\theta})$, $\tilde{\theta}_h = (\sqrt{\delta^*}\tilde{\alpha}_h', \sqrt{1-\delta^*}\tilde{\zeta}_h')'$, which satisfies their Assumption S.2.1(c) because $\|\tilde{\theta} - \theta_n\| \leq \|\tilde{\psi} - \psi_n\|$. With $\tilde{\delta} \neq \delta^*$, the difference between the weighted distances to two centers h and k equals the difference between the stacked Euclidean distances plus $(\tilde{\delta} - \delta^*)\big[\|r_i - \tilde{\alpha}_h\|^2 - \|r_i - \tilde{\alpha}_k\|^2 - \|z_i - \tilde{\zeta}_h\|^2 + \|z_i - \tilde{\zeta}_k\|^2\big]$ with $r_i = y_i - x_i\tilde{\beta}$, and the bracket is a continuously differentiable function of the parameters whose gradient is of order $s_n + \|\tilde{v}_i\|$, exactly as the perturbation induced by $\tilde{\psi} - \psi_n$ in the proof of their Lemma S.2.3. The weight therefore enters the sandwich argument of that lemma as one more coordinate of $\tilde{\psi}$, with the rate $n^{-1/4-\eta}$, and the conclusions of their Theorem S.2.1 hold for the stacked problem. Part (a) is their part (a), since the pairs at minimal distance in the stacked metric are the pairs of $\mathcal{N}$. Part (b) is their part (b) applied to the stacked error: the centering is $\pi_g q_{h|g} \rho_n (\theta_{n,h} - \theta_{n,g})$ with $\rho_n = \frac{1}{2} + m_1(s_n d^*/2)/(s_n d^*)$, where $m_1(c) = \mathbb{E}[\tilde{v}_{i1} - c \mid \tilde{v}_{i1} > c] = \sigma\lambda(c/\sigma) - c$ is the mean residual life of a $N(0, \sigma^2)$ variable, so that $\rho_n = \sigma\lambda(c_n)/(s_n d^*) = \lambda(c_n)/(2c_n)$, which is the exact offset (15) at δ^* evaluated at the drifting separation, and projecting on the two blocks of $\theta_{n,h} - \theta_{n,g}$ gives the two displays. For part (c), the within-group variance of the outcome view under $\hat{g}$ differs from the one under the true groups by the contribution of the misclassified units, whose number is $O_p(\sqrt{n})$ by part (a) and each of which contributes a residual of order s_n^2 to the sum of squares, so the difference is $O_p(n^{-1/2} s_n^2)$ after division by nT, while the variance under the true groups is within $O_p(n^{-1/2})$ of σ_v^2 by the law of large numbers and the $\sqrt{n}$-consistency of the oracle estimator. The same holds for the proxy view, and the ratio $\hat{\delta} = \hat{\sigma}_u^2/(\hat{\sigma}_v^2 + \hat{\sigma}_u^2)$ inherits the rate. Since $s_n^2 = O(\log n)$ under Gaussian tails, $n^{-1/2} s_n^2 = o(n^{-1/4-\eta})$ for every $\eta < 1/4$. The final statement is Theorem 1 and Corollary 1 of González-Casasús (2026*a*), whose Theorem S.2.2 re-establishes them along the drifting design. □

Two remarks. First, the proposition covers the precision weight and any preliminary weight with the rate of Assumption S.1(c), and part (c) shows that the weight updated from a partition with the properties in (a) has that rate, which is what makes the fixed-point condition on $\tilde{\delta}$ self-consistent. The parameters updated from such a partition are within $O_p(n^{-1/2} s_n)$ of the truth by the same argument, since the misclassified units are a fraction of order $n^{-1/2}$ at a distance of order s_n, so the induction covers any fixed number of updates of the weight and the parameters, and the following corollary covers the fixed point. For a fixed weight $\delta \neq \delta^*$, the stacked error is Gaussian but not spherical, so the rule is an elliptical k-means rule. The probabilities and offsets with population centers are those of Theorem 1, and the sandwich argument of González-Casasús (2026*a*), which uses only the Gaussian tail of the discriminant and the rate of the parameters, goes through with the discriminant D_i of the proof of Theorem 1 in place of the Euclidean one; the details are omitted. Second, the one-half offset in part (b) is the local misclassification value $\bar{\rho}(\delta^*)$ of Theorem 1(c), reached along the design because $c_n \to \infty$. The overshoot $\rho_n - \frac{1}{2}$ is the finite-sample correction displayed in Figure 2.

Corollary S.1 (Local solutions of Algorithm 1)**.** *Let Assumption S.1(a)–(b) hold together with the sampling conditions of Assumption 1(b)–(c) of González-Casasús (2026a) for the outcome equation, let $r_n = n^{-1/4-\eta/2}$ for some $\eta \in (0, 1/2)$, and let $\mathsf{T}(\tilde{\psi}, \tilde{\delta})$ denote the parameters and the weight that Step 2 of Algorithm 1 returns from the partition assigned by Step 3 with parameters $\tilde{\psi}$ and weight $\tilde{\delta}$, ties being broken by a fixed rule. Then* T

maps the ball $\mathcal{B}_n = \{(\tilde{\psi}, \tilde{\delta})\colon \|\tilde{\psi} - \psi_n\| \le r_n,\ |\tilde{\delta} - \delta^*| \le r_n\}$ *into itself with probability tending to one, every sequence of iterations of Steps 2 and 3 started in* $\mathcal{B}_n$ *converges in finitely many steps to a fixed point in* $\mathcal{B}_n$*, and every such fixed point satisfies Assumption S.1(c) with exponent* $\eta/2$*, so that the conclusions of Proposition S.1 hold for it.*

Proof of Corollary S.1. The argument is that of Corollary S.2.2 of González-Casasús (2026*a*), with the weight as one more coordinate. On the event $\{\max_i \|\tilde{v}_i\| \le \Lambda_0 s_n\}$, whose probability tends to one for Λ_0 large by the Gaussian tail and a union bound (their Lemma S.2.1(b)), the perturbation of the weighted distances induced by $(\tilde{\psi}, \tilde{\delta}) \in \mathcal{B}_n$ relative to (ψ_n, δ^*) is bounded by $Cr_n(s_n + \|\tilde{v}_i\|) \le C' r_n s_n$ for every unit, as in the proof of Proposition S.1, so their sandwich Lemma S.2.3 holds simultaneously for all $(\tilde{\psi}, \tilde{\delta}) \in \mathcal{B}_n$ with radii that do not depend on $(\tilde{\psi}, \tilde{\delta})$. The sandwich sums are therefore the same for every point of $\mathcal{B}_n$, and the proof of Proposition S.1 shows that its conclusions (a) and (b) hold uniformly over $\mathcal{B}_n$, while the set of units misclassified under any partition assigned from $\mathcal{B}_n$ is contained in a set of cardinality $O_p(\sqrt{n})$ that does not depend on the point. The least squares parameters computed from such a partition are within $O_p(n^{-1/2} s_n)$ of ψ_n, uniformly over $\mathcal{B}_n$, since the misclassified units are $O_p(\sqrt{n})$ in number and at a distance of order s_n from their profiles, and the updated weight is within $O_p(n^{-1/2} s_n^2)$ of δ^*, uniformly, by the argument of part (c). Both rates are $n^{-1/2+o(1)} = o(r_n)$ because $s_n \asymp \sqrt{\log n}$ and $\eta < 1/2$, so $\mathsf{T}(\mathcal{B}_n) \subset \mathcal{B}_n$ with probability tending to one. Steps 2 and 3 weakly decrease (6), there are finitely many partitions, and the parameters and the variances are unique functions of the partition while the assignment is a function of the parameters and the weight under the tie-breaking rule, so an iteration started in $\mathcal{B}_n$ reaches a fixed point in finitely many steps; all its iterates lie in $\mathcal{B}_n$, hence so does the fixed point, which is Assumption S.1(c) with $\|\tilde{\psi} - \psi_n\| \le r_n$ and $|\tilde{\delta} - \delta^*| \le r_n$, that is, with exponent $\eta/2$. □

S.2 More Than Two Groups

Let $G \ge 3$ and consider the precision-weighted rule, so that by Proposition 1 the stacked error is spherical with variance σ^2. The centers are those of the drifting design of Assumption S.1(b), $\theta_{n,g} = s_n \theta_g^0$ with $s_n \to \infty$, so that the limit statements below are along the local misclassification design of González-Casasús (2026*a*), and the index n is omitted. A unit of group g is assigned to h when $\tilde{v}_i \in C_{hg} = \{w\colon \|\theta_g + w - \theta_h\| \le \|\theta_g + w - \theta_k\|\ \forall k \ne h\}$, the Voronoi cell of θ_h translated by $-\theta_g$, which is a convex polyhedron contained in the half-space $H_{hg} = \{w\colon (\theta_h - \theta_g)'w \ge \|\theta_h - \theta_g\|^2/2\}$. Its point closest to the origin is the midpoint $m_{hg} = (\theta_h - \theta_g)/2$ if and only if $m_{hg} \in C_{hg}$, that is, if no center θ_k is closer to m_{hg} than θ_h is, which is the condition that the open ball with diameter $[\theta_g, \theta_h]$ contains no other center. The condition is invariant to the scale s_n, so it is a condition on $\theta_1^0, \dots, \theta_G^0$. Pairs of centers that are nearest neighbors always satisfy it, since every point of that ball is closer to θ_g than θ_h is.

Proposition S.2 (More than two groups)**.** *Let Assumptions 1 and 2 hold for the precision-weighted rule with population centers along the design of Assumption S.1(b), with* $G \ge 3$ *and an external proxy with* $s_{ab} = 0$*. Then, for every ordered pair* $h \ne g$*,*

(a) (Pairwise bound) *Whenever* $a_{hg}, b_{hg} > 0$*,*

$$p_{h|g}(\delta^*) = \mathbb{P}(\tilde{v}_i \in C_{hg}) \le \bar{p}_{h|g}(\delta^*) = \bar{\Phi}\big(\mathrm{SNR}_{hg}(\delta^*)/2\big) < \min\{\bar{p}_{h|g}(1), \bar{p}_{h|g}(0)\},$$

and the first inequality is an equality up to a factor $1 + o(1)$ *as* $s_n \to \infty$ *for the pairs that satisfy the*

ball condition at δ^*. The same holds at $\delta = 1$ and $\delta = 0$ for the pairs that satisfy the ball condition at those weights, in which case $p_{h|g}(\delta^*) < \min\{p_{h|g}(1), p_{h|g}(0)\}$ for n large;

(b) (Offsets) *If the pair (g, h) satisfies the ball condition, then $\mathbb{E}[v_i \mid \tilde{v}_i \in C_{hg}, g_i = g] = \rho_{hg}\Delta_{hg}$ with $\rho_{hg} \to 1/2$ as $s_n \to \infty$, and $\mathbb{E}[x_i - \mu_g \mid \tilde{v}_i \in C_{hg}, g_i = g] = 0$;*

(c) (Bias) $\mathrm{Bias}(\hat{\beta}(\delta^*)) = \sum_{(a,b)}(1-\rho_{ab})\kappa^0_{ab}p_{a|b}(\delta^*)$ *with* $\kappa^0_{ab} = S^{-1}\pi_b(\mu_a - \mu_b)'(\alpha_a - \alpha_b)$ *(the sensitivity κ_{ab} of (12) without the offset factor), and the flows that violate the ball condition have probabilities of smaller order than the nearest-neighbor flows as $s_n \to \infty$. For bounds $\bar{p}_b$ on the outgoing probabilities $\sum_{a\neq b} p_{a|b}$, the worst-case bias $\sum_b \bar{p}_b \max_{a\neq b} \frac{1}{2}\|\kappa^0_{ab}\|$ is smallest under the precision weight with bounds $\bar{p}_b(\delta^*) = \sum_{a\neq b}\bar{p}_{a|b}(\delta^*)$, each term of which is strictly smaller than the corresponding pairwise bound at $\delta = 1$ or $\delta = 0$.*

Proof of Proposition S.2. (a) $C_{hg} \subset H_{hg}$ gives $p_{h|g} \leq \mathbb{P}(\tilde{v}_i \in H_{hg}) = \bar{\Phi}(\|m_{hg}\|/\sigma) = \bar{\Phi}(\mathrm{SNR}_{hg}(\delta^*)/2)$ by the proof of Proposition 3(c), and the strict inequality is Theorem 1(b). Under the ball condition, the constraints $k \neq g, h$ are slack at m_{hg}, so in a neighborhood of m_{hg} the boundary of C_{hg} is the hyperplane bounding H_{hg}; as $s_n \to \infty$, the cell $C_{hg} = s_n C^0_{hg}$ recedes from the origin at the rate s_n while the error distribution is fixed, and the Gaussian measure of a receding convex set is concentrated near its closest point to the origin, so $\mathbb{P}(\tilde{v}_i \in C_{hg}) = \bar{\Phi}(\|m_{hg}\|/\sigma)(1 + o(1))$, which is Proposition S.2.1 of González-Casasús (2026*a*). (b) By the same proposition, $\mathbb{E}[\tilde{v}_i \mid \tilde{v}_i \in C_{hg}] = m_{hg} + o(\|m_{hg}\|)$ under the ball condition, and the first T coordinates of m_{hg} are $\sqrt{\delta^*}\Delta_{hg}/2$, which gives the offset in the original scale; the covariate shift is zero because x_i is independent of $\tilde{v}_i$ given g_i. (c) The bias expression is (12) with $s_{ab} = 0$. For a pair that violates the ball condition, the point of C_{hg} closest to the origin is farther from the origin than m_{hg}, so its probability is $\bar{\Phi}(\|p^*_{hg}\|/\sigma)(1 + o(1))$ with $\|p^*_{hg}\| > \|m_{hg}\|$, which is of smaller order than the probability of the nearest-neighbor pairs, whose $\|m_{hg}\|$ is minimal. The worst-case bias over outgoing bounds is the sharp bound of Proposition 2 in González-Casasús (2026*a*), which is nondecreasing in each $\bar{p}_b$, and the bounds $\bar{p}_b(\delta^*)$ dominate the outgoing probabilities by part (a), each term being smaller than at $\delta \in \{0, 1\}$. □

The proposition is the content of Remark 3. Every flow has a smaller pairwise bound under the precision weight than under either single-view rule, the bounds are attained to first order by the flows that satisfy the ball condition, the offset of those flows is one half, and the bias set built from outgoing bounds shrinks accordingly, which is the sense in which the ordering (18) extends to $G \geq 3$. The comparison of the misclassification probabilities themselves, rather than of their bounds, requires the ball condition at the three weights, since the stacked centers, and hence the Voronoi cells, depend on the weight. The signed bias need not be ordered, since different flows can have sensitivities κ_{ab} of different signs. For a fixed $\delta \neq \delta^*$, the same argument applies to the whitened error $\Lambda^{-1/2}\tilde{v}_i$ with $\Lambda = \mathrm{diag}(\delta\sigma_v^2 I_T, (1-\delta)\sigma_u^2 I_L)$ and the whitened cells, whose pairwise half-spaces reproduce the discriminant of the proof of Theorem 1, and the ball condition becomes an ellipsoid condition in the original coordinates.

S.3 Partially Valid Proxies

Theorem 1 maintains that the proxy shares the partition of the outcome equation. This section relaxes it in the direction that the misaligned designs of Supplemental Appendix S.5 simulate, and shows that

the geometry of Section 3.2 describes the departure as well. Take $G = 2$ and suppose that, within each group, a fraction $\pi \in [0, 1/2)$ of the units (a scalar, not to be confused with the group shares π_g), drawn independently of the errors and the covariates, has its proxy centered at the other group's center, $z_i = \zeta_h + u_i$ for a unit of group g, $h \neq g$. The proxy is valid for the remaining units and misaligned for these. At $\pi = 1/2$, the proxy carries no information about the groups, and beyond it the labels are reversed, so $\pi < 1/2$ is without loss of generality. As in Theorem 1, the computation conditions on population centers, the profiles α_g and the centers ζ_g of the proxy's own partition, from which the centers of the shared-partition fit differ by an amount of order π.

Fix the pair and drop its subscripts, so that a, b, $\tau(\delta)$, $V(\delta)$, $c(\delta)$, D_i and $\rho(\delta)$ are as in Theorem 1 and its proof, and write $p_A(\delta) = \bar{\Phi}(c(\delta))$. Define the misaligned counterparts

$$\tau_m(\delta) = \delta a - (1-\delta)b, \qquad c_m(\delta) = \frac{\tau_m(\delta)}{2\sqrt{V(\delta)}}, \qquad p_M(\delta) = \bar{\Phi}(c_m(\delta)), \qquad \rho_m(\delta) = \frac{\delta\sigma_v^2\,\lambda(c_m(\delta))}{\sqrt{V(\delta)}}, \tag{S.1}$$

and the bias coefficients of the two kinds of units,

$$B_A(\delta) = (1-\rho(\delta))p_A(\delta), \qquad B_M(\delta) = (1-\rho_m(\delta))p_M(\delta).$$

Proposition S.3 (Partially valid proxy)**.** *Let Assumption 2 hold with $G = 2$ and $a, b > 0$, and let a fraction $\pi \in [0, 1/2)$ of the units of each group be misaligned. For every $\delta \in [0, 1]$, the following hold.*

(a) (Misaligned units) *A misaligned unit of group g is assigned to h on the event $\{D_i > \tau_m(\delta)/2\}$, which has probability $p_M(\delta)$, and its expected outcome shock on that event is $\rho_m(\delta)\Delta_{hg}$. The function c_m is strictly increasing on $[0, 1]$, with $c_m(0) = -c(0)$, $c_m(1) = c(1)$ and*

$$c_m(\delta^*) = \frac{\mathrm{SNR}(1)^2 - \mathrm{SNR}(0)^2}{2\sqrt{\mathrm{SNR}(1)^2 + \mathrm{SNR}(0)^2}}.$$

(b) (Rates and bias) *The misclassification probability of the rule is $p(\delta;\pi) = (1-\pi)p_A(\delta) + \pi p_M(\delta)$, and under Assumption 1 for the rule, with $s_{ab} = 0$ and the offsets of Theorem 1(c) and of part (a), its first-order bias is $\mathrm{Bias}(\hat{\beta}(\delta)) = B(\delta;\pi)\varkappa$ with $B(\delta;\pi) = (1-\pi)B_A(\delta) + \pi B_M(\delta)$ and $\varkappa$ as in Corollary 1. Both are the GFE values at $\delta = 1$ for every π.*

(c) (Weights) *$\partial p(\delta;\pi)/\partial\delta = -\pi\phi(c_m(\delta^*))c_m'(\delta^*) < 0$ at $\delta = \delta^*$ when $\pi > 0$, so the weight that minimizes the misclassification probability exceeds δ^*. If $\rho(1) < 1$, then $\partial p(\delta;\pi)/\partial\delta$ and $\partial B(\delta;\pi)/\partial\delta$ at $\delta = 1$ are both positive multiples of $1-2\pi$, so for every $\pi < 1/2$ there are weights below one at which the rule has a smaller misclassification probability and a smaller first-order bias than the GFE rule.*

(d) (Tolerance) *Suppose $\rho(\delta) \leq 1$, so that $B_A(\delta) \geq 0$. For $\delta < 1$, $B_M(\delta) > B_A(\delta)$, and the rule with weight δ has a smaller first-order bias than the GFE rule, in absolute value, if and only if*

$$\pi < \bar{\pi}(\delta) = \frac{B_A(1) - B_A(\delta)}{B_M(\delta) - B_A(\delta)},$$

which is positive if and only if $B_A(\delta) < B_A(1)$. If $\rho(1) < 1$, $\bar{\pi}(\delta) \to 1/2$ as $\delta \to 1$.

Proof of Proposition S.3. (a) For a misaligned unit of group g, the squared distances to the two pairs of centers are $\delta\|v_i\|^2 + (1-\delta)\|u_i + \nabla\|^2$ and $\delta\|v_i - \Delta\|^2 + (1-\delta)\|u_i\|^2$. Expanding as in the proof of Theorem 1, the second is smaller if and only if $\delta(a - 2\Delta' v_i) < (1-\delta)(b + 2\nabla' u_i)$, that is, $D_i > \tau_m(\delta)/2$. Since $D_i \sim N(0, V(\delta))$, the probability is $\bar{\Phi}(c_m(\delta))$, and the Gaussian regression of v_i on D_i in that proof gives $\mathbb{E}[v_i \mid D_i > \tau_m(\delta)/2] = \delta\sigma_v^2\Delta\,\mathbb{E}[D_i \mid D_i > \tau_m(\delta)/2]/V(\delta) = \rho_m(\delta)\Delta$. The values at the endpoints are immediate, and at δ^*, with $x = \delta^*\sigma_v^2 = (1-\delta^*)\sigma_u^2$, $\tau_m(\delta^*) = x(a/\sigma_v^2 - b/\sigma_u^2)$ and $V(\delta^*) = x^2(a/\sigma_v^2 + b/\sigma_u^2)$. For the monotonicity, differentiate $\tau_m/\sqrt{V}$: its derivative has the sign of $(a+b)V(\delta) - \frac{1}{2}\tau_m(\delta)V'(\delta)$, and with $V'(\delta) = 2\delta a\sigma_v^2 - 2(1-\delta)b\sigma_u^2$ a direct expansion gives

$$(a+b)V(\delta) - \tfrac{1}{2}\tau_m(\delta)V'(\delta) = ab\left(\delta\sigma_v^2 + (1-\delta)\sigma_u^2\right) > 0.$$

(b) A unit of group g is aligned with probability $1-\pi$, in which case it is misclassified on $E_{hg}(\delta)$, and misaligned with probability π, in which case it is misclassified on $\{D_i > \tau_m(\delta)/2\}$; the misalignment draw is independent of D_i, so the probabilities add. The flow from g to h has probability $p(\delta;\pi)$ and its average outcome shock is the mixture $[(1-\pi)p_A(\delta)\rho(\delta) + \pi p_M(\delta)\rho_m(\delta)]\Delta_{hg}/p(\delta;\pi)$, so $p_{h|g}\gamma_{hg} = [(1-\pi)B_A(\delta) + \pi B_M(\delta)]\Delta_{hg}$, and the two flows combine as in the proof of Corollary 1. At $\delta = 1$, the proxy plays no role, $c_m(1) = c(1)$ and $\rho_m(1) = \rho(1)$.

(c) At δ^*, $c'(\delta^*) = 0$ by Theorem 1(b), so the derivative of $p(\delta;\pi)$ reduces to the misaligned term, and $c_m'(\delta^*) > 0$ by part (a). At $\delta = 1$, the same computation as in part (a) for $\tau/\sqrt{V}$ gives $c'(1) = -(b/a)c(1)$, while $c_m'(1) = (b/a)c(1)$, so $\partial p(1;\pi)/\partial\delta = (1-2\pi)(b/a)c(1)\phi(c(1))$. For the bias, write $\rho(\delta) = \delta\sigma_v^2\lambda(c(\delta))/\sqrt{V(\delta)}$. The factor $\delta\sigma_v^2/\sqrt{V(\delta)}$ has zero derivative at $\delta = 1$, because $V(1) = \frac{1}{2}V'(1) = a\sigma_v^2$, so that, using $\sigma_v^2/\sqrt{V(1)} = 1/(2c(1))$,

$$\rho'(1) = \frac{\lambda'(c(1))c'(1)}{2c(1)}, \qquad \rho_m'(1) = \frac{\lambda'(c(1))c_m'(1)}{2c(1)} = -\rho'(1).$$

Hence

$$B_A'(1) = -B_M'(1) = \frac{b}{a}\left[\tfrac{1}{2}\lambda'(c(1))\bar{\Phi}(c(1)) + (1-\rho(1))c(1)\phi(c(1))\right],$$

which is positive when $\rho(1) < 1$, since $\lambda' > 0$, and $\partial B(1;\pi)/\partial\delta = (1-2\pi)B_A'(1)$.

(d) For $\delta < 1$, $c_m(\delta) < c(\delta)$ because $b > 0$, so $p_M(\delta) > p_A(\delta)$ and, λ being increasing, $\rho_m(\delta) < \rho(\delta) \le 1$, which gives $B_M(\delta) > B_A(\delta) \ge 0$; both biases are then nonnegative multiples of $\varkappa$, so the comparison in absolute value is the comparison of the coefficients. Since $B(1;\pi) = B_A(1)$, the inequality $B(\delta;\pi) < B_A(1)$ is $\pi(B_M(\delta) - B_A(\delta)) < B_A(1) - B_A(\delta)$, which is the claim. As $\delta \to 1$, both numerator and denominator of $\bar{\pi}(\delta)$ vanish, and by part (c) their derivatives at one are $-B_A'(1)$ and $-2B_A'(1)$, so the ratio tends to $1/2$. □

Part (a) is the counterpart of Theorem 1(a) and (c) for a unit whose two views disagree. The residual view pulls it towards its outcome group with strength δa and the proxy view towards the other group with strength $(1-\delta)b$, so the unit follows the proxy with probability above one half whenever $\delta < b/(a+b)$ and its outcome group otherwise. At the precision weight, the comparison is between the two separation-to-noise ratios. When the views are equally informative, a misaligned unit is a coin flip at δ^*, and its outcome shock, when it is misclassified, absorbs only a small part of the gap, so the misaligned units behave much like the units misclassified by a proxy-based rule. This is why the tolerance at the precision weight is small,

of the order of the misclassification probability of the GFE rule itself, which is what the two-view rule trades the misaligned units against.

Parts (c) and (d) are the constructive side. Misalignment moves both the accuracy-maximizing and the bias-minimizing weights towards the residual view, and the tolerance $\bar{\pi}(\delta)$ increases with the weight, with limit one half, the point at which the proxy stops carrying group information. A partially valid proxy is therefore still useful, at a weight that leans towards the residuals, and the loss from using δ^* when the proxy is misaligned is the price of the precision weight's optimality under validity. The estimated weight does not correct for misalignment on its own. A misaligned unit kept in its outcome group inflates $\hat{\sigma}_u^2$ and one pulled into the proxy's group inflates $\hat{\sigma}_v^2$, and at δ^* the two effects nearly cancel. The sensitivity analysis over δ that Section 6 reports, and the validity diagnostic, are the tools for this case. Supplemental Appendix S.5 evaluates the proposition in the baseline design of Section 5 and compares it with the misaligned designs of Table S.3.

S.4 Several Regressors

Let $K \geq 2$ and partition $x_i = (x_i^{(1)}, x_i^{(2)})$, where the scalar regressor $x_i^{(1)} = \mu_{g_i}^{(1)} + e_i$ is used as the proxy, $z_i = x_i^{(1)}$, and $x_i^{(2)} = \mu_{g_i}^{(2)} + e_i^{(2)} \in \mathbb{R}^{T\times(K-1)}$ collects the remaining regressors. Write $e_i^{(2)} = e_i\gamma' + w_i$, where $\gamma \in \mathbb{R}^{K-1}$ is the coefficient of the linear projection of the deviations of the other regressors on the deviation of the proxy regressor and w_i is the projection residual.

Proposition S.4 (Regressor-assisted classification with several regressors)**.** *Let $G = 2$, let $e_i \sim N(0, \sigma_u^2 I_T)$ and $v_i \sim N(0, \sigma_v^2 I_T)$ be independent conditional on g_i, let w_i be independent of (e_i, v_i) given g_i, and let Assumption 1 hold for $\hat{g}(\delta)$ with the probabilities and the local misclassification offsets of Theorem 1. Define*

$$m = \sum_{t=1}^{T} (\mu_{2t} - \mu_{1t})(\alpha_{2t} - \alpha_{1t}) \in \mathbb{R}^K,$$

with first component m_1 and remaining components $m_2 \in \mathbb{R}^{K-1}$. Then

$$\operatorname{Bias}\big(\hat{\beta}(\delta)\big) = p(\delta)\, S^{-1} \begin{pmatrix} (1 - \bar{\rho}(\delta) - \bar{\rho}^z(\delta))m_1 \\ (1 - \bar{\rho}(\delta))m_2 - \bar{\rho}^z(\delta)\, m_1\gamma \end{pmatrix}, \tag{S.2}$$

and at the precision weight

$$\operatorname{Bias}\big(\hat{\beta}(\delta^*)\big) = \tfrac{1}{2} p(\delta^*)\, S^{-1} \begin{pmatrix} 0 \\ m_2 - m_1\gamma \end{pmatrix}. \tag{S.3}$$

Proof of Proposition S.4. The misclassification event depends on (v_i, e_i) only, so by Theorem 1(c) the units misclassified from b to a have shock offset $\bar{\rho}(\delta)(\alpha_a - \alpha_b)$ and proxy-regressor shift $\bar{\rho}^z(\delta)(\mu_a^{(1)} - \mu_b^{(1)})$ in the local misclassification limit, while the shift of the other regressors is $\mathbb{E}[e_i^{(2)} \mid E_{ab}] = \mathbb{E}[e_i \mid E_{ab}]\gamma' + \mathbb{E}[w_i \mid E_{ab}] = \bar{\rho}^z(\delta)(\mu_a^{(1)} - \mu_b^{(1)})\gamma'$, since w_i is independent of the event. Hence $s_{ab} = \bar{\rho}^z(\delta)(\mu_a^{(1)} - \mu_b^{(1)})(1, \gamma')$, and substituting in (12) and summing the two symmetric flows as in the proof of Corollary 1 gives (S.2). At δ^*, $\bar{\rho} = \bar{\rho}^z = 1/2$ by (16). □

The first component of the numerator in (S.3) is zero, so the bias generated by the regressor that carries the group signal is removed, and the other components are halved, net of the part of their contrast

that the proxy regressor's noise explains. Because S^{-1} is not diagonal in general, the bias of $\hat{\beta}_1$ itself is $\frac{1}{2}p(\delta^*)[S^{-1}]_{12}(m_2 - m_1\gamma)$, which vanishes when the regressors are uncorrelated within groups or when $m_2 = m_1\gamma$. The condition $m_2 = m_1\gamma$ holds when the group-mean contrasts of the other regressors are the γ-multiples of those of the proxy regressor, $\mu_h^{(2)} - \mu_g^{(2)} = (\mu_h^{(1)} - \mu_g^{(1)})\gamma'$ for every pair, where γ is the projection coefficient of the deviations $e_i^{(2)}$ on e_i. A shared group structure up to a scale that differs from γ does not deliver it. The condition is satisfied when a single factor drives both the group means and the within-group deviations of all regressors with common loadings, which is a special case of the factor structure of Chetverikov and Manresa (2022), and then the first-order bias of the whole vector vanishes at the precision weight.

S.5 Further Simulation Results

Table S.1 reports the design of Section 5.3, in which the regressor itself is the proxy.

Table S.1: The proxy as an included regressor

	Misclassification		Offsets		$\hat{\beta}$				
	MC	Theory	$\hat{\rho}_y$	$\hat{\rho}_x$	Bias	SD	RMSE	Coverage	$\hat{\delta}$
				$\xi = 0.5$					
Oracle	–	–	–	–	0.000	0.051	0.051	0.94	
GFE	0.375	0.236	0.42	-0.02	0.140	0.063	0.153	0.15	
Proxy	0.029	0.028	-0.01	0.58	0.013	0.051	0.053	0.94	
CAS	0.022	0.020	0.56	0.57	-0.004	0.051	0.052	0.94	0.36
Combined	–	–	–	–	0.267	0.099	0.284	0.03	
				$\xi = 1$					
Oracle	–	–	–	–	-0.001	0.050	0.050	0.94	
GFE	0.107	0.075	0.58	-0.02	0.082	0.070	0.108	0.54	
Proxy	0.030	0.028	0.00	0.58	0.025	0.052	0.058	0.92	
CAS	0.009	0.008	0.56	0.55	-0.003	0.050	0.050	0.94	0.36
Combined	–	–	–	–	0.139	0.107	0.175	0.32	
				$\xi = 1.5$					
Oracle	–	–	–	–	-0.002	0.050	0.050	0.95	
GFE	0.019	0.016	0.56	-0.02	0.024	0.059	0.064	0.88	
Proxy	0.029	0.028	0.00	0.58	0.037	0.055	0.066	0.88	
CAS	0.002	0.002	0.54	0.53	-0.002	0.050	0.050	0.95	0.36
Combined	–	–	–	–	0.012	0.082	0.083	0.79	

Note: Baseline design with $z_i = x_i$. $\hat{\rho}_y$ and $\hat{\rho}_x$ are the realized offsets of the outcome shocks and of the regressor noise of the misclassified units, relative to the respective gaps. Other entries as in Table 1, with $\delta^* = 0.36$.

Table S.2 varies the group structure of the proxy relative to the outcome equation in the baseline design of Section 5.1 with $\xi = 1$. In the three-group designs, the third profile and the third proxy center are placed at the same pairwise distance from the other two as the baseline pair, so that every pair of groups is as separated as in the two-group baseline. In the coarser design, model groups 2 and 3 share a proxy center. In the finer designs, each of the two model groups is split at random into two subgroups with proxy centers at ζ_1, $\zeta_1 + 0.5d$ and $\zeta_1 + d$, $\zeta_1 + 1.5d$, with $d = 1.085$ the baseline proxy gap, and the model is estimated either with the two model groups or with the four proxy groups.

Table S.2: Finer and coarser group structure in the proxy

	Misclassification			Bias			RMSE			
	GFE	Proxy	CAS	GFE	Proxy	CAS	GFE	Proxy	CAS	$\hat{\delta}$
Design 1	0.281	0.148	0.053	0.213	0.210	0.045	0.240	0.218	0.077	0.20
Design 2	0.291	0.375	0.139	0.220	0.319	0.119	0.246	0.322	0.148	0.20
Design 3	0.108	0.134	0.045	0.084	0.192	0.046	0.110	0.201	0.078	0.21
Design 4	0.156	0.165	0.066	0.132	0.198	0.062	0.156	0.206	0.091	0.20
Oracle ($G = 3$)				-0.003			0.052			
Oracle ($G = 2$)				0.000			0.051			

Note: Baseline design with $\xi = 1$, $N = 100$, $T = 7$, 2,000 replications. Design 1 has three groups and a valid proxy, estimated with $G = 3$ groups. Design 2 has a coarser proxy, with two centers for the three groups, and is estimated with $G = 3$ groups. Design 3 has a finer proxy, with four centers for two groups, estimated with $G = 2$ groups. Design 4 has a finer proxy and is estimated with $G = 4$ groups. In Design 4, the misclassification rate maps the four estimated labels onto the two true groups in the most favorable way. The last two rows are the oracle estimators with the true groups.

Table S.3 reports the baseline design under several departures from the validity of the proxy, and under an overspecified number of groups. In the misaligned designs, a fraction π of the units, drawn at random, is assigned to the other group in the proxy equation only. In the design without group structure, the two proxy centers coincide. In the finer-structure design of this table, half of the units of the second group have a third proxy center $\zeta_3 = \zeta_1 + 2.17$, while the outcome equation keeps two groups and the model is estimated with $G = 2$. In the last design, the proxy is valid and the model is estimated with $G = 3$ groups, and the misclassification rate is computed by mapping the three estimated labels onto the two true ones in the most favorable way.

Table S.3: Invalid proxies and overspecified number of groups

	Misclassification			Bias			RMSE			
	GFE	Proxy	CAS	GFE	Proxy	CAS	GFE	Proxy	CAS	$\hat{\delta}$
Valid proxy	0.107	0.083	0.026	0.082	0.135	0.023	0.109	0.148	0.061	0.20
Misaligned, $\pi = 0.05$	0.109	0.126	0.056	0.085	0.186	0.068	0.110	0.194	0.094	0.20
Misaligned, $\pi = 0.15$	0.108	0.208	0.132	0.086	0.259	0.164	0.112	0.264	0.181	0.19
Misaligned, $\pi = 0.30$	0.110	0.332	0.281	0.085	0.318	0.279	0.110	0.321	0.290	0.18
No group structure	0.107	0.461	0.126	0.083	0.345	0.103	0.109	0.347	0.132	0.20
Finer structure (3 groups)	0.109	0.101	0.034	0.085	0.149	0.041	0.112	0.162	0.077	0.22
Overspecified $G = 3$	0.145	0.106	0.037	0.115	0.150	0.032	0.139	0.162	0.068	0.20
Oracle				-0.001			0.051			

Note: Baseline design with $\xi = 1$, $N = 100$, $T = 7$, 2,000 replications, with the same draws as Table 1 in the first row. In the misaligned designs, a fraction π of the units, drawn at random, belongs to the other group in the proxy equation. The proxy without group structure has a single center. The finer proxy splits the second group at random into two subgroups of equal size in expectation, one with the original center and one with a center a further distance d away, estimated with $G = 2$. The overspecified design estimates the baseline design with $G = 3$ groups, mapping the three estimated labels onto the two true groups in the most favorable way. Bias and RMSE of $\hat{\beta}$ relative to $\beta = 1$. The final row is the oracle estimator with the true groups.

The precision weight is close to 0.2 in every design of Table S.3. It measures the noise of the proxy and not its validity, so an invalid proxy receives the same weight as a valid one. With 5% of misaligned units, the CAS estimator still improves on GFE in bias and RMSE, with 15% it is worse, and with 30% it is about as bad as the proxy-only rule. Proposition S.3 in Supplemental Appendix S.3 accounts for these numbers. In the baseline design the two views are equally informative, so a misaligned unit is a coin flip at δ^* with an offset $\rho_m(\delta^*) = 0.20$ against 0.59 for an aligned unit, and the tolerance of the precision weight is

$\bar{\pi}(\delta^*) = 0.044$. The CAS estimator still beats GFE at $\pi = 0.05$ because the estimated centers raise the GFE misclassification rate above its population value. The formula $B(\delta^*;\pi)\varkappa$ with $\varkappa = 2.10$ gives 0.018, 0.059, 0.142 and 0.265 for $\pi = 0$, 0.05, 0.15 and 0.30, against realized biases of 0.023, 0.068, 0.164 and 0.279. The tolerance increases with the weight, from 0.044 at δ^* to 0.25 at $\delta = 0.5$ and 0.43 at $\delta = 0.8$, and the bias-minimizing weight, computed with the exact offsets, is 0.28 under validity, 0.43 at $\pi = 0.05$, 0.56 at $\pi = 0.15$ and 0.73 at $\pi = 0.30$, with a bias at that weight of 0.0073, 0.0136, 0.0196 and 0.0241 in the units of B, below the GFE value of 0.0259 in every case. The estimated weight barely reacts to the misalignment, moving from 0.20 to 0.19 at $\pi = 0.15$ and 0.18 at $\pi = 0.30$, because a misaligned unit kept in its outcome group inflates $\hat{\sigma}_u^2$ and one pulled into the proxy's group inflates $\hat{\sigma}_v^2$, and at δ^* the two effects nearly cancel. A proxy without group structure leads the k-means solution to fit idiosyncratic noise in the proxy dimensions, and the two-view rule misclassifies slightly more units than GFE, with a correspondingly larger bias. Overspecifying G affects the CAS estimator little, whereas the GFE estimator deteriorates in this short panel. The asymptotic robustness of GFE to overspecification requires the extra groups to fit noise rather than structure in the covariates, which is not yet the case with $T = 7$, while the proxy pins down the partition.

Table S.4 reports the size and power of the validity diagnostic of Section 2.4, Table S.5 the coverage and length of the bias-aware intervals of Remark 2, and Table S.6 the designs with non-Gaussian, heteroskedastic and serially correlated errors. Section 5.3 discusses them.

Table S.4: Size and power of the validity diagnostic

	Statistic			Rejection rate		Misclassification		
	LR	LR_y	LR_z	5%	10%	GFE	Proxy	CAS
Valid proxy	39.8	21.5	18.4	0.051	0.105	0.107	0.083	0.026
Misaligned, $\pi = 0.05$	62.3	34.8	27.6	0.449	0.592	0.108	0.124	0.056
Misaligned, $\pi = 0.15$	99.1	66.4	32.7	0.930	0.966	0.108	0.208	0.130
Misaligned, $\pi = 0.30$	132.8	115.6	17.2	0.988	0.995	0.108	0.332	0.282
No group structure	97.7	6.2	91.5	0.180	0.326	0.107	0.459	0.128
Valid proxy, $\xi = 0.5$	74.0	65.0	9.0	0.071	0.168	0.377	0.082	0.064

Note: Baseline design with $\xi = 1$ (and $\xi = 0.5$ in the last row), $N = 100$, $T = 7$, 2,000 replications, $B = 199$ parametric bootstrap draws per replication, with 10 random starts per bootstrap fit, the first from the fitted partition. LR is the statistic (10) and LR_y and LR_z its two terms (averages over replications). Rejection rates at the 5% and 10% levels. The misclassification columns are the rates of the three partitions that enter the statistic.

S.6 Minimum Wage and County Employment: Data and Further Results

Data. County employment is the private-ownership annual average employment of the Quarterly Census of Employment and Wages (Bureau of Labor Statistics), from the open-data API for 2014–2019 and from the annual single files for 2012–2013. The outcome is food services and drinking places (NAICS 722, aggregation level 75) and the proxy is manufacturing (NAICS 31–33, level 74). Cells suppressed for confidentiality are treated as missing, and counties with a missing or zero value in any year in either series are dropped, which leaves 1,976 of the 3,144 counties. The minimum wage is the annual average of the state minimum wage of Vaghul and Zipperer (2016) (release 1.4.0), floored at the federal minimum. Sub-state minimum wages are not used. Population is the Census Bureau vintage-2019 county estimate. All variables enter in logs and are demeaned by county, so that (25) has county and grouped time effects, and the proxy is demeaned in the same way. Standard errors clustered by state are computed from the within-$(\hat{g}, t)$ least squares residuals

Table S.5: Coverage and length of 90% bias-aware intervals

		Misclassification		Standard		Sym., $\varsigma = 1$		Sym., $\varsigma = 2$		Rec., $\varsigma = 2$		Sym., $\varsigma = 2$, ϖ		Rec., $\varsigma = 2$, ϖ	
		Bound	Rate	Cov.	Len.	Cov.	Len.	Cov.	Len.	Cov.	Len.	Cov.	Len.	Cov.	Len.
$\xi = 0.5$	GFE	0.169	0.374	0.118	0.124	0.153	0.134	0.209	0.153	0.208	0.134	0.428	0.297	0.412	0.200
	CAS	0.049	0.065	0.790	0.155	0.810	0.164	0.851	0.186	0.837	0.164	0.887	0.208	0.866	0.172
$\xi = 1$	GFE	0.085	0.107	0.455	0.148	0.624	0.205	0.802	0.294	0.790	0.205	0.956	0.453	0.942	0.284
	CAS	0.022	0.026	0.810	0.164	0.837	0.172	0.887	0.194	0.877	0.172	0.901	0.203	0.888	0.175
$\xi = 1.5$	GFE	0.019	0.020	0.795	0.166	0.830	0.180	0.914	0.212	0.892	0.180	0.936	0.224	0.909	0.185
	CAS	0.006	0.006	0.869	0.166	0.874	0.167	0.883	0.171	0.891	0.167	0.885	0.172	0.891	0.167

Note: Baseline design, $N = 100$, $T = 7$, 2,000 replications. "Bound" is the average over replications of the largest calibrated bound on the outgoing misclassification probabilities of a group (safety factor one) and "Rate" the actual misclassification rate. Standard interval based on the large-T clustered standard errors, and symmetric and recentered bias-aware intervals of González-Casasús (2026*a*) with calibrated bounds on the outgoing misclassification probabilities, offset one half and safety factor ς; the last four columns inflate the bias set by $(1 - \varpi\bar{p})^{-2}$ with $\varpi = 1.5$ and $\bar{p}$ the largest calibrated bound. For CAS, the calibration uses the stacked distances (Remark 2). Same draws as Table 1.

Table S.6: Non-Gaussian, heteroskedastic and serially correlated errors

	Misclassification			Bias			RMSE			Coverage		
	GFE	Proxy	CAS	GFE	Proxy	CAS	GFE	Proxy	CAS	GFE	CAS	$\hat{\delta}$
Gaussian	0.107	0.083	0.026	0.082	0.135	0.023	0.109	0.148	0.061	0.55	0.89	0.20
$t(3)$, both views	0.126	0.090	0.056	0.088	0.122	0.040	0.150	0.149	0.110	0.65	0.83	0.20
$t(3)$, outcome	0.124	0.082	0.042	0.088	0.134	0.028	0.149	0.150	0.091	0.65	0.87	0.22
AR(1) hetero., Euclidean	0.252	0.216	0.142	0.143	0.257	0.080	0.155	0.266	0.106	0.14	0.59	0.21
AR(1) hetero., whitened	0.212	0.133	0.115	0.028	0.042	0.017	0.044	0.055	0.039	0.79	0.86	0.50
Oracle				-0.001			0.051			0.94		

Note: Baseline design with $\xi = 1$, $N = 100$, $T = 7$, 2,000 replications. The $t(3)$ errors are scaled to the baseline variances. In the heteroskedastic designs, the standard deviation of the errors rises linearly across periods, from 0.47 to 1.42 times the baseline after normalizing the average variance, and the errors follow an AR(1) within unit with coefficient 0.5, in both views. The whitened rule premultiplies each view by the inverse Cholesky factor of its within-group covariance matrix estimated from a first CAS fit (Remark 4). Coverage refers to the nominal 95% interval based on the large-T clustered standard errors. The oracle row refers to the Gaussian design.

with the estimated groups treated as known.

Auxiliary results. Table S.7 adds the proxy-only estimates and the population coefficient to Table 2. Classification on manufacturing employment alone gives an elasticity of -0.03 at every G, with a standard error of 0.05, that is, the two-way fixed-effects estimate, because a partition that ignores the restaurant residuals leaves the state-level variation of the minimum wage untouched by the group effects. The population coefficient lies between 0.6 and 1.1 across estimators and G. Table S.8 reports the diagnostic and Table S.9 the bias-aware intervals, whose recentered versions behave as the symmetric ones in Table 2. The versions with the finite-sample inflation $\varpi = 1.5$ widen the CAS intervals moderately for $G \leq 7$, to $[-0.141, 0.006]$ at $G = 4$, and become uninformative beyond. For GFE, they are uninformative from $G = 4$ on, at $G = 4$ to 6 because of the sensitivities, with a factor $(1 - \varpi\bar{p})^{-2}$ between 2.5 and 5.5, and from $G = 7$ on because the factor exceeds 50 when the calibrated bound with the safety factor approaches 0.6. The inflation is a device for small bounds, and the large bounds of the GFE partition in this panel are themselves the finding.

Table S.7: Minimum wage: GFE, proxy-only and CAS estimates

	GFE		Proxy		CAS			Diff. vs.	
G	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\delta}$	GFE	Proxy
1	-0.032	1.082							
	(0.049)	(0.054)							
2	-0.095	0.769	-0.030	1.062	-0.036	0.638	0.82	0.14	0.34
	(0.021)	(0.037)	(0.048)	(0.056)	(0.025)	(0.043)			
3	-0.050	0.758	-0.034	1.057	-0.050	0.747	0.80	0.40	0.41
	(0.013)	(0.032)	(0.049)	(0.057)	(0.023)	(0.045)			
4	-0.064	0.992	-0.031	1.054	-0.068	0.904	0.76	0.36	0.50
	(0.010)	(0.028)	(0.049)	(0.056)	(0.025)	(0.039)			
5	-0.020	0.835	-0.032	1.056	-0.043	0.909	0.79	0.44	0.51
	(0.011)	(0.022)	(0.049)	(0.058)	(0.015)	(0.031)			
6	-0.050	0.751	-0.031	1.055	-0.032	0.961	0.77	0.49	0.41
	(0.006)	(0.018)	(0.049)	(0.057)	(0.018)	(0.036)			
7	-0.064	0.826	-0.031	1.048	-0.021	0.949	0.75	0.55	0.42
	(0.006)	(0.021)	(0.049)	(0.056)	(0.017)	(0.032)			
8	-0.073	0.820	-0.031	1.039	-0.045	1.023	0.76	0.53	0.49
	(0.007)	(0.023)	(0.049)	(0.056)	(0.012)	(0.031)			
9	-0.069	0.857	-0.033	1.046	-0.032	1.066	0.74	0.57	0.58
	(0.007)	(0.016)	(0.049)	(0.058)	(0.014)	(0.039)			
10	-0.073	0.837	-0.029	1.045	-0.080	0.963	0.73	0.62	0.53
	(0.008)	(0.017)	(0.049)	(0.059)	(0.019)	(0.032)			

Note: Coefficients of log minimum wage ($\hat{\beta}_1$) and log population ($\hat{\beta}_2$) in (25) for the GFE ($\delta = 1$), proxy-only ($\delta = 0$) and CAS (QML weight) estimators, $G = 1$ (two-way fixed-effects) to 10. Standard errors clustered by state in parentheses. "Diff." is the share of counties classified differently by CAS and by GFE or the proxy-only rule.

Sample. The sample conditions on disclosure in every year of both the restaurant and the manufacturing series, which drops a third of the counties, the smallest ones. On the 2,158 counties with a complete restaurant series, the two-way fixed-effects elasticity is -0.023 (state-clustered standard error 0.05), against -0.032 in the estimation sample. Within the estimation sample, the point estimates place the elasticity in the small counties. When the sample is split by the disclosure of a third, high-wage sector (professional, scientific and technical services, NAICS 54), which selects the 1,246 larger counties with a median population of 81,000 against 29,000 for the other 730, the two-way fixed-effects elasticity is -0.004 (state-clustered standard error 0.048) in the larger counties and -0.096 (0.059) in the smaller ones, a difference of 0.09 with a standard error of 0.08, so the split is suggestive rather than conclusive.

S.7 Income and Democracy: Further Results

This appendix collects further results for the income and democracy application of Section 6.2. It reports the estimates of Table 3 as a function of G (Figure S.1) and as a function of the weight at $G = 4$ (Figure S.2), the estimated groups at $G = 4$ (Figures S.3 and S.4), the information criteria for the number of groups in both applications (Table S.10), the validity diagnostic of Section 2.4 for the savings rate (Table S.11), the whitened two-view rule (Table S.12), the median age of the population and lagged income as alternative proxies (Tables S.13 and S.14), the bias-aware intervals of Remark 2 (Table S.15), and the relation between the CAS estimates and the local analysis of the GFE estimates in González-Casasús (2026*a*).

Estimated groups. Figure S.3 shows the estimated time effects, the group averages of democracy and income, and the estimated savings-rate centers for $G = 4$, with the groups numbered in increasing order of

Table S.8: Minimum wage: validity diagnostic

G	LR	LR_y	LR_z	p-value	Diff. GFE/Proxy
2	4300.8	955.9	3344.9	0.990	0.46
3	7063.2	3303.2	3760.0	0.120	0.48
4	7863.5	5262.0	2601.5	0.195	0.56
5	8989.5	4457.1	4532.4	0.085	0.63
6	9805.4	5559.8	4245.6	0.030	0.65
7	10282.0	6558.7	3723.2	0.075	0.70
8	10764.4	5981.3	4783.1	0.040	0.69
9	11186.6	7065.3	4121.3	0.120	0.72
10	11914.8	7641.0	4273.8	0.035	0.74

Note: Statistic (10) and its two terms, with the bootstrap p-value based on $B = 199$ draws in which the within-group residuals of the fitted two-view model are resampled by county, with 10 random starts, the first from the fitted partition, and 10 variable neighborhood search steps per bootstrap fit; with 100 starts and 100 steps the p-values are 0.995, 0.220, 0.240, 0.170, 0.050, 0.075, 0.025, 0.125 and 0.030. The last column is the share of counties classified differently by the GFE and the proxy-only partitions.

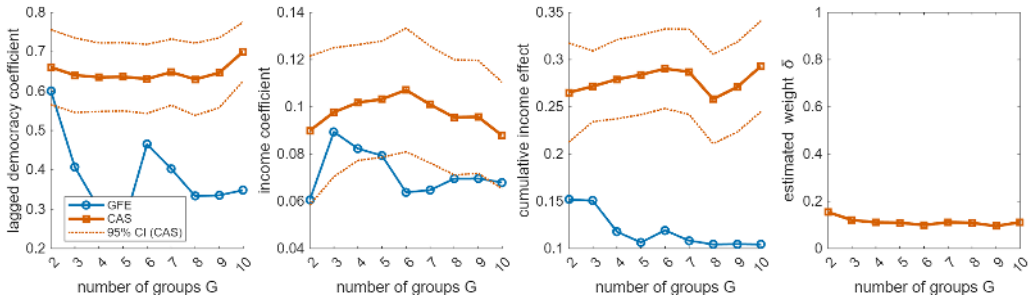


Figure S.1: Coefficients of lagged democracy and income by number of groups

Note: Point estimates of the GFE estimator (blue) and of the CAS estimator (red) with 95% confidence intervals for the CAS estimator based on large-T clustered standard errors (dotted), and the estimated weight $\hat{\delta}$ (right panel).

their average democracy index, as in the map of Figure S.4. The savings-rate centers are well separated and nearly flat, at about 0.03, 0.34, 0.16 and 0.24, which is what drives the partition. Group 1, with 17 countries, mostly in sub-Saharan Africa, has a savings rate close to zero, the lowest income and a democracy index of about 0.2 until 1990 that rises to 0.5 in the 1990s. Group 2, with 11 countries, among them China, Japan, Singapore, Norway and several oil exporters, has the highest savings rate, an income level close to the top group and a democracy index that stays around 0.5 and declines at the end of the sample. Groups 3 and 4, with 28 and 34 countries, are separated by their savings rates, 0.16 against 0.24, and both mix countries at different levels of development. The United States, the United Kingdom, Argentina, India and several East African countries are in group 3, and Canada, France, Spain, Ireland, Denmark, Sweden, Finland, Australia, Korea, Brazil, Mexico, South Africa and Nigeria are in group 4. Group 4 has the higher average income and a democracy index that rises steadily from 0.57 to 0.81, and group 3 an index that rises from 0.5 to 0.65 in the 1980s. The time effects of the four groups move within a band of 0.13 until 1990, apart from 1985, when group 3 peaks and the band widens to 0.21, and diverge afterwards, upward for group 1 and downward for the high-savings group. In contrast with the partition of Bonhomme and Manresa (2015), which separates early and late democratic transitions, the partition here is a classification by savings behavior, correlated with the level of development at the bottom and at the top of the distribution but not in the middle, and the democratic transitions of the 1990s appear within group 1 rather than as a group of their own.

Table S.9: Minimum wage: bias-aware 90% intervals for the elasticity

G		Standard	Symmetric	Recentered	Symmetric, $\varpi = 1.5$	Bound
2	GFE	[-0.128, -0.061]	[-0.143, -0.047]	[-0.144, -0.067]	[-0.163, -0.026]	0.095
	CAS	[-0.078, 0.006]	[-0.078, 0.006]	[-0.079, 0.005]	[-0.078, 0.006]	0.083
3	GFE	[-0.072, -0.028]	[-0.080, -0.020]	[-0.081, -0.032]	[-0.100, 0.000]	0.126
	CAS	[-0.088, -0.013]	[-0.092, -0.009]	[-0.094, -0.016]	[-0.102, 0.002]	0.103
4	GFE	[-0.081, -0.047]	[-0.132, 0.005]	[-0.107, 0.005]	[-0.216, 0.088]	0.123
	CAS	[-0.109, -0.027]	[-0.120, -0.016]	[-0.121, -0.031]	[-0.141, 0.006]	0.104
5	GFE	[-0.038, -0.002]	[-0.485, 0.444]	[-0.059, 0.444]	[-1.301, 1.260]	0.135
	CAS	[-0.067, -0.018]	[-0.078, -0.007]	[-0.078, -0.013]	[-0.114, 0.028]	0.144
6	GFE	[-0.060, -0.039]	[-0.678, 0.578]	[-0.126, 0.578]	[-3.478, 3.378]	0.191
	CAS	[-0.061, -0.002]	[-0.079, 0.015]	[-0.077, 0.015]	[-0.131, 0.067]	0.146
7	GFE	[-0.074, -0.054]	[-0.718, 0.590]	[-0.141, 0.590]	[-45.920, 45.792]	0.294
	CAS	[-0.049, 0.006]	[-0.067, 0.024]	[-0.062, 0.024]	[-0.120, 0.078]	0.148
8	GFE	[-0.084, -0.062]	[-0.643, 0.496]	[-0.155, 0.496]	[-28.193, 28.046]	0.286
	CAS	[-0.064, -0.025]	[-0.142, 0.053]	[-0.107, 0.053]	[-0.347, 0.258]	0.155
9	GFE	[-0.081, -0.057]	[-0.625, 0.486]	[-0.177, 0.486]	[-43.134, 42.996]	0.296
	CAS	[-0.055, -0.009]	[-0.213, 0.149]	[-0.105, 0.149]	[-0.680, 0.616]	0.164
10	GFE	[-0.086, -0.060]	[-0.704, 0.559]	[-0.203, 0.559]	[-72.775, 72.630]	0.303
	CAS	[-0.112, -0.049]	[-0.256, 0.095]	[-0.187, 0.095]	[-0.954, 0.793]	0.193

Note: Standard intervals based on the state-clustered standard errors, and symmetric and recentered bias-aware intervals of González-Casasús (2026*a*) with calibrated bounds on the outgoing misclassification probabilities, offset one half and safety factor two, without and with the finite-sample inflation $\varpi = 1.5$. "Bound" is the largest calibrated outgoing probability across groups (safety factor one). For GFE at $G \geq 7$, the inflation factor $(1 - \varpi\bar{p})^{-2}$ exceeds 50, because $\varpi\bar{p}$ is above 0.85, and the inflated intervals are reported for completeness only.

Number of groups. Table S.10 reports two information criteria for the number of groups in each application. For the GFE fit, the criterion is that of Bonhomme and Manresa (2015), $\hat{\sigma}_v^2(G)/\hat{\sigma}_v^2(10) + (GT + N + K)\ln(NT)/(NT)$, their residual sum of squares plus penalty divided by the residual variance at the largest number of groups, which leaves the minimizer unchanged. For the CAS fit, the criterion is the Gaussian quasi-likelihood (6) with the two variances fixed at their values at $G = 10$, $[NT\hat{\sigma}_v^2(G)/\hat{\sigma}_v^2(10) + NL\hat{\sigma}_u^2(G)/\hat{\sigma}_u^2(10) + (G(T+L)+N+K)\ln(N(T+L))]/(N(T+L))$, so that the two views enter with the same normalization as in the estimator. The fixed variances make the criterion a penalized sum of squares that compares fits on one scale, as in Bonhomme and Manresa (2015), at the cost of normalizing by the most noise-fitting variances. The "profiled" versions replace the fixed variances by the fitted ones, $NT\ln\hat{\sigma}_v^2(G)$ and $NT\ln\hat{\sigma}_v^2(G) + NL\ln\hat{\sigma}_u^2(G)$, with the same penalties. In the income and democracy panel, the GFE criterion selects $G = 10$ in both versions, whereas the CAS criterion selects $G = 5$ with fixed variances and $G = 4$ with profiled variances, with values within 0.02 of the minimum for $G = 4$ to 7 in the first version. In the county panel all four criteria decrease monotonically up to $G = 10$.

Validity of the proxy. The disagreement between the views raises the question of Section 2.4, namely whether the savings rate and the democracy residuals share one partition at all. Table S.11 reports the statistic (10) and its bootstrap p-value for $G = 2, \ldots, 10$, with the within-group residuals of the fitted model resampled by country, which preserves their serial dependence. The statistic is dominated by the outcome term, since the proxy-only partition is close to the CAS partition, and it grows with G, from 111 at $G = 2$ to 575 at $G = 10$, because the GFE partition fits more of the idiosyncratic variation as the groups become smaller. The bootstrap reproduces this noise fitting under the maintained model, and the p-values are 0.28 and 0.30 at $G = 2$ and 3, 0.075 and 0.09 at $G = 4$ and 5, 0.035 at $G = 6$ and at most 0.015 from $G = 7$ on. The bootstrap fits use a tenth of the optimization budget of the sample fits, and raising

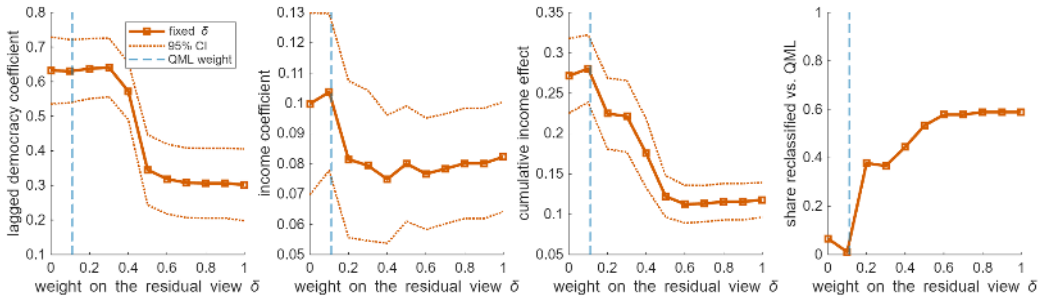


Figure S.2: Estimates as a function of the weight, $G = 4$

Note: CAS estimator with the weight on the residual view fixed at $\delta \in \{0, 0.1, \ldots, 1\}$, with 95% confidence intervals based on large-T clustered standard errors (dotted). The weight $\delta = 0$ classifies on the savings rate only and $\delta = 1$ is the GFE estimator. The vertical dashed line is the QML weight. The right panel is the share of countries classified differently from the QML solution.

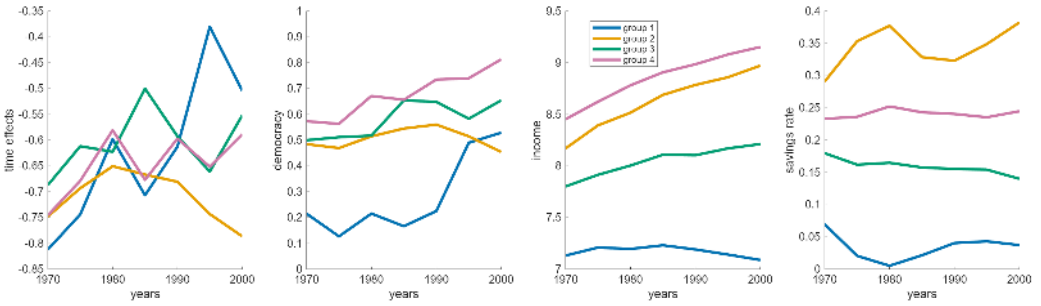


Figure S.3: Patterns of heterogeneity, $G = 4$

Note: Covariate-assisted estimates with $G = 4$. From left to right, estimated time effects $\hat{\alpha}_{gt}$, group averages of the democracy index and of lagged log income per capita, and estimated savings-rate centers $\hat{\zeta}_g$.

it tenfold moves the p-values within bootstrap noise, to 0.065 at $G = 6$ and at most 0.02 from $G = 7$ on, so the rejections are not an artifact of under-optimized bootstrap fits. The shared partition is therefore not rejected at the number of groups used in the main analysis, although the evidence is not strong, and it is rejected when the number of groups is large relative to the sample, where both single-view partitions fit noise and the loss of fit from imposing a common one exceeds what the bootstrap generates. Three qualifications apply. First, the calibration matters. With Gaussian draws, the p-value at $G = 4$ is 0.005, because independent draws understate the serial dependence of the residuals that the GFE rule fits within countries, so the bootstrap statistic is too small. The residual draws are the relevant ones for a bounded index with persistent shocks, but the difference shows that the diagnostic is sensitive to the bootstrap design in a panel of seven periods. Second, the bootstrap holds lagged democracy fixed at its sample values, as the fitted model treats it as a regressor. A recursive design that regenerates the lagged outcome from the fitted dynamics is the alternative, and it is not pursued here. Third, the power of the diagnostic is limited at this sample size, as the simulations with $N = 100$ and $T = 7$ indicate. In those simulations, a misalignment of 5% of the units is detected less than half of the time. The median age as a second proxy, discussed below, makes the limitation concrete, since it is not rejected either at $G = 4$ ($p = 0.25$) although its partition differs from that of the savings rate for 60% of the countries, so that at most one of the two can share the partition of the residual view. The diagnostic is thus a check that the savings rate passes at $G = 4$ rather than a confirmation of its validity, which remains the maintained assumption.

Table S.10: Information criteria for the number of groups

	Income and democracy				Minimum wage			
	Fixed variances		Profiled		Fixed variances		Profiled	
G	GFE	CAS	GFE	CAS	GFE	CAS	GFE	CAS
2	3.641	2.581	−2.373	−3.457	3.133	2.365	−4.529	−4.259
3	3.294	2.309	−2.480	−3.565	2.847	2.177	−4.689	−4.355
4	3.072	2.246	−2.556	**−3.581**	2.682	2.018	−4.795	−4.450
5	2.921	**2.226**	−2.613	−3.575	2.554	1.948	−4.886	−4.511
6	2.805	2.231	−2.665	−3.556	2.474	1.873	−4.948	−4.563
7	2.738	2.234	−2.695	−3.538	2.403	1.813	−5.006	−4.612
8	2.706	2.264	−2.707	−3.503	2.346	1.772	−5.057	−4.651
9	2.671	2.286	−2.728	−3.481	2.296	1.728	−5.103	−4.689
10	**2.657**	2.314	**−2.739**	−3.447	**2.259**	**1.701**	**−5.139**	**−4.715**

Note: Criteria defined in the text, for the GFE fit ($\delta = 1$) and the CAS fit (QML weight), from the estimates of Tables 3 and S.7. The minimum of each column is in bold. In the minimum wage panel, $N = 1{,}976$ and $T = L = 8$.

Table S.11: Validity diagnostic for the savings-rate proxy

G	LR	LR_y	LR_z	p-value	Disagreement GFE/Proxy
2	111.1	106.9	4.2	0.275	0.36
3	194.1	192.4	1.8	0.295	0.57
4	270.1	255.6	14.5	0.075	0.59
5	324.6	295.4	29.2	0.090	0.64
6	389.3	367.8	21.5	0.035	0.67
7	431.9	352.7	79.1	0.015	0.66
8	485.3	386.5	98.8	0.010	0.68
9	528.2	455.9	72.3	0.010	0.68
10	575.1	423.3	151.7	0.005	0.71

Note: Statistic (10) and its two terms, with the bootstrap p-value based on $B = 199$ draws in which the within-group residuals of the fitted two-view model are resampled by country, with 50 random starts, the first from the fitted partition, and 50 variable neighborhood search steps per bootstrap fit; with 500 starts and 250 steps the p-values are 0.385, 0.320, 0.095, 0.080, 0.065, 0.020, 0.005, 0.015 and 0.005. The last column is the share of countries classified differently by the GFE and the proxy-only partitions. For $G = 4$, the p-value with Gaussian draws is 0.005, and the diagnostic for the median age as the only proxy, with residual draws, gives $\text{LR} = 282.0$ ($\text{LR}_y = 282.0$, $\text{LR}_z = 0.0$) and $p = 0.245$.

Serial dependence. Section 5.3 recommends the whitened rule of Remark 4 when the errors are not approximately spherical within unit, and the difference between the Gaussian and the residual bootstrap in the validity diagnostic indicates that the residuals of this specification are persistent. Table S.12 reports the CAS estimates when both views are premultiplied by the inverse Cholesky factors of their within-group covariance matrices, estimated from the Euclidean fit, together with the GFE estimates on the whitened outcome. At $G = 4$, the whitened partition differs from the Euclidean one for 3% of the countries and the estimates are 0.65, 0.10 and 0.27, so the results of Table 3 do not depend on the Euclidean metric. For $G \geq 6$, the whitened partition departs more from the Euclidean one, for 28% to 46% of the countries, and the persistence estimate moves to between 0.53 and 0.62, above the whitened GFE estimates of 0.32 to 0.50, by a wide margin except at $G = 9$ and 10. On the whitened scales, the two views have unit within-group variances by construction, so the estimated weight is between 0.50 and 0.63.

Alternative proxies. The data set of Acemoglu et al. (2008) contains the median age of the population, a marker of the demographic transition, which the political demography literature links to transitions

Table S.12: Whitened two-view rule

	Whitened CAS				Diff. vs.		Whitened GFE
G	$\hat{\beta}_1$	$\hat{\beta}_2$	Cumulative	$\hat{\delta}$	CAS	GFE	$\hat{\beta}_1$
2	0.664	0.093	0.278	0.51	0.10	0.37	0.538
	(0.040)	(0.013)	(0.022)				
3	0.553	0.099	0.221	0.55	0.26	0.37	0.470
	(0.039)	(0.012)	(0.018)				
4	0.652	0.095	0.272	0.50	0.03	0.56	0.432
	(0.040)	(0.012)	(0.022)				
5	0.625	0.098	0.261	0.50	0.14	0.57	0.377
	(0.042)	(0.012)	(0.021)				
6	0.623	0.076	0.202	0.55	0.38	0.49	0.396
	(0.039)	(0.011)	(0.023)				
7	0.616	0.069	0.179	0.55	0.46	0.44	0.318
	(0.040)	(0.011)	(0.026)				
8	0.530	0.077	0.163	0.59	0.38	0.44	0.403
	(0.037)	(0.009)	(0.016)				
9	0.558	0.090	0.204	0.58	0.28	0.52	0.498
	(0.041)	(0.010)	(0.020)				
10	0.537	0.060	0.130	0.63	0.43	0.49	0.473
	(0.041)	(0.008)	(0.015)				

Note: CAS estimates with the savings rate as proxy after premultiplying each view, and the covariates, by the inverse of the Cholesky factor of the within-group covariance matrix of the corresponding residuals from the Euclidean CAS fit of Table 3 (Remark 4). "Diff. vs." are the shares of countries classified differently from the Euclidean CAS and GFE partitions. The last column is the GFE estimate of β_1 on the whitened outcome. Large-T clustered standard errors in parentheses. 1,000 random starts and 500 variable neighborhood search steps.

to democracy (Cincotta, 2009, Weber, 2013, Dyson, 2013). Table S.13 reports the CAS estimates with the median age as the only proxy and with the savings rate and the median age as two proxy blocks, each with its own precision weight. With the median age as the only proxy, the persistence estimate is 0.61 to 0.66 for every G, with 0.62 at $G = 4$, so a proxy of a different nature produces the same departure from the GFE values. The income coefficient is smaller, 0.04 to 0.07 against 0.10 to 0.11 with the savings rate, with 0.05 at $G = 4$, and the cumulative effect is 0.10 to 0.19, with 0.14 at $G = 4$, close to the GFE value. With the two proxies jointly, the estimates at $G = 4$ are 0.66, 0.07 and 0.21. The two proxies thus agree on the persistence of democracy and disagree on the size of the income effect, and the partitions they select differ from each other for about half of the countries, as much as each of them differs from the GFE partition. At most one of the two can therefore be close to the partition of the residual view, and the diagnostic, which rejects neither at $G = 4$, does not have the power to tell which, as discussed in the validity paragraph. Lagged income is itself an included regressor, so it can serve as a proxy in the sense of Section 3.4, with the second reading of Remark 6. Table S.14 reports the estimates with lagged income as the only proxy and the diagnostic for every G. The persistence estimate is again 0.65 to 0.67, with 0.65 at $G = 4$, while the income coefficient falls to 0.015 at $G = 4$, with a standard error of 0.024, and lies between 0.01 and 0.08 for $G \geq 4$, with standard errors of about 0.03, so the cumulative effect is 0.04 to 0.22. Classifying on income trajectories assigns countries to income classes, so the group effects absorb the between-class variation of income and the coefficient is identified from the within-class variation, which is small over seven periods. Theorem 2 does not cover this estimate, since the specification has a second regressor, the income deviations are serially correlated rather than spherical, and the strict exogeneity of income that the theorem requires is doubtful. The cancellation it describes applies at most to the channel of income itself, as Proposition S.4 shows, and only approximately. What can be tested is whether the income partition is

the partition of the outcome equation, and the diagnostic does not reject this at the 5% level for $G \leq 6$, with p-values between 0.085 and 0.19, and rejects it from $G = 7$ on. The income partition differs from the savings-rate partition for 20% to 67% of the countries and from the median-age partition for 27% to 59% (not reported in the tables). The three proxies therefore agree on the persistence of democracy, between 0.6 and 0.7 under every proxy and every G against 0.25 to 0.60 under GFE, and disagree on the income coefficient, 0.10 with the savings rate, 0.05 with the median age and 0.02 with income at $G = 4$, so that the cumulative effect ranges from 0.04 to 0.28. The diagnostic rejects none of the three at $G = 4$ although their partitions differ pairwise for about half of the countries, so at most one of them can share the partition of the residual view and the diagnostic does not have the power to tell which. In the reading of Remark 6, the non-rejection for income would be consistent with a fixed-effects specification in which income classes are the latent groups, but that reading requires the income deviations to be independent of both partitions, which a predetermined and serially correlated regressor does not satisfy. The reading is therefore at most suggestive, with the same qualification about power, and with the further caveat that lagged income is predetermined rather than strictly exogenous. The persistence of democracy is thus the finding that survives the choice of proxy, and the size of the income effect is not identified by this sample without a stand on which proxy is valid.

Table S.13: Median age as an alternative proxy

	Median age as proxy					Savings rate and median age				
				Diff. vs.					Diff. vs.	
G	$\hat{\beta}_1$	$\hat{\beta}_2$	Cumulative	GFE	CAS	$\hat{\beta}_1$	$\hat{\beta}_2$	Cumulative	GFE	CAS
2	0.657	0.067	0.194	0.31	0.36	0.657	0.067	0.194	0.31	0.36
	(0.049)	(0.015)	(0.033)			(0.049)	(0.015)	(0.033)		
3	0.622	0.054	0.142	0.42	0.49	0.656	0.071	0.206	0.46	0.43
	(0.048)	(0.016)	(0.039)			(0.049)	(0.019)	(0.044)		
4	0.624	0.051	0.135	0.54	0.60	0.659	0.071	0.207	0.52	0.53
	(0.047)	(0.016)	(0.039)			(0.049)	(0.019)	(0.045)		
5	0.616	0.037	0.097	0.59	0.59	0.624	0.074	0.198	0.58	0.57
	(0.051)	(0.017)	(0.044)			(0.047)	(0.018)	(0.043)		
6	0.623	0.044	0.118	0.63	0.62	0.592	0.053	0.129	0.60	0.53
	(0.047)	(0.018)	(0.049)			(0.049)	(0.018)	(0.045)		
7	0.635	0.047	0.129	0.61	0.63	0.598	0.065	0.162	0.59	0.52
	(0.045)	(0.018)	(0.048)			(0.048)	(0.018)	(0.042)		
8	0.619	0.046	0.120	0.60	0.59	0.600	0.063	0.157	0.62	0.53
	(0.048)	(0.018)	(0.046)			(0.048)	(0.018)	(0.044)		
9	0.622	0.051	0.135	0.58	0.67	0.599	0.052	0.131	0.64	0.56
	(0.042)	(0.017)	(0.046)			(0.047)	(0.018)	(0.044)		
10	0.608	0.059	0.151	0.58	0.66	0.572	0.057	0.134	0.64	0.63
	(0.045)	(0.017)	(0.043)			(0.049)	(0.016)	(0.037)		

Note: CAS estimates with the median age of the population (1970–2000, from the same data set, with the seven observations of Taiwan, which are missing, imputed with the sample mean) as the only proxy, and with the savings rate and the median age as two proxy blocks with their own precision weights. "Diff. vs. GFE" and "Diff. vs. CAS" are the shares of countries classified differently from the GFE partition and from the savings-rate CAS partition of Table 3, after relabeling. Large-T clustered standard errors in parentheses. All estimators use 1,000 random starts and 500 variable neighborhood search steps.

Table S.13 reports the CAS estimates with the median age of the population as the only proxy and with the savings rate and the median age as two proxy blocks, each with its own precision weight, for $G = 2, \ldots, 10$. The median age varies slowly within countries and widely across them, from under 16 to over 35 years, so its within-group variance is small relative to the distances between its group centers, and in the two-block

fit it dominates the partition. The two-block partition is close to the median-age partition, with which it coincides at $G = 2$, and far from the savings-rate partition. The precision weights themselves are not reported because they depend on the units of each proxy, to which the estimator is invariant. The median-age partition disagrees with the savings-rate partition for 36% to 67% of the countries and with the GFE partition for 31% to 63%. The diagnostic of Section 2.4 applied to the median age at $G = 4$, with residual draws, gives LR $= 282.0$, almost entirely from the outcome term, and a p-value of 0.245, so it rejects neither proxy at $G = 4$ although their partitions are far apart. The persistence estimate is 0.61 to 0.66 under the median age and 0.57 to 0.66 under the two blocks, in line with the 0.63 to 0.70 under the savings rate and well above the GFE range, while the income coefficient is 0.04 to 0.07 and 0.05 to 0.07 against 0.09 to 0.11 under the savings rate, with standard errors of the same size under every proxy.

Table S.14: Lagged income as the proxy

	Estimates			Diff. vs.			Validity diagnostic				
G	$\hat{\beta}_1$	$\hat{\beta}_2$	Cumulative	GFE	Savings	Median age	LR	LR_y	LR_z	p-value	Diff. GFE/Proxy
2	0.667	0.080	0.240	0.22	0.20	0.27	105.6	100.2	5.4	0.185	0.24
	(0.048)	(0.016)	(0.036)								
3	0.668	0.052	0.157	0.47	0.48	0.32	188.8	187.4	1.4	0.155	0.47
	(0.050)	(0.018)	(0.057)								
4	0.651	0.015	0.042	0.48	0.49	0.40	266.3	256.5	9.8	0.125	0.49
	(0.049)	(0.024)	(0.068)								
5	0.652	0.027	0.077	0.52	0.57	0.46	352.4	344.4	8.0	0.115	0.53
	(0.052)	(0.025)	(0.074)								
6	0.653	0.020	0.057	0.59	0.62	0.54	418.3	418.3	0.0	0.085	0.59
	(0.052)	(0.029)	(0.085)								
7	0.658	0.030	0.088	0.56	0.62	0.58	476.3	448.9	27.5	0.040	0.58
	(0.051)	(0.031)	(0.090)								
8	0.651	0.076	0.219	0.59	0.58	0.57	517.9	485.3	32.6	0.030	0.60
	(0.050)	(0.031)	(0.092)								
9	0.672	0.043	0.130	0.58	0.67	0.59	578.4	500.0	78.4	0.020	0.62
	(0.049)	(0.032)	(0.101)								
10	0.663	0.031	0.093	0.59	0.62	0.59	625.6	582.7	42.9	0.005	0.61
	(0.052)	(0.031)	(0.096)								

Note: CAS estimates with lagged log income per capita, an included regressor, as the only proxy. "Diff. vs." are the shares of countries classified differently from the GFE partition and from the savings-rate and median-age CAS partitions, after relabeling. The diagnostic columns report the statistic (10), its two terms, the bootstrap p-value with $B = 199$ draws in which the within-group residuals of both equations are resampled jointly by country and lagged income, being the proxy, is regenerated from its group centers before the democracy index, and the share of countries classified differently by the GFE partition and by the k-means partition of the income trajectories. Large-T clustered standard errors in parentheses. All estimators use 1,000 random starts and 500 variable neighborhood search steps.

Table S.14 reports the estimates with lagged income as the only proxy, which Section 6.2 and the paragraph on alternative proxies above discuss. The k-means partition of the income trajectories is a partition into income classes, and the CAS partition coincides with it or nearly so, since the proxy term LR_z of the statistic is small at every G and zero at $G = 6$. The persistence estimate is 0.65 to 0.67 for every G, and the income coefficient is 0.02 to 0.08, against 0.09 to 0.11 with the savings rate and 0.04 to 0.07 with the median age.

Bias-aware inference. Table S.15 reports the 90% bias-aware intervals of González-Casasús (2026*a*) for β_1 and β_2 under GFE and under CAS, with the calibrated bounds of Remark 2 and the safety factor of two used in that paper, in two versions. The baseline version uses the sensitivities (12) with the one-half offset and no covariate shift. Lagged democracy is a predetermined regressor, however, and under a residual-based rule the countries misclassified from group b into group a have lagged outcomes shifted

by the offset of their past shocks, a shift that Remark 1 of González-Casasús (2026*a*) computes recursively from the offset and the persistence parameter and that their Supplemental Appendix S.4.5 adds to the sensitivities as a plug-in correction. The two-view rule has the same outcome offset at the precision weight, so the same correction applies to CAS with its own profiles, and the second version of the table uses it. The GFE intervals reproduce those of that paper in both versions. At $G = 4$, the baseline symmetric interval for the persistence parameter widens from the standard $[0.215, 0.389]$ to $[0.123, 0.480]$, because every flow pulls $\hat{\beta}_1$ towards the pooled value, and the shift correction removes most of that exposure, leaving $[0.207, 0.396]$. For the income coefficient, the baseline interval, $[0.066, 0.098]$, is the standard one, and the correction makes the bias set one-sided and widens it to $[0.057, 0.107]$, a worst-case upward bias of 16%. Under CAS, the bias-aware intervals are nearly identical to the standard ones in both versions, $[0.561, 0.710]$ against $[0.562, 0.708]$ for β_1 and $[0.080, 0.124]$ against $[0.081, 0.122]$ for β_2 at $G = 4$, because the calibrated outgoing misclassification probabilities of the two-view rule are small. The calibration is computed from the stacked distances and presumes a valid proxy, so the CAS intervals are conditional on validity, and on the absence of a covariate shift for income, and do not account for a misaligned savings rate, which the estimated weight does not detect either. The savings rate is measured precisely relative to the distances between its group centers, so the stacked distances leave little ambiguity in the assignment of most countries. The CAS intervals for the persistence parameter lie entirely above the GFE intervals, including the bias-aware ones, for $G = 3$ to 5 and for $G = 7$ to 10, overlap them at $G = 2$ and, at $G = 6$, overlap the baseline bias-aware interval marginally while lying above the shift-corrected one. For the income coefficient, the intervals of the two estimators overlap at every G except 6, with the CAS intervals shifted upward, so the difference in the cumulative effect comes mostly from the persistence parameter. The finite-sample inflation of the sensitivities discussed in Section 5.3 widens the corrected intervals by at most 0.015 at each end for GFE and 0.01 for CAS, since the calibrated bounds are small.

Relation to the local analysis of the GFE estimates. The direction of the change relative to GFE deserves comment in the light of the bias-aware analysis of González-Casasús (2026*a*), who applies the local misclassification framework to the GFE estimates of this specification and finds that, around the GFE partition, classification error can only have inflated the income coefficient, by at most 16% under calibrated misclassification bounds and the shift-corrected sensitivities discussed above, and the persistence parameter by at most 0.02 under the same correction, against a two-sided exposure of 0.09 under the baseline sensitivities. The CAS estimates move both coefficients up, not down, and by far more. There is no contradiction, because the two exercises condition on different partitions and answer different questions. The local analysis takes the residual-based partition as approximately correct and asks how far a few misclassified units, at most a tenth of the sample in the most exposed group, can move the estimate, with sensitivities computed at the GFE profiles, which track the democracy transitions. The CAS estimator asks instead whether the residual-based partition is the one generated by the latent types when a second view of those types is available, and the answer is that the two views disagree for more than half of the countries. A disagreement of this size is not a perturbation of the GFE partition, and the local bias set, which describes the consequences of the GFE classifier's own errors, does not bound the difference between estimators that rest on different partitions. What the local analysis does establish is that, if the GFE partition were the right one, the auxiliary covariate would have little to add, since the misclassification bounds it calibrates are small. The CAS results therefore stand or fall with the validity of the savings rate as a proxy, which is the maintained assumption discussed above, and not with the size of the bias of the GFE estimator around

its own partition.

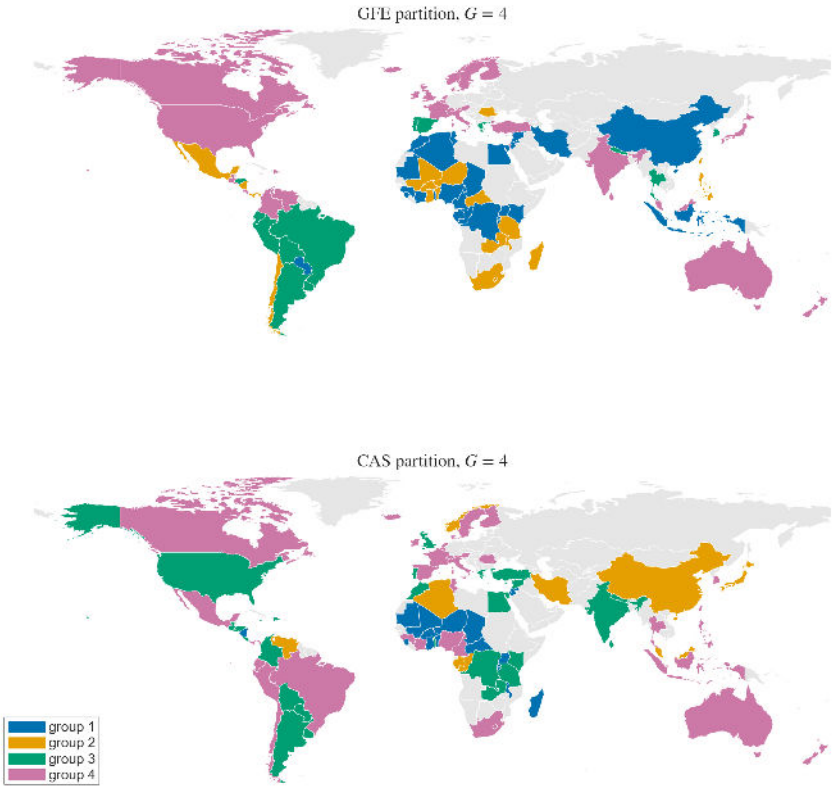


Figure S.4: Estimated groups, $G = 4$

Note: GFE partition (top) and CAS partition with the savings rate as proxy (bottom). In each panel, the groups are numbered in increasing order of the average democracy index of their members. Countries outside the balanced panel are in gray. Map polygons from the replication files of Bonhomme and Manresa (2015).

Table S.15: Bias-aware 90% confidence intervals

			Baseline sensitivities		Shift-corrected		Corrected, $\varpi = 1.5$
G		Standard	Symmetric	Recentered	Symmetric	Recentered	Symmetric
			Lagged democracy β_1				
2	GFE	[0.534, 0.667]	[0.508, 0.693]	[0.506, 0.655]	[0.533, 0.668]	[0.532, 0.665]	[0.533, 0.669]
	CAS	[0.581, 0.741]	[0.581, 0.741]	[0.580, 0.740]	[0.581, 0.741]	[0.581, 0.741]	[0.581, 0.741]
3	GFE	[0.323, 0.490]	[0.257, 0.556]	[0.257, 0.472]	[0.321, 0.492]	[0.317, 0.485]	[0.317, 0.496]
	CAS	[0.561, 0.720]	[0.561, 0.720]	[0.559, 0.718]	[0.561, 0.720]	[0.560, 0.719]	[0.561, 0.720]
4	GFE	[0.215, 0.389]	[0.123, 0.480]	[0.123, 0.370]	[0.207, 0.396]	[0.202, 0.380]	[0.193, 0.410]
	CAS	[0.562, 0.708]	[0.561, 0.710]	[0.558, 0.705]	[0.562, 0.709]	[0.561, 0.707]	[0.562, 0.709]
5	GFE	[0.174, 0.335]	[0.089, 0.420]	[0.089, 0.317]	[0.169, 0.340]	[0.165, 0.328]	[0.163, 0.346]
	CAS	[0.564, 0.709]	[0.563, 0.710]	[0.562, 0.708]	[0.564, 0.709]	[0.563, 0.708]	[0.564, 0.709]
6	GFE	[0.397, 0.534]	[0.370, 0.561]	[0.369, 0.534]	[0.396, 0.535]	[0.393, 0.531]	[0.395, 0.536]
	CAS	[0.557, 0.705]	[0.546, 0.716]	[0.542, 0.696]	[0.556, 0.706]	[0.552, 0.700]	[0.555, 0.707]
7	GFE	[0.335, 0.471]	[0.262, 0.544]	[0.262, 0.478]	[0.328, 0.478]	[0.324, 0.465]	[0.324, 0.482]
	CAS	[0.578, 0.719]	[0.571, 0.725]	[0.572, 0.725]	[0.577, 0.719]	[0.576, 0.718]	[0.577, 0.720]
8	GFE	[0.264, 0.403]	[0.215, 0.451]	[0.215, 0.416]	[0.261, 0.405]	[0.258, 0.398]	[0.258, 0.409]
	CAS	[0.553, 0.707]	[0.537, 0.723]	[0.536, 0.707]	[0.551, 0.709]	[0.547, 0.703]	[0.547, 0.714]
9	GFE	[0.266, 0.405]	[0.215, 0.456]	[0.215, 0.410]	[0.263, 0.408]	[0.260, 0.401]	[0.261, 0.410]
	CAS	[0.572, 0.721]	[0.534, 0.759]	[0.534, 0.723]	[0.565, 0.728]	[0.560, 0.714]	[0.555, 0.738]
10	GFE	[0.278, 0.419]	[0.235, 0.462]	[0.234, 0.421]	[0.276, 0.421]	[0.273, 0.415]	[0.275, 0.422]
	CAS	[0.638, 0.763]	[0.624, 0.777]	[0.622, 0.764]	[0.635, 0.766]	[0.632, 0.759]	[0.632, 0.769]
			Income β_2				
2	GFE	[0.042, 0.079]	[0.038, 0.084]	[0.037, 0.076]	[0.032, 0.089]	[0.032, 0.075]	[0.021, 0.100]
	CAS	[0.063, 0.116]	[0.063, 0.117]	[0.064, 0.117]	[0.063, 0.117]	[0.064, 0.117]	[0.063, 0.117]
3	GFE	[0.071, 0.108]	[0.069, 0.109]	[0.073, 0.110]	[0.064, 0.115]	[0.064, 0.104]	[0.055, 0.124]
	CAS	[0.075, 0.121]	[0.074, 0.121]	[0.076, 0.122]	[0.074, 0.121]	[0.076, 0.122]	[0.074, 0.121]
4	GFE	[0.067, 0.098]	[0.066, 0.098]	[0.067, 0.098]	[0.057, 0.107]	[0.057, 0.095]	[0.047, 0.117]
	CAS	[0.081, 0.122]	[0.080, 0.124]	[0.083, 0.125]	[0.080, 0.123]	[0.083, 0.124]	[0.080, 0.124]
5	GFE	[0.064, 0.095]	[0.063, 0.096]	[0.063, 0.095]	[0.055, 0.103]	[0.055, 0.092]	[0.048, 0.110]
	CAS	[0.083, 0.124]	[0.082, 0.125]	[0.084, 0.126]	[0.082, 0.124]	[0.084, 0.125]	[0.082, 0.124]
6	GFE	[0.052, 0.075]	[0.052, 0.076]	[0.051, 0.075]	[0.049, 0.079]	[0.048, 0.074]	[0.047, 0.081]
	CAS	[0.085, 0.129]	[0.080, 0.134]	[0.088, 0.135]	[0.084, 0.131]	[0.087, 0.132]	[0.083, 0.131]
7	GFE	[0.051, 0.078]	[0.050, 0.080]	[0.049, 0.078]	[0.045, 0.085]	[0.044, 0.076]	[0.042, 0.088]
	CAS	[0.080, 0.122]	[0.079, 0.123]	[0.079, 0.124]	[0.079, 0.123]	[0.081, 0.123]	[0.078, 0.124]
8	GFE	[0.056, 0.083]	[0.055, 0.084]	[0.055, 0.083]	[0.053, 0.086]	[0.052, 0.081]	[0.050, 0.090]
	CAS	[0.075, 0.116]	[0.069, 0.122]	[0.076, 0.123]	[0.073, 0.118]	[0.076, 0.119]	[0.069, 0.122]
9	GFE	[0.057, 0.082]	[0.057, 0.082]	[0.057, 0.082]	[0.054, 0.085]	[0.054, 0.081]	[0.052, 0.087]
	CAS	[0.076, 0.116]	[0.064, 0.127]	[0.076, 0.127]	[0.072, 0.119]	[0.077, 0.120]	[0.068, 0.123]
10	GFE	[0.056, 0.080]	[0.056, 0.080]	[0.056, 0.080]	[0.054, 0.082]	[0.053, 0.079]	[0.052, 0.084]
	CAS	[0.069, 0.107]	[0.064, 0.112]	[0.070, 0.112]	[0.067, 0.109]	[0.070, 0.110]	[0.065, 0.111]

Note: Standard intervals based on the large-T clustered standard errors, and symmetric and recentered bias-aware intervals of González-Casasús (2026*a*) with calibrated bounds on the outgoing misclassification probabilities (safety factor $\varsigma = 2$, offset one half), with the baseline sensitivities and with the sensitivities corrected for the covariate shift of lagged democracy (their Supplemental Appendix S.4.5). The last column adds the finite-sample inflation of the bias set by $(1 - \varpi\bar{p})^{-2}$ with $\varpi = 1.5$ and $\bar{p}$ the largest calibrated bound. For CAS, the calibration uses the stacked distances. The GFE rows reproduce the intervals reported in that paper for $G = 4$ and $G = 8$.