Confidence Set for Group Membership[*]

Andreas Dzemski[†] and Ryo Okui[‡]

December 16, 2023

Abstract

Our confidence set quantifies the statistical uncertainty from data-driven group assignments in grouped panel models. It covers the true group memberships jointly for all units with pre-specified probability and is constructed by inverting many simultaneous unit-specific one-sided tests for group membership. We justify our approach under $N, T \to \infty$ asymptotics using tools from high-dimensional statistics, some of which we extend in this paper. We provide Monte Carlo evidence that the confidence set has adequate coverage in finite samples. An empirical application illustrates the use of our confidence set.

[*] The authors would like to thank Stèphane Bonhomme (Editor), a co-editor, two anonymous referees, Otilia Boldea, Christoph Breunig, Le-Yu Chen, Elena Erosheva, Eric Gautier, Hidehiko Ichimura, Hiroaki Kaido, Hiroyuki Kasahara, Kengo Kato, Toru Kitagawa, Arthur Lewbel, Artem Prokhorov, Adam Rosen, Myung Hwan Seo, Katsumi Shimotsu, Liangjun Su, Michael Vogt, Wendun Wang, Wuyi Wang, Martin Weidner, Yoon-Jae Whang and seminar participants at the Centre for Panel Data Analysis Symposium at the University of York, St Gallen, HKUST, SUFE, Sydney Econometric Reading Group, Xiamen, CUHK Workshop on Econometrics, Chinese Academy of Sciences, Asian Meeting of Econometric Society 2017, STJU, Workshop on Advances in Econometrics 2017 at Hakodate, SNU, Academia Sinica, ESEM 2017, Berlin Humboldt, CFE-CMStatistics 2017, International Panel Data Conference, Tsinghua, Fudan, Bonn, Hanyang, Lund, Exeter, Montreal, Cambridge, Panel data workshop Amsterdam, Barcelona GSE summer forum and 2021 Nanyang Econometric Workshop for valuable comments. We are particularly grateful to one anonymous referee whose comment led to correcting an error in a previous version of the manuscript. Sophie Li and Heejun Lee provided excellent research assistance. A part of this research was done while Okui was at Kyoto University, Vrije Universiteit Amsterdam, NYU Shanghai, and Seoul National University. This work is supported by JSPS KAKENHI Grant number 15H03329, 16K03598, 22H00833, 22K20154 and 23H00804, and Jan Wallanders och Tom Hedelius stiftelse samt Tore Browaldhs stiftelse grant number P19-0079.

[†] Department of Economics, University of Gothenburg, P.O. Box 640, SE-405 30 Gothenburg, Sweden. Email: andreas.dzemski@economics.gu.se

[‡] Corresponding Author. Graduate School of Economics and Faculty of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyou-ku, Tokyo, 113-0033, Japan. Email: okuiryo@e.u-tokyo.ac.jp;

Keywords: Panel data, clustering, confidence set, joint one-sided test, high-dimensional statistics.

JEL codes: C23, C33, C38

1 Introduction

Clustering units into discrete groups is one of the oldest problems in statistics (Pearson 1896). It has received interest in the recent econometric literature on grouped panel models (Hahn and Moon 2010; Lin and Ng 2012; Bonhomme and Manresa 2015; Sarafidis and Weber 2015; Ando and Bai 2016; Vogt and Linton 2017; Su, Shi, and Phillips 2016; Wang, Phillips, and Su 2018; Vogt and Schmid 2021; Lu and Su 2017; Gu and Volgushev 2019; Liu et al. 2020; Wang and Su 2021; Mammen, Wilke, and Zapp 2022; Mehrabani 2022; Mugnier 2022; Chetverikov and Manresa 2022; Yu, Gu, and Volgushev 2023; Mugnier 2023).

In grouped panel models, a data-driven clustering algorithm is used to estimate a latent group structure. As statistical procedures, clustering algorithms suffer from sampling errors and produce a noisy version of the true group structure. The existing literature gives little guidance on how to assess clustering uncertainty in a given application. Inferential theory for grouped panel models has focused on group characteristics, but is underdeveloped for assessing the uncertainty about individual group memberships (see e.g. McLachlan and Peel 2004).

In this paper, we quantify the statistical uncertainty about the true group memberships in a grouped panel model. As far as we know, we are the first to propose and justify a rigorous frequentist method to evaluate clustering uncertainty.

In grouped panel models, individual regression curves are heterogeneous and exhibit a grouped pattern. All units belonging to the same group face the same regression curve. Group memberships are unobserved and estimated by a clustering algorithm. Clustering uncertainty means that the algorithm may misclassify some units and assign them an incorrect regression curve.

We propose a confidence set for group membership that quantifies clustering uncertainty jointly for all units in the panel. For a panel of N units, an element of a joint confidence set is an N-dimensional vector that specifies a group assignment for every unit. Our confidence set gathers all N-dimensional vectors of group assignments that are “not ruled out by the data” and is guaranteed to contain the vector of the true group memberships with a pre-specified probability, say 95%.

The latent groups in a grouped panel model have no natural labels and can only be identified up to a permutation. For notational convenience, we write our confidence set using an arbitrary ordering of the groups. We interpret this as a shorthand for linking units to regression curves. For example, suppose that there is a "group 1" with a slope coefficient of 0.2 and a "group 2" with a slope coefficient of 0.8. If our confidence set rules out that unit i belongs to "group 1", then we take this to mean that unit i does not face a slope coefficient of 0.2. This interpretation does not depend on the ordering of the groups. Similarly, if our confidence set determines that units i and j belong to different groups, then we take this to mean that they face different slope coefficients.

This interpretation of a confidence set assumes that the data are rich enough to recover the group-specific coefficients. If the data do not provide any clue about the group-specific coefficients, then we are not able to statistically examine group memberships either. On the other hand, if the data are known to be "very rich" and group memberships are guaranteed to be estimated correctly, then our confidence set is not needed.

Settings between these two extreme scenarios are relevant in practice. In Monte Carlo experiments calibrated to their empirical application, Bonhomme and Manresa (2015) find that units are frequently misclassified, whereas group-specific coefficients are estimated precisely (see Table S.III in their supplemental appendix). Dzemski and Okui (2021) provide a theoretical framework to explain this observation.[1] They assume that unit i faces an error term with unit-specific variance σ_i^2. Units with small σ_i are classified reliably. Units with large σ_i are potentially misclassified. Dzemski and Okui (2021) show that the group-specific coefficients can be estimated consistently if the proportion of potentially misclassified units is sufficiently small. This is the main setting that we have in mind for applications of our confidence set. It is less restrictive than assuming, as is typically done in the literature, that both group-specific coefficients and all group assignments can be reliably estimated.

Our empirical application illustrates how our confidence set provides new economic insights. We follow Wang, Phillips, and Su (2019) who estimate a panel model that allows the effect of a minimum wage on unemployment to vary between US states, depending on the assignment of each state to one of four latent groups. This group assignment is potentially estimated with error. We use our confidence set to identify, up to a small pre-specified error probability, states without clustering uncertainty. In terms of the framework discussed above, these are states with low values of σ_i. For these states, we

[1] For mathematical convenience, the theoretical analysis of clustered panel models often proceeds under assumptions that rule out any misclassification in the asymptotic limit (see e.g. Bonhomme and Manresa 2015; Vogt and Linton 2017).

can identify the state-specific effects of the minimum wage.

Our confidence set can also be used to enhance a plot of the estimated groups by adding information about clustering uncertainty. We illustrate this in our empirical application. Providing a plot of the estimated groups is standard practice.[2] This is true even in applications where the group structure is considered merely a nuisance parameter.[3]

An alternative to our frequentist approach is Bayesian inference. Fully parametric grouped-panel models are finite mixtures models that can be estimated by the EM algorithm (Dempster, Laird, and Rubin 1977). The E-step of the EM algorithm computes unit-wise posterior probabilities for group membership. These are valid if the units in the panel are independently drawn from the assumed parametric distribution. Our frequentist approach is more general. We do not assume a parametric distribution of the error term and show that our approach is valid for error distributions in a broad nonparametric class. We also allow for cross-sectional dependence, non-random patterns of heteroscedasticity, and non-random group assignments. Another advantage of our procedure is that it allows for joint inference on the N-dimensional vector of all group memberships, whereas unit-wise posterior probabilities only address uncertainty about the group membership of a single unit.

Quantifying the uncertainty about the true group structure is a high-dimensional inference problem. The N-dimensional vector of true group memberships is high-dimensional since its size grows as $N \to \infty$. To construct a confidence set for this high-dimensional parameter, we invert a test of the many moment inequalities that characterize group memberships. Testing many moment inequalities is the problem considered in Chernozhukov, Chetverikov, and Kato (2019). Their test is based on a single test statistic, whereas ours combines many simultaneous group membership tests for individual units. The advantage of our approach is that it can be inverted without running a computationally infeasible exhaustive search over the space of all partitions.

Our confidence set is valid in the presence of weak-dependence and serial correlation, whereas Chernozhukov, Chetverikov, and Kato (2019) assume independence over time. We account for serial correlation by using a heteroskedasticity–and–autocorrelation–robust (HAC) variance estimator (Andrews 1991; Newey and West 1987) when constructing the

[2] For example, see Figure 2 in Wang, Phillips, and Su (2019), Figure 2 in Bonhomme and Manresa (2015) and Figure 6 in Wang, Phillips, and Su (2018).

[3] For example, time-varying unobserved heterogeneity can be controlled by imposing a latent group structure with group-specific time fixed-effects. In this context, the heterogeneity in the fixed-effect is a nuisance parameter similar to the interacted fixed effects in, for example, Moon and Weidner (2023).

unit-wise test statistics.

The asymptotic analysis of our procedure accounts for the high-dimensional nature of our setting and allows for weak time-dependence. We build on Chang, Chen, and Wu (2023) who provide results for HAC estimators and a high-dimensional central limit theorem for dependent data. Our setting requires extending their approach in different ways. Specifically, we use a Nasarov-type inequality (Chernozhukov, Chetverikov, and Kato 2017) to control for the effect of parameter estimation. Moreover, we develop a regularization scheme for the HAC estimates. This regularization scheme allows us to control the estimation error in our data-driven critical values by using a comparison bound based on Li and Shao (2002).

We provide several extensions of our method. First, we suggest alternative critical values. These are slightly conservative but much easier to compute than our benchmark critical values. Second, we show that HAC estimation is not needed if there is no serial correlation. In this case, test statistics based on the usual variance estimators provide valid confidence sets and are easier to implement than those based on the HAC estimator. Third, we propose a two-step method, called unit selection, to shrink the cardinality of the confidence set when there are many units for which clustering uncertainty is low. The two-step procedure identifies and discards such units. We then construct a confidence set for the remaining units. The method accounts for errors in the unit selection and provides a valid confidence set. This additional error control can inflate the confidence set if unit selection does not eliminate sufficiently many units.

The remainder of this paper is organized as follows. Section 2 introduces the grouped panel model. Section 3 defines our confidence set for group membership. Section 4 proves the asymptotic validity of our confidence sets. Section 5 discusses the extensions of our method. Section 6 provides an empirical application. Section 7 presents Monte Carlo simulations that investigate the validity and power of our confidence set based on simulation designs inspired by our empirical application. Section 8 concludes.

An R package implementing the methods proposed in this paper is included in the replication package that is published together with this article.

2 Model

We observe panel data $(y_{it}, w_{it}', x_{it}')'$ for units $i = 1, \ldots, N$ and time periods $t = 1, \ldots, T$, where y_{it} is a scalar dependent variable and w_{it} and x_{it} are covariate vectors. Unit i belongs to group $g_i^0 \in \mathbb{G} = \{1, \ldots, G\}$. Group memberships are unobserved. The data are generated from the model

$$y_{it} = w_{it}' \theta^w + x_{it}' \theta_{g_i^0} + \sigma_i v_{it}, \tag{1}$$

where v_{it} is a noise term with variance one and is potentially serially correlated, and σ_i is a latent heteroscedasticity parameter. The slope coefficient θ^w on w_{it} is common to all units. The slope coefficient on x_{it} is group-specific and given by θ_g for units i belonging to group $g \in \mathbb{G}$.

We assume that the regressors $(w_{it}', x_{it}')'$ are uncorrelated with the contemporaneous error term $\sigma_i v_{it}$,

$$\mathbb{E}\left[(x_{it}', w_{it}')' \sigma_i v_{it}\right] = 0. \tag{2}$$

This assumption does not rule out predetermined regressors such as lagged dependent variables.

Different estimation strategies for estimating the common and group-specific coefficients (θ^w and θ_g) have been proposed in the literature. For example, Bonhomme and Manresa (2015) estimate slope coefficients $\hat{\theta}^w, \hat{\theta}_1, \ldots, \hat{\theta}_G$ and group memberships $\hat{g}_1, \ldots, \hat{g}_N$ simultaneously by solving the least-squares problem

$$(\hat{\theta}^w, \hat{\theta}_1, \ldots, \hat{\theta}_G, \hat{g}_1, \ldots, \hat{g}_N) = \operatorname{argmin}_{\substack{\theta^w, \theta_1, \ldots, \theta_G \\ g_1, \ldots, g_N \in \mathbb{G}}} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{it} - w_{it}' \theta^w - x_{it}' \theta_{g_i}\right)^2$$

via the *k-means* algorithm. The choice of squared loss is justified under the orthogonality condition (2). The estimators in Su, Shi, and Phillips (2016) and Wang, Phillips, and Su (2018) augment a squared loss function by a penalization scheme that imposes the grouped structure.

We assume that the number of groups G is either pre-specified or consistently estimated. Consistent estimates of G can be obtained, for example, by using the information criteria proposed by Bonhomme and Manresa (2015) or Su, Shi, and Phillips (2016) or by employing the testing procedure in Lu and Su (2017). For simplicity of exposition, we describe our procedure for known G. Our procedure remains valid if G is replaced by a consistent estimator.

Remark 1. *For ease of exposition, we describe our procedures for balanced panels. The extension to unbalanced panels is straightforward.*

3 Confidence set for group membership

This section describes our confidence set for group membership. First, we define the formal requirements for an asymptotically valid confidence set for group membership. Second, we show that each group allocation corresponds to a set of moment inequalities and that a confidence set can be obtained by inverting a test of these inequalities. Finally, we introduce the test statistic and critical values.

3.1 Definition

A joint confidence set of group membership at confidence level $1-\alpha$ is a random set $\widehat{C}_\alpha$ of vectors in $\mathbb{G}^N$ that satisfies

$$\liminf_{N,T\to\infty} \inf_{P\in\mathbb{P}_N} P\left(\left\{g_i^0\right\}_{1\le i\le N} \in \widehat{C}_\alpha\right) \ge 1-\alpha, \tag{3}$$

where $\mathbb{P}_N$ is a class of data-generating processes. If we observe $\{g_i\}_{1\le i\le N} \in \widehat{C}_\alpha$, then the group structure that assigns unit $i = 1, \ldots, N$ to group g_i is not ruled out by the data at confidence level $1-\alpha$. Inequality (3) ensures that the confidence set is asymptotically valid in the sense that it rules out the population partition at most with probability α.

We impose uniform validity over sequences P_N on $\mathbb{P}_N$. Changing the data-generating process along the asymptotic sequence allows for σ_i that diverge as $N \to \infty$ for some units i, rendering these units potentially misclassified in the limit. Data-generating processes that are constant in N cannot model asymptotic clustering uncertainty (Dzemski and Okui 2021).

We construct our joint confidence set by combining unit-wise marginal confidence sets. This approach is computationally simple and can be tabulated and visualized easily. The marginal confidence set for unit i is computed by inverting a test for group membership

$$\widehat{C}_{\alpha,N,i} = \left\{g \in \mathbb{G} : \widehat{T}_i(g) \le \hat{c}_{\alpha,N,i}(g)\right\} \cup \{\hat{g}_i\},$$

where $\hat{g}_i$ denotes an estimator of the group membership of unit i, $\widehat{T}_i(g)$ is a test statistic and $\hat{c}_{\alpha,N,i}$ is a unit-specific and data-dependent critical value. Test statistic and critical value are defined in Sections 3.3 and 3.4, respectively.

By explicitly adding $\hat{g}_i$ to the marginal confidence set, we guarantee that the joint confidence set is never empty and can always be interpreted as containing the estimated group structure padded by a margin of error. For the typical unit, inverting the test already includes the unit's estimated group membership in its marginal confidence set.[4]

Our joint confidence set is given by the Cartesian product of the unit-wise confidence sets:

$$\widehat{C}_\alpha = \bigtimes_{1 \leq i \leq N} \widehat{C}_{\alpha,N,i}.$$

We use Bonferroni correction to control dependence between units and compute each unit-wise marginal confidence set at a nominal level of $1 - \alpha/N$.

In principle, it is possible to construct a joint confidence set directly without first computing unit-wise confidence sets. This can be accomplished by inverting a *joint* test for group membership. Testing group memberships for all groups simultaneously avoids the possible power loss from Bonferroni correction. However, inverting the test to obtain the confidence set requires testing all G^N possible groupings. This task is computationally infeasible unless N is very small. In contrast, our approach carries out only $G \times N$ tests and is feasible even if N is large.

An additional advantage of Bonferroni correction is that it produces confidence sets that are easy to report and interpret. The joint confidence set can be fully described by reporting the marginal unit-wise confidence sets without enumerating all N-dimensional vectors contained in $\widehat{C}_\alpha$. Interpreting a potentially large collection of such high-dimensional vectors would be challenging.

The Bonferroni correction renders our confidence set robust to any kind of cross-sectional dependence. This correction is only minimally conservative if the unit-wise confidence sets are approximately independent, i.e., if

$$P\left(\{g_i^0\}_{1 \leq i \leq N} \in \widehat{C}_\alpha\right) = \prod_{1 \leq i \leq N} P\left(g_i^0 \in \widehat{C}_{\alpha,N,i}\right) + p_\Delta, \tag{4}$$

where p_Δ is a small number. Approximate independence holds if units are cross-sectionally independent and N and T are large enough to estimate the group-specific coefficients precisely.

Theorem 1. *Let $0 < \alpha < 1$ and suppose that the unit-wise confidence sets satisfy*

$$P\left(g_i^0 \in \widehat{C}_{\alpha,N,i}\right) = 1 - \alpha/N \tag{5}$$

[4] In our simulations, we find that the probability of the test rejecting the estimated group membership to be very close to, but not equal to, zero (see Supplemental Material C.2).

and that condition (4) *holds. Then*

$$\alpha - \frac{\alpha^2}{2} - p_\Delta \leq P\left(\{g_i^0\}_{1\leq i\leq N} \notin \widehat{C}_\alpha\right) \leq \alpha - \frac{\alpha^2}{2}\left(1 - \frac{\alpha}{3} + \frac{1}{N}\left(1 - \frac{\alpha}{N}\right)^{-2}\right) + p_\Delta.$$

For example, suppose that $N \geq 8$ and $1 - \alpha = 0.9$ and that conditions (4) and (5) hold with $p_\Delta \approx 0$. Theorem 1 implies that the Bonferroni correction inflates the joint coverage probability by only about 0.5-0.55%.

Remark 2. *Our approach can be adapted to produce joint confidence sets for subsets of units. For $1 \leq K < N$, suppose that the researcher is only interested in the K first units. A joint confidence set for these units is given by*

$$\bigtimes_{1\leq i\leq K} \left\{g \in \mathbb{G} : \widehat{T}_i(g) \leq \hat{c}_{\alpha,K,i}(g)\right\} \cup \{\hat{g}_i\}.$$

For inference on the first K units, this confidence set is more powerful than the joint confidence set for all units. This is because less Bonferroni correction is needed when testing fewer units and therefore $\hat{c}_{\alpha,K,i}(g) < \hat{c}_{\alpha,N,i}(g)$ (see the definition of the critical values in Section 3.4 below).

Remark 3. *For applications where we are interested in inference on a single pre-specified unit i, a confidence set for the group membership of i is given by $\widehat{C}_{\alpha,1,i}$.*

3.2 Motivation of our test of group membership

Our approach to testing the group membership hypothesis $H_0 : g_i^0 = g$ is based on

$$d_{it}(g,h) = \frac{1}{2}\left((y_{it} - w_{it}'\theta^w - x_{it}'\theta_g)^2 - (y_{it} - w_{it}'\theta^w - x_{it}'\theta_h)^2 + (x_{it}'(\theta_g - \theta_h))^2\right).$$

The first two terms on the right-hand side are squared residuals representing the fit of assigning unit i to group g and the fit of assigning unit i to group h, respectively. The third term applies moment re-centering and ensures that $d_{it}(g,h)$ has mean zero under the null hypothesis. This can be seen by re-writing $d_{it}(g,h)$ as

$$d_{it}(g,h) = -\sigma_i v_{it} x_{it}'(\theta_g - \theta_h) + \left(\theta_g - \theta_{g_i^0}\right)' x_{it}x_{it}'(\theta_g - \theta_h).$$

The first term on the right-hand side has mean zero under the orthogonality assumption (2). If the null hypothesis is true, i.e., if $g_i^0 = g$, then the second term vanishes and $\sum_{t=1}^T \mathbb{E}[d_{it}(g,h)] = 0$ for all $h \neq g$.

If $\sum_{t=1}^{T}\mathbb{E}[x_{it}x_{it}']$ has full rank and the null hypothesis is false, then $\sum_{t=1}^{T}\mathbb{E}\left[d_{it}(g,h)\right] > 0$ for some $h \neq g$. The strict inequality holds because for $h = g_i^0 \in \mathbb{G} \setminus \{g\}$ the second term in $d_{it}(g,h)$ is a quadratic form with strictly positive mean.

In summary, testing $H_0 : g_i^0 = g$ is equivalent to testing

$$H_0' : \sum_{t=1}^{T}\mathbb{E}[d_{it}(g,h)] = 0 \quad \text{for all } h \in \mathbb{G} \setminus \{g\}$$

against

$$H_1' : \text{there exists } h \in \mathbb{G} \setminus \{g\} \text{ such that } \quad \sum_{t=1}^{T}\mathbb{E}[d_{it}(g,h)] > 0.$$

This is a one-sided significance test for a vector of moments.

3.3 Test statistic

Our test statistic $\widehat{T}_i(g)$ for unit $i = 1, \ldots, N$ is the maximum of $(G-1)$ statistics $\widehat{D}_i(g,h)$ that test a hypothesized group membership g against alternative group assignments $h \neq g$:

$$\widehat{T}_i(g) = \max_{h \in \mathbb{G} \setminus \{g\}} \widehat{D}_i(g,h).$$

$\widehat{D}_i(g,h)$ tests group g against group h based on the restriction $\sum_{t=1}^{T}\mathbb{E}[d_{it}(g,h)] = 0$ from the previous section and is equal to the t-statistic

$$\widehat{D}_i(g,h) = \frac{\sum_{t=1}^{T}\hat{d}_{it}(g,h)/\sqrt{T}}{\sqrt{\widehat{\Xi}_i(g,h,h)}},$$

where $\hat{d}_{it}(g,h)$ is a sample counterpart of $d_{it}(g,h)$ that replaces the true slope coefficients θ^w and θ_g by their estimated values $\hat{\theta}^w$ and $\hat{\theta}_g$,

$$\hat{d}_{it}(g,h) = \frac{1}{2}\left(\left(y_{it} - w_{it}'\hat{\theta}^w - x_{it}'\hat{\theta}_g\right)^2 - \left(y_{it} - w_{it}'\hat{\theta}^w - x_{it}'\hat{\theta}_h\right)^2 + \left(x_{it}'\left(\hat{\theta}_g - \hat{\theta}_h\right)\right)^2\right),$$

and $\widehat{\Xi}_i(g,h,h)$ is an estimator of the long-run variance of $d_{it}(g,h)$.

The long-run variance estimator is kernel-based as in Andrews (1991) and Newey and West (1987). In order to define the estimator, let $\widehat{H}_{ij}(g,h,h')$ denote the sample covariance of order j between $\hat{d}_{it}(g,h)$ and $\hat{d}_{it}(g,h')$,

$$\widehat{H}_{ij}(g,h,h') = \frac{1}{T}\sum_{t=|j|+1}^{T}\left(\hat{d}_{i,t+\min(0,j)}(g,h) - \bar{\hat{d}}_i(g,h)\right)\left(\hat{d}_{i,t-\max(0,j)}(g,h') - \bar{\hat{d}}_i(g,h')\right), \tag{6}$$

where $\bar{\hat{d}}_i(g,h) = \frac{1}{T}\sum_{t=1}^{T}\hat{d}_{it}(g,h)$. The long-run variance-covariance estimator is given by

$$\widehat{\Xi}_i(g,h,h') = \sum_{j=-T+1}^{T-1} K\left(\frac{j}{\kappa_N}\right)\widehat{H}_{ij}(g,h,h'),$$

where $K(\cdot)$ is a kernel function and κ_N is a bandwidth parameter. In our simulation studies and empirical application, we use the quadratic spectral (QS) kernel (see, for example, Andrews 1991) given by

$$K_{QS}(x) = \frac{25}{12\pi^2 x^2}\left(\frac{\sin(6\pi x/5)}{6\pi x/5} - \cos(6\pi x/5)\right).$$

We select a data-driven bandwidth $\hat{\kappa}_N$ by using the following algorithm adapted from Chang, Chen, and Wu (2023):

Step A: For $g \in \mathbb{G}$ and $h \in \mathbb{G}\setminus\{g\}$, take each unit i such that $\hat{g}_i = g$ and fit an $AR(1)$-model on $(\hat{d}_{it}(g,h))_{t=1}^{T}$. Let $\hat{\rho}_{igh}$ denote the estimated autoregressive coefficients and $\hat{\sigma}^2_{igh}$ the estimated variance of the innovation.

Step B: Select the bandwidth

$$\hat{\kappa}_N = 1.3221\left(T \times \frac{\sum_{i=1}^{N}\sum_{g\in\mathbb{G}}\sum_{h\in\mathbb{G}\setminus\{g\}}\hat{\rho}^2_{igh}\hat{\sigma}^4_{igh}/(1-\hat{\rho}^2_{igh})^8}{\sum_{i=1}^{N}\sum_{g\in\mathbb{G}}\sum_{h\in\mathbb{G}\setminus\{g\}}\hat{\sigma}^4_{igh}/(1-\hat{\rho}^2_{igh})^4}\right)^{1/5}.$$

3.4 Critical values

The critical value $\hat{c}_{\alpha,N,i}(g)$ is computed from the multivariate t-distribution (MVT) in $(G-1)$ dimensions.[5] In the case of two groups ($G=2$), this distribution is equal to

[5] Using the multivariate t-distribution instead of a Gaussian distribution improves the finite sample performance of our confidence set if T is small.

Student's t-distribution and our critical value is given by

$$c_{\alpha,N,i}(g) = \sqrt{\frac{T}{T-1}} t_{T-1}^{-1}\left(1 - \frac{\alpha}{N}\right),$$

where $t_{T-1}^{-1}(p)$ denotes the p-quantile of Student's t-distribution with $(T-1)$ degrees of freedom. This critical value is straightforward to evaluate in most statistical software packages.

For $G \geq 3$, the computation of the critical value is more involved and requires the estimation of unit-specific correlation matrixes. In Section 5, we discuss a conservative approximation of the critical value that is easy to implement and independent of the data.

The critical value is given by

$$\hat{c}_{\alpha,N,i}(g) = c_{\alpha,N}\left(\rho(\widehat{\Omega}_i(g), \epsilon_N)\right) = \sqrt{\frac{T}{T-1}}\left(t_{\max,\rho(\widehat{\Omega}_i(g),\epsilon_N),T-1}\right)^{-1}\left(1 - \frac{\alpha}{N}\right),$$

where $t_{\max,\Omega,T-1}$ denotes the distribution function of the maximal entry of a centered random vector with multivariate t-distribution with scale matrix Ω and $(T-1)$ degrees of freedom, $\widehat{\Omega}_i(g)$ is an estimator of the correlation matrix of the moment inequalities, ρ is a regularization function and ϵ_N is a regularization parameter.[6]

To define the estimated correlation matrix $\widehat{\Omega}_i(g)$ for given g and i, map $j, j' = 1, \ldots, G-1$ to $h, h' \in \mathbb{G}$ such that h and h' give the jth and j'th element of the vector $\mathbb{G}/\{g\}$, respectively.[7] $\widehat{\Omega}_i(g)$ is given by the $(G-1) \times (G-1)$ matrix with entry (j, j') equal to

$$\left(\widehat{\Omega}_i(g)\right)_{j,j'} = \frac{\widehat{\Xi}_i(g, h, h')}{\sqrt{\widehat{\Xi}_i(g, h, h)\widehat{\Xi}_i(g, h', h')}}.$$

For a $(G-1) \times (G-1)$ correlation matrix Ω, the regularization function ρ is defined as

$$\rho(\Omega, \epsilon) = \operatorname{diag}^{-1/2}\left(\Omega + \epsilon^*(\Omega, \epsilon) I_{G-1}\right)\left(\Omega + \epsilon^*(\Omega, \epsilon) I_{G-1}\right)\operatorname{diag}^{-1/2}\left(\Omega + \epsilon^*(\Omega, \epsilon) I_{G-1}\right),$$

where I_{G-1} is the identity matrix in $\mathbb{R}^{G-1}$, $\operatorname{diag}(A)$, for a matrix A, returns a diagonal matrix of the same dimension as A with the diagonal entries equal to the diagonal entries

[6] The distribution function of the multivariate t-distribution can be efficiently approximated by modern algorithms (Genz 1992). Implementations exist for Stata (Grayling and Mander 2016) and R (Azzalini and Genz 2016).

[7] More formally, $h = j - \mathbf{1}_{\{j>g\}}$ and $h' = j' - \mathbf{1}_{\{j'>g\}}$, where $\mathbf{1}_{\{\cdot\}}$ is the indicator function.

of A, and

$$\epsilon^*(\Omega, \epsilon) = \max\{0, \epsilon - (1 - \max_{i<j} \Omega_{ij})\}.$$

We set the regularization parameter equal to $\epsilon_N = 0.01$. For robustness, we also conduct simulations with other values of ϵ_N and find that the exact choice of ϵ_N is not crucial to the validity of our method. Regularization is a technical tool needed to prove the asymptotic validity of our confidence set.

Our regularization scheme bounds pairwise correlations away from one but may output a singular matrix. A related regularization scheme in Andrews and Barwick (2012) bounds the resulting matrix away from singularity.

4 Asymptotic validity of our confidence set

Our asymptotic framework is of the long-panel variety and takes both the number of units N and the number of time periods T to infinity. In particular, we consider asymptotic sequences in which $T = T(N)$, where $T(\cdot)$ is increasing, but its exact form is unspecified except for conditions given in the statement of the theorems. In many panel data sets, the number of units far exceeds the number of time periods. We replicate this feature along the asymptotic sequence by allowing N to diverge at a much faster rate than T.

We consider a sequence $\mathbb{P}_N$ of classes of probability measures. All our theoretical results hold uniformly over the sequence $\mathbb{P}_N$. For a probability measure P, let $\mathbb{E}_P$ denote the expectation operator that integrates with respect to measure P. The parameters θ^w, $\{\theta_g\}_{g\in\mathbb{G}}$ and σ_i depend potentially on N. For notational convenience, we keep this dependence implicit.

The number of latent groups G is fixed and does not depend on N.

To state our assumptions, we define the matrix $\Omega_i(g_i^0)$ which provides the population counterpart to $\widehat{\Omega}_i(g_i^0)$. For unit i, define the population long-run covariance of the averages of $d_{it}(g_i^0, h)$ and $d_{it}(g_i^0, h')$ as

$$\Xi_i(h, h') = \sum_{j=-T+1}^{T-1} H_{ij}(h, h'),$$

where

$$H_{it}(h,h') = \frac{1}{T}\sum_{t=|j|+1}^{T} \mathbb{E}_P[d_{i,t+\min(0,j)}(g_i^0,h)d_{i,t-\max(0,j)}(g_i^0,h')].$$

For $i = 1,\ldots,N$, $\Omega_i(g_i^0)$ is the $(G-1)\times(G-1)$ correlation matrix with entries

$$\left(\Omega_i(g_i^0)\right)_{j,j'} = \frac{\Xi_i(h,h')}{\sqrt{\Xi_i(h,h)\Xi_i(h',h')}},$$

where the convention relating (j,j') to (h,h') is defined in footnote 7.

We now state our assumptions for the asymptotic validity of our joint confidence set.

Assumption 1. *1. (Regressors are uncorrelated with the error term) The linear panel model satisfies the orthogonality assumption* (2)*.*

2. (Number of groups) The number of latent groups G is known and fixed along the asymptotic sequence.

3. (Estimation of auxiliary parameters) There are vanishing sequences $r_{\theta,N}$ and $a_{\theta,N}$ such that

$$\sup_{P\in\mathbb{P}_N} P\left(\|\hat{\theta}^w - \theta^w\| \vee \max_{g\in\mathbb{G}}\|\hat{\theta}_g - \theta_g\| > r_{\theta,N}\right) \le a_{\theta,N}.$$

4. (Full rank) Let $\lambda_{\min}(\cdot)$ denote the smallest eigenvalue of its argument. There is a finite constant $C_\lambda > 0$ such that

$$\inf_{P\in\mathbb{P}_N}\min_{1\le i\le N}\lambda_{\min}\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}_P[v_{it}v_{is}x_{it}x_{is}']\right) \ge C_\lambda^{-1}.$$

5. (Group separation) $\iota_N \equiv \min_{g\in\mathbb{G}}\min_{h\in\mathbb{G}\setminus\{g\}}\|\theta_g - \theta_h\| > 0$.

6. (Exponential tail bound) There exist constants a and $d_1 > 1$ such that $P(|Z| > z) < \exp(-(z/a)^{d_1})$ for sufficiently large z, where Z is any component of the random vector $(x_{it}', w_{it}', v_{it})$.

7. (Mixing sequence) For every $i = 1,2,\ldots,$ the sequence $(x_{it}', w_{it}', v_{it})_{t=1}^{T}$ is an α-mixing sequence with mixing coefficient α_i satisfying $\sup_i \alpha_i[k] \le \exp(1 - bk^{d_2})$ for some $b > 0$ and $d_2 > 0$.

8. *(Stationarity) For every $i = 1, 2, \dots,$ $\{x'_{it}, w'_{it}, v_{it}\}_{t=1}^{T}$ is a strictly stationary time series.*

9. *(Correlation of moment inequalities)* $\min_{1 \le i \le N} \min_{1 \le j, j' \le G-1} \left(\Omega_i(g_i^0)\right)_{j,j'} > -1 + 4\epsilon_N/3$*, where ϵ_N is the regularization parameter for $\rho(\cdot, \cdot)$.*

Assumption 1.1 requires the regressors to be uncorrelated with the contemporaneous error term. It is a much weaker exogeneity assumption than, for example, strict exogeneity and allows for a rich set of regressors, including lagged dependent variables. Assumption 1.2 requires the number of groups to be consistently estimated. This condition is weak and can be guaranteed by using an appropriate procedure for choosing the number of groups (e.g. Lu and Su 2017; Vogt and Schmid 2021).

Assumption 1.3 requires the estimators $\hat{\theta}^w$ and $\hat{\theta}_g$ to be consistent for θ^w and θ_g, respectively, at a rate that vanishes as fast as or faster than $r_{\theta,N}$. Theorem 2 below require $r_{\theta,N}$ to vanish faster than $(T \log N)^{-1/2}$. Therefore, $\hat{\theta}^w$ and $\hat{\theta}_g$ have to converge faster than $T^{-1/2}$. Since $T^{-1/2}$ is the rate obtained by estimators based on time-series regression within units, it is important to choose estimators that also exploit cross-sectional variation such as the *k-means* estimator (Bonhomme and Manresa 2015) or the estimators in Su, Shi, and Phillips (2016) and Wang, Phillips, and Su (2018). These estimators are known to be $\sqrt{NT}$-consistent under assumptions that rule out misclassification in the limit. Dzemski and Okui (2021) study consistency of the estimated coefficients under weaker assumptions. They distinguish between units with $\sigma_i \prec \sqrt{T/\log N}$ that can be classified reliably and noisy units with $\sigma_i \succeq \sqrt{T/\log N}$ that are potentially misclassified in the limit. They show that the *k-means* estimator is $\sqrt{NT}$-consistent if the proportion of noisy units is sufficiently small. In Section 7, we complement this theoretical result by simulating designs where *k-means* misclassifies units but still recovers group-specific coefficients at a rate that is faster than $\sqrt{T}$.

The full-rank condition in Assumption 1.4 ensures that the denominator of $\widehat{D}_i(g, h)$ is not too close to zero. The term in the minimum eigenvalue function is the population long-run covariance matrix of $v_{it}x_{it}$. The assumption restricts v_{it} (which has variance one) but does not limit the magnitude of the error term $\sigma_i v_{it}$ in our panel model (1). It allows conditional heteroskedasticity of v_{it} given x_{it}.

Assumption 1.5 maintains that groups are unique in the sense that there are not two groups that share the same coefficient values. The minimal distance between any two groups is measured by ι_N and is allowed to vanish asymptotically, provided that it satisfies additional rate conditions stated below. Vanishing group separation is an asymptotic modeling device to study settings where groups are distinct but difficult to distinguish. Most existing results that establish asymptotic properties of estimators of the group-specific coefficient assume strict group separation, i.e., that ι_N is bounded away from zero. This makes it difficult to verify Assumption 1.3 if $\iota_N \to 0$. This difficulty can be overcome by using our result for *k-means* estimation in Supplemental Material D, which gives a consistency rate under vanishing group separation.

Assumptions 1.6-1.8 restrict the distribution of the time series $\{x_{it}', w_{it}', v_{it}\}_{t=1}^{T}$. Assumption 1.6 imposes exponential decay of the tails of the marginal distributions. Assumption 1.7 restricts the time-series dependence of the data by imposing exponential decay of the mixing coefficients. This assumption rules out processes with long memory. Assumption 1.8 requires the time series to be stationary. We use the stationarity assumption primarily to show that certain long-run variances are bounded. It can be replaced by other conditions that bound the long-run variances.

Assumption 1.9 rules out that the correlation matrix $\Omega_i(g_i^0)$ contains entries that are too close to negative one. This assumption does not rule out singularity of $\Omega_i(g_i^0)$. Singularity occurs mechanically in our setting whenever $G > p + 1$, where p is the dimension of x_{it}.[8] No restrictions are placed on positive correlations, which is important in settings with vanishing group separation, where groups h and h' have similar coefficients and hence highly positively correlated moment inequalities.[9] Our regularization approach controls positive correlations that are close to one.

We now introduce the last assumption that restricts the choice of kernel function. The validity of this assumption is under complete control of the researcher and does not depend on the underlying data. It is satisfied by the QS kernel (see Andrews 1991).

Assumption 2. *The kernel function $K(\cdot) : \mathbb{R} \to [-1, 1]$ is continuously differentiable with bounded derivatives on $\mathbb{R}$ and satisfies (i) $K(0) = 1$, (ii) $K(x) = K(-x)$ for any $x \in \mathbb{R}$, (iii) $\int_{-\infty}^{\infty} |K(x)|dx < \infty$, and (iv) $K(x) \precsim |x|^{-\vartheta}$ as $|x| \to \infty$ for some constant $\vartheta > 1$.*

The following theorem establishes the validity of our joint confidence set. In the limit, it covers the true group membership at least with pre-specified probability $1 - \alpha$.

[8] Our testing approach is designed to be able to handle singular correlation matrices. In contrast, e.g., the quasi-likelihood ratio statistic used in Kudo (1963) is not defined for singular correlation matrices.

[9] In Supplemental Material D, we develop a theoretical framework based on local alternatives to study settings with very similar groups.

Theorem 2. *Let $\mathbb{P}_N$ be a class of probability measures that satisfy Assumption 1 with identical choices of a, b, d_1, and d_2. Assume that there are finite constants $0 < \delta_1 < \delta_2$ and $1 < k_1 \leq k_2$ such that $T^{\delta_1} \leq N \leq o(1)T^{\delta_2}$ and $(\log N)^{-k_2} \leq \epsilon_N \leq (\log N)^{-k_1}$. Let Assumption 2 hold with $\kappa_N \asymp T^{\rho}$, where $0 < \rho < (\vartheta - 1)/(3\vartheta - 2)$.*[10] *In addition, assume*

$$r_{\theta,N}\sqrt{T \log N} = o\left(1 \wedge \iota_N \min_{1 \leq i \leq N} \sigma_i\right). \tag{7}$$

Then,

$$\liminf_{N,T\to\infty} \inf_{P\in\mathbb{P}_N} P\left(\left\{g_i^0\right\}_{1\leq i\leq N} \in \widehat{C}_\alpha\right) \geq 1 - \alpha.$$

In addition to the aforementioned assumptions, this theorem introduces some rate conditions. The first condition restricts the relative magnitudes of T and N. It accommodates both "short panels" where T is small relative to N and "long panels" where T is large relative to N. The second condition requires the regularization parameter ϵ_N to vanish at a sufficiently slow rate. The third rate condition controls the rate at which the bandwidth sequence κ_N diverges. This rate condition is automatically satisfied for the QS kernel if the bandwidth is chosen by the procedure described in Section 3.3. Finally, condition (7) restricts the rate of convergence of the estimators $\hat{\theta}^w$ and $\hat{\theta}_g$.

We now discuss the latter condition in the context of two examples. For both examples, we assume that σ_i is bounded away from zero uniformly over units i. For the first example, groups are well-separated, i.e., ι_N is a positive constant and does not depend on N. Existing results for clustering with well-separated groups suggest $r_{\theta,N} = (NT)^{-1/2}\zeta_N$ with $\zeta_N \to \infty$ (Bonhomme and Manresa 2015; Su, Shi, and Phillips 2016; Wang, Phillips, and Su 2018).[11] Under this convergence rate, condition (7) is equivalent to $\log N/N = o(1)$ and hence trivially satisfied. For the second example, group separation is vanishing with $\iota_N = T^{-1/2+e}$ for $0 < e < 1/2$. In Theorem D.1 in Supplemental Material D.3, we show $r_{\theta,N} = (NT)^{-1/2}\zeta_N$, under some technical conditions. Then, condition (7) becomes $T^{1-2e}\log N/N = o(1)$. This condition restricts the relative magnitudes of N and T. In particular, the weaker group separation is, i.e., the closer e is to zero, the larger N has to be relative to T. It is sufficient that $N \geq T^{\delta_1}$ with $\delta_1 > 1 - 2e$.

A caveat to the calculations in the previous paragraph is that the results for the rate of consistency of the *k-means* estimator rely on assumptions that rule out misclassification in the limit. In the discussion following Theorem D.1 in Supplemental Material D.3, we

[10] For sequences a_N and b_N we write $a_N \asymp b_N$ if and only if $a_N = O(b_N)$ and $b_N = O(a_N)$.

[11] The sequence ζ_N can go to infinity at any slow rate.

indicate how this limitation can be overcome based on the approach in Dzemski and Okui (2021). However, we leave the formal proof to future research.

Remark 4. *Theorem 2 above and Theorem 3 –Theorem 5 stated below in Section 5 continue to hold if G is replaced by a consistent estimator $\widehat{G}_N$.*

Remark 5. *Our method can be extended to panel models with unit fixed effects by interpreting model (1) as representing the fixed-effect transformed model. This application of our procedure can be theoretically justified but is not covered by the asymptotic results in this section. Extending the results to the fixed-effect model requires some modifications to our arguments, which may be lengthy.*

To derive the asymptotic behavior of the test statistic under the transformed model, we have to examine $-\sigma_i \dot{v}_{it} \dot{x}'_{it}(\theta_g - \theta_h)$, where $\dot{v}_{it} = v_{it} - \sum_{s=1}^T v_{is}/T$ and $\dot{x}_{it} = x_{it} - \sum_{s=1}^T x_{is}/T$. When x_{it} is strictly exogenous, i.e., $\mathbb{E}[v_{it} \mid x_{i1}, \ldots, x_{iT}] = 0$, the orthogonality condition (2) holds also in the transformed model. However, the mixing condition in Assumption 1.7 may not be satisfied. For example, if v_{it} is i.i.d. over time, the correlation between $\dot{v}_{it}$ and $\dot{v}_{is}$ for $t \neq s$ is $1/T$ regardless of the distance between s and t. The mixing condition requires serial correlation between distant time periods to vanish and is therefore not satisfied for $\dot{v}_{it}$ under fixed T as required by Assumption 1.7. Mixing still holds asymptotically as $T \to \infty$, and our proof has to be modified to show that this is sufficient. When x_{it} is merely predetermined, i.e. $\mathbb{E}[v_{it} \mid x_{i1}, \ldots, x_{it}] = 0$, the transformed model is not guaranteed to satisfy (2) because

$$\begin{aligned}\sum_{t=1}^T \mathbb{E}\left[\dot{v}_{it}\dot{x}'_{it}\right] =& \mathbb{E}\left[\sum_{t=1}^T v_{it}x_{it} - T\left(\sum_{s=1}^T v_{is}/T\right)\left(\sum_{s=1}^T x'_{is}/T\right)\right] \\ =& T^{-1}\mathbb{E}\left[\left(\sum_{s=1}^T v_{is}\right)\left(\sum_{s=1}^T x'_{is}\right)\right]\end{aligned}$$

may not be zero. In many cases, including the panel AR(1) model with $x_{it} = y_{i,t-1}$, it holds that $(\sum_{s=1}^T v_{is}/T)(\sum_{s=1}^T x'_{is}/T) = O_P(1/T)$ for each i (see, e.g., Nickell 1981; Hahn and Kuersteiner 2002). The expected uniform order of this term is $O_P(T^{-1}\log N)$. Inspection of the proof of Theorem 2 reveals that a term of this order is asymptotically negligible.

5 Extensions

This section discusses several extensions of our procedure. We first demonstrate the possibility of simplifying the procedure at the cost of power and/or losing robustness against serial correlation. We also propose a two-step method to increase the power of our confidence set.

5.1 A simpler procedure with conservative critical values

The implementation of our confidence set can be greatly simplified by using different critical values that are slightly conservative but can be computed without estimating and regularizing a covariance matrix. We call these the *SNS critical values*, borrowing a term from Chernozhukov, Chetverikov, and Kato (2019) who propose similar critical values for a high-dimensional testing problem and justify them using the theory of self-normalized sums (SNS).

For $G = 2$, the SNS critical values are identical to our critical values. For $G \geq 3$, the SNS critical values are an upper bound to our critical values and will always yield a weakly larger confidence set. The SNS critical values are given by

$$\hat{c}^{\text{SNS}}_{\alpha,N,i}(g) = c^{\text{SNS}}_{\alpha,N} = \sqrt{\frac{T}{T-1}} t^{-1}_{T-1}\left(1 - \frac{\alpha}{(G-1)N}\right)$$

with $t^{-1}_{T-1}(p)$ defined in Section 3.4. The factor $(G-1)$ carries out a Bonferroni correction to account for the $(G-1)$ moment inequalities that are simultaneously tested for unit i. The critical values do not depend on i or g; identical values can be used for all i and g. Let $\widehat{C}^{\text{SNS}}_{\alpha}$ denote the confidence set computed by applying our procedure with SNS critical values. The following result establishes the asymptotic validity of this confidence set.

Theorem 3. *Let $\mathbb{P}_N$ be a class of probability measures that satisfy Assumptions 1.1–1.8 with identical choices of a, b, d_1, and d_2. Assume that there are finite constants $0 < \delta_1 < \delta_2$ such that $T^{\delta_1} \leq N \leq o(1)T^{\delta_2}$. Let Assumption 2 hold with $\kappa_N \asymp T^{\rho}$, where $0 < \rho < (\vartheta - 1)/(3\vartheta - 2)$, and assume that condition* (7) *holds. Then,*

$$\liminf_{N,T\to\infty} \inf_{P\in\mathbb{P}_N} P\left(\left\{g^0_i\right\}_{1\leq i\leq N} \in \widehat{C}^{SNS}_{\alpha}\right) \geq 1 - \alpha.$$

5.2 A simpler procedure under no serial correlation

Another simplification of our procedure is possible in the absence of serial correlation. Suppose that, for each unit i, the time series $\{x_{it}v_{it}\}_{t=1}^{T}$ is serially uncorrelated.

Assumption 3. *For every $i = 1, 2, \ldots$ and all $s, t = 1, 2, \ldots$ such that $s \neq t$, $\mathbb{E}[v_{is}v_{it}x_{is}x_{it}] = 0$.*

Under this assumption

$$\Xi_i(g, h, h') = \operatorname{cov}\left(d_{it}(g, h), d_{it}(g, h')\right)$$

can be consistently estimated by $\widehat{\Xi}_i(g, h, h')$ by setting

$$K\left(\frac{j}{\kappa_N}\right) = \begin{cases} 0 & \text{if } j \neq 0 \\ 1 & \text{if } j = 0. \end{cases} \tag{8}$$

Then $\widehat{\Xi}_i(g, h, h')$ takes the simple form

$$\widehat{\Xi}_i(g, h, h') = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{d}_{it}(g, h) - \bar{\hat{d}}_i(g, h)\right)\left(\hat{d}_{it}(g, h') - \bar{\hat{d}}_i(g, h')\right)$$

and the test statistic is given by

$$\widehat{D}_i(g, h) = \frac{\sum_{t=1}^{T}\hat{d}_{it}(g, h)/\sqrt{T}}{\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\hat{d}_{it}(g, h) - \bar{\hat{d}}_i(g, h)\right)^2}}. \tag{9}$$

Critical values are computed by $\hat{c}_{\alpha,N,i}(g)$ with the simplified version of $\widehat{\Omega}_i(g)$. The following result states the conditions for the validity of the simplified procedure. In contrast to Theorem 2, stationarity is not required.

Theorem 4. *Let $\mathbb{P}_N$ be a class of probability measures that satisfy Assumptions 1.1–1.7 and Assumption 1.9 with identical choices of a, b, d_1, and d_2. Assume that there are finite constants $0 < \delta_1 < \delta_2$ and $1 < k_1 \leq k_2$ such that $T^{\delta_1} \leq N \leq o(1)T^{\delta_2}$ and $(\log N)^{-k_2} \leq \epsilon_N \leq (\log N)^{-k_1}$. Let Assumption 3 and condition (8) hold. In addition, suppose that*

$$r_{\theta,N}\sqrt{T\log N} = o\left(1 \wedge \iota_N \wedge \min_{1\leq i\leq N}\sigma_i\right). \tag{10}$$

Then, $\widehat{C}_\alpha$ *based on* (8) *satisfies*

$$\liminf_{N,T\to\infty} \inf_{P\in\mathbb{P}_N} P\left(\left\{g_i^0\right\}_{1\le i\le N} \in \widehat{C}_\alpha\right) \ge 1-\alpha.$$

Combining the test statistic (9) under no serial correlation with SNS critical values yields a particularly simple procedure that can be implemented with minimal programming effort.

5.3 Increasing power by using a two-step procedure

For units that are very easy to classify (i.e., have very small σ_i), we estimate the true group memberships with probability strictly larger than $1-\alpha/N$. This renders our confidence set conservative. By using the information provided by the units that are easy to classify, we may be able to shrink the marginal confidence sets for the units that are difficult to classify (i.e., have large σ_i). This idea inspires our two-step procedure that we call *unit selection*.

The key part of our two-step procedure is detecting units that are easy to classify. For these units, we report singleton marginal confidence sets. We then compute our joint confidence set on the sub-sample of remaining units. We slightly adjust the nominal level of the confidence set to control for classification error in unit selection. Unit selection can increase the power of the confidence set because it carries out fewer simultaneous tests than our one-step procedure. For example, if unit selection eliminates $N/3$ units, the resulting confidence set is based on Bonferroni correction to adjust for $2N/3$ rather than N simultaneous tests.

For our two-step procedure we assume that $\hat{g}_i$ minimizes squared loss, i.e., we assume

$$\sum_{t=1}^{T} \hat{d}_{it}^{U}(\hat{g}_i, h) \le 0 \quad \text{for all } h \in \mathbb{G}, \tag{11}$$

where

$$\hat{d}_{it}^{U}(g,h) = (y_{it} - w_{it}'\hat{\theta}^w - x_{it}'\hat{\theta}_g)^2 - (y_{it} - w_{it}'\hat{\theta}^w - x_{it}'\hat{\theta}_h)^2.$$

This requirement is automatically satisfied if the grouped panel model is estimated by *k-means* clustering (e.g. Bonhomme and Manresa 2015).

Our algorithm for unit selection identifies a unit as easy to classify if it satisfies two conditions that we call *moment selection* and *hypothesis selection*.[12] A unit i satisfies the moment selection criterion if we detect substantial slackness in inequality (11) for all $h \neq \hat{g}_i$. A unit i satisfies the hypothesis selection criterion if all group memberships $h \neq \hat{g}_i$ are rejected.

The unit selection procedure is parameterized by $\beta \in [0, \alpha/3)$. The larger β, the more unit selection is carried out. Let

$$\widehat{D}_i^U(g,h) = \frac{\sum_{t=1}^T \hat{d}_{it}^U(g,h)/\sqrt{T}}{\sqrt{\widehat{\Xi}_i^U(g,h,h)}},$$

where $\widehat{\Xi}_i^U(g,h,h)$ is defined as $\widehat{\Xi}_i(g,h,h)$ with $\hat{d}_{it}^U$ replacing $\hat{d}_{it}$. $\widehat{D}_i^U(g,h)$ is a counterpart to $\widehat{D}_i(g,h)$ that does not adjust for the mean under the null hypothesis.

The first step of our two-step procedure carries out moment selection by computing the set

$$\widehat{M}_i(g) = \left\{ h \in \mathbb{G} \setminus \{g\} \mid \widehat{D}_i^U(g,h) > -2c_{\beta,N}^{\mathrm{SNS}} \right\}$$

for $g \in \mathbb{G}$ and $i = 1, \ldots, N$. This set gives the selected moment inequalities for the hypothesis $H_0 : g_i^0 = g$. For units i for which $\widehat{M}_i(g)$ is empty, we have strong evidence that $g_i^0 = g$. These units satisfy the moment selection criterion for elimination in the first step. Condition (11) ensures that $\widehat{M}_i(g)$ is never empty for $g \neq \hat{g}_i$. This property ensures that moment selection does not eliminate misclassified units.

The second step of our two-step procedure is given by the following algorithm that carries out hypothesis selection:

Step 2.A: Set $s = 0$ and $H_i(0) = \mathbb{G}$.

Step 2.B: Set $\widehat{N}(s) = \sum_{i=1}^N \max_{g \in H_i(s)} \mathbf{1}\{\widehat{M}_i(g) \neq \emptyset\}$.

Step 2.C: Set

$$H_i(s+1) = \left\{ g \in \mathbb{G} \mid \widehat{T}_i(g) \leq \hat{c}_{\alpha-2\beta, \hat{N}(s), i}(g) \right\} \cup \{\hat{g}_i\}.$$

[12] The term "moment selection" is borrowed from the literature on testing moment inequalities (see, e.g., Chernozhukov, Chetverikov, and Kato 2019; Andrews and Soares 2010; Andrews and Barwick 2012; Romano, Shaikh, and Wolf 2014; Canay and Shaikh 2017). In this literature, moment selection reduces the power loss from possibly slack moment inequalities by identifying inequalities that are "obviously" satisfied. Our use is different. We use moment selection to reduce the power loss from running many simultaneous tests by identifying units that are "obviously" correctly classified.

Step 2.D: If $H_i(s+1) = H_i(s)$ for all i, then exit the algorithm. Otherwise, set $s = s+1$ and go to Step 2.B.

Step 2.A initializes the algorithm by designating all possible group assignments as hypotheses that have to be tested. Step 2.B counts the number $\widehat{N}(s)$ of units that are not easy to classify. Unit i is easy to classify if and only if $\widehat{M_i}(\hat{g}_i)$ is empty (moment selection) and $H_i(s) = \{\hat{g}_i\}$ (hypothesis selection). Step 2.C carries out hypothesis selection with critical values adjusted for $\widehat{N}(s)$ simultaneous tests of group membership. $H_i(s+1)$ gives a preliminary marginal confidence set for unit i after s iterations of hypothesis selection. Step 2.D checks the convergence of the algorithm.[13] If the algorithm has converged after $s = s^*$ iterations, the final joint confidence set is given by

$$\widehat{C}_{\text{sel},\alpha,\beta} = \bigtimes_{1 \le i \le N} H_i(s^*).$$

The second-step confidence set is calculated at nominal confidence level $1 - \alpha + 2\beta$ (see the definition of $H_i(s+1)$ in Step 2.C). The adjustment by 2β represents the cost of unit selection and controls for two possible errors at the first step. The first error is estimating an incorrect group membership for a unit that is easy to classify “in population”. The second error is erroneously declaring a unit as easy to classify.

Unit selection increases the power of the confidence set if its benefits (decreasing the number of units at the second step) outweigh the cost (adjustment of nominal level at the second step). If sufficiently many units can be eliminated, then $\widehat{C}_{\text{sel},\alpha,\beta}$ is more powerful (“smaller”) than the corresponding one-step confidence set $\widehat{C}_\alpha$. If too few units are eliminated, then a two-step confidence set can be slightly more conservative (“larger”) than the corresponding one-step confidence set.

We establish the validity of the two-step procedure under the assumption of no serial correlation (Assumption 3 above) and the following additional assumption.

Assumption 4. *1. The vector $(\theta^{w\prime}, \{\theta_g'\}'_{g\in\mathbb{G}})$ is contained in a compact parameter space Θ.*

2. Let p denote the dimension of x_{it}. For any $k_1, k_2, k_3 = 1, \ldots, p$,

$$\mathbb{E}[\sigma_i v_{it} x_{it,k_1} x_{it,k_2} x_{it,k_3}] = 0.$$

The first part of this assumption is standard. The second part imposes an orthogonality

[13] The algorithm always converges since $\widehat{N}(s)$ and the cardinality of $H_i(s)$ are decreasing in s.

condition on triples of time periods. It is stronger than the assumption of no serial correlation between pairs of time periods but weaker than independence across time. We now state the formal result for the asymptotic validity of the two-step procedure.

Theorem 5. *Let $\mathbb{P}_N$ be a set of probability measures that satisfy Assumptions 1, 3 and 4 with identical choices of a, b, d_1, and d_2. Assume that there are finite constants $0 < \delta_1 < \delta_2$ and $1 < k_1 \leq k_2$ such that $T^{\delta_1} \leq N \leq o(1)T^{\delta_2}$ and $(\log N)^{-k_2} \leq \epsilon_N \leq (\log N)^{-k_1}$. Let Assumption 3 and condition* (8) *hold and $0 < 3\beta < \alpha < 1$. If condition* (10) *is satisfied, then*

$$\liminf_{N,T\to\infty} \inf_{P\in\mathbb{P}_N} P\left(\left\{g_i^0\right\}_{1\leq i\leq N} \in \widehat{C}_{\mathrm{sel},\alpha,\beta}\right) \geq 1-\alpha.$$

6 Empirical application: Heterogeneous effects of a minimum wage

In this section, we illustrate our procedure by revisiting the work of Wang, Phillips, and Su (2019). They estimate a grouped panel model to study heterogeneous effects of a minimum wage in the restaurant sector. Their analysis builds on Dube, Lester, and Reich (2010) who employ a similar panel model but do not allow for effect heterogeneity. We assess the clustering uncertainty of the group memberships estimated in Wang, Phillips, and Su (2019) by computing confidence sets for group memberships.

We use the panel data described in Dube, Lester, and Reich (2010). It contains quarterly data for 1380 US counties, ranging from the first quarter of 1990 to the second quarter of 2006. The grouped panel model estimated in Wang, Phillips, and Su (2019) is given by

$$\log(\texttt{emp}_{ict}) = \theta_{g_i^0,1}\log(\texttt{mw}_{ict}) + \theta_{g_i^0,2}\log(\texttt{pop}_{ict}) + \theta_{g_i^0,3}\log(\texttt{emp}_{ict}^{\texttt{TOT}}) + \phi_c + \tau_t + \sigma_i v_{ict}$$

for state $i = 1,\ldots,51$, county $c = 1,\ldots,n_i$ and time period $t = 1,\ldots,T = 66$, where n_i is the number of counties in state i. The variable $\texttt{emp}_{ict}$ gives employment in the restaurant sector, $\texttt{mw}_{ict}$ gives the minimum wage and $\texttt{emp}_{ict}^{\mathrm{TOT}}$ gives total employment in all sectors. Finally, ϕ_c is a county fixed effect, τ_t is a time fixed effect, and $\sigma_i v_{ict}$ is an idiosyncratic error term.

Su, Shi, and Phillips (2016) propose an information criterion to select the number of groups G in a grouped-panel regression and prove its consistency. Wang, Phillips, and

	$g = 1$		$g = 2$		$g = 3$		$g = 4$	
	coef	se	coef	se	coef	se	coef	se
$\theta_{g,1}$	0.55	0.05	-0.03	0.04	0.06	0.04	-0.25	0.04
$\theta_{g,2}$	0.63	0.07	0.60	0.06	0.34	0.07	0.47	0.06
$\theta_{g,3}$	0.51	0.04	0.61	0.03	0.41	0.03	0.53	0.03

Table 1: Estimates for the group-specific slope coefficients (coef) and corresponding standard errors (se). Standard errors clustered at the county level.

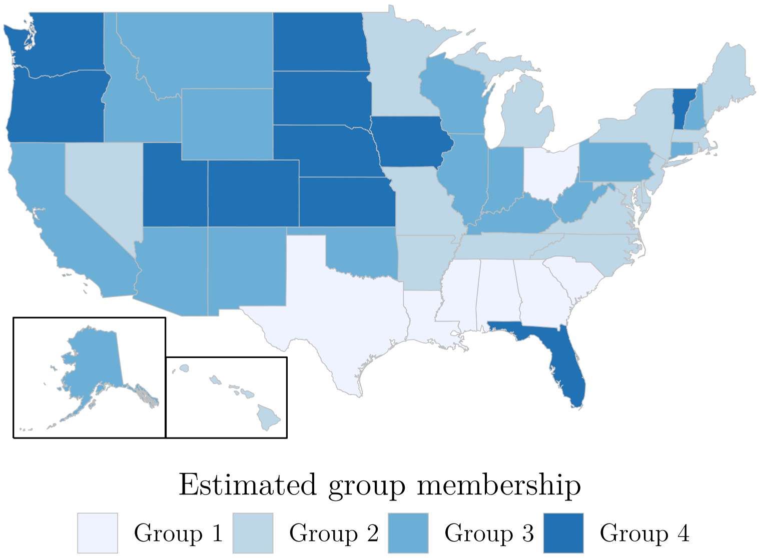

Figure 1: Estimated clusters.

Su (2019) apply this criterion to the data and select $G = 4$. Table 1 gives their CLasso (Su, Shi, and Phillips 2016) estimates for the slope coefficients for the four groups.

There are two groups with estimated positive effects of the minimum wage on employment and two groups with negative effects (see $\theta_{g,1}$ in the table).

Based on these estimates for the slope coefficients, we estimate group memberships by running one update step of the *k-means* algorithm. This updating step ensures that the estimated group memberships satisfy inequality (11) and produces estimated group memberships that are identical to the CLasso estimates in all but six states. Estimated group memberships are displayed in Figure 1 and reported in Table B.1 in the Supplemental Material. Note that Figure 1 is slightly different from Figure 2 in Wang, Phillips, and Su (2019) due to the additional *k-means* step.

To translate the panel model to our framework, we identify states (subscript i) as the cross-sectional dimension and counties and quarters (subscripts c and t) jointly as the second ("time") dimension. We compute our confidence sets based on the fixed-effect transformation

$$\widetilde{\texttt{lemp}}_{ict} = \theta_{g_s^0,1}\widetilde{\texttt{lmw}}_{ict} + \theta_{g_s^0,2}\widetilde{\texttt{lpop}}_{ict} + \theta_{g_s^0,3}\widetilde{\texttt{lemp}}_{ict}^{\texttt{TOT}} + \tilde{\tau}_t + \sigma_i\tilde{v}_{ict}. \tag{12}$$

Here, $\widetilde{\texttt{lemp}}_{ict} = \log(\texttt{lemp}_{ict}) - \sum_{t'=1}^{T}\log(\texttt{lemp}_{ict'})/T$ and $\widetilde{\texttt{lmw}}_{ict}$, $\widetilde{\texttt{lpop}}_{ict}$, and $\widetilde{\texttt{lemp}}_{ict}^{\texttt{TOT}}$ are defined similarly. Moreover, $\tilde{\tau}_t = \tau_t - \sum_{t'=1}^{T}\tau_{t'}/T$ and $\tilde{v}_{ict} = v_{ict} - \sum_{t'=1}^{T}v_{ict'}/T$. Remark 5 above heuristically justifies applying our procedure to fixed-effect-transformed data.

Our second panel dimension comprises both cross-sectional variation between counties and time-series variation between different quarters. To adapt our definition of the estimated long-run variance to this setting, we redefine $\widehat{H}_{ij}$ in (6) as

$$\begin{aligned}&\widehat{H}_{ij}(g,h,h')\\ =&\frac{1}{Tn_i}\sum_{c=1}^{n_i}\sum_{t=|j|+1}^{T}\left(\hat{d}_{ic,t+\min(0,j)}(g,h) - \bar{\hat{d}}_{ic}(g,h)\right)\left(\hat{d}_{ic,t+\min(0,j)}(g,h) - \bar{\hat{d}}_{ic}(g,h')\right),\end{aligned}$$

where $\bar{\hat{d}}_{ic} = \sum_{t=1}^{T}\hat{d}_{ict}/T$ and $\hat{d}_{ict}$ is defined as in Section 3.3 with the double index ct replacing t. We use the following modified algorithm for bandwidth selection:

(A) For $g \in \mathbb{G}$ and $h \in \mathbb{G}\setminus\{g\}$, take each state i such that $\hat{g}_i = g$ and compute $\hat{d}_{ict}(g,h)$ for all $c = 1,\ldots,n_i$ and $t = 1,\ldots,T$. Fit an $AR(1)$-model on $(\hat{d}_{ict}(g,h))_{t=1}^{T}$ and let $\hat{\rho}_{icgh}$ denote the estimated autoregressive coefficients and $\hat{\sigma}_{icgh}^2$ the estimated variance of the innovation.

(B) Estimate the bandwidth sequence by

$$\hat{\kappa}_N = 1.3221\left(T \times \frac{\sum_{i=1}^{N}\sum_{c=1}^{n_i}\sum_{g\in\mathbb{G}}\sum_{h\in\mathbb{G}\setminus\{g\}}\hat{\rho}_{icgh}^2\hat{\sigma}_{icgh}^4/(1-\hat{\rho}_{icgh}^2)^8}{\sum_{i=1}^{N}\sum_{c=1}^{n_i}\sum_{g\in\mathbb{G}}\sum_{h\in\mathbb{G}\setminus\{g\}}\hat{\sigma}_{icgh}^4/(1-\hat{\rho}_{icgh}^2)^4}\right)^{1/5}$$

This algorithm yields $\hat{\kappa}_N = 21.37$.

We compute the joint-confidence set at level $1-\alpha = 0.95$. For the regularization parameter, we set $\epsilon_N = 0.01$. Our results are robust to different choices of ϵ_N.[14]

The full joint confidence set is reported in Table B.1 in the Supplemental Material. Marginal confidence sets for a subset of states are given in Table 2. The cardinality of

[14] For $\epsilon_N = 0, 0.05$, we obtain a confidence set that differs only in the group assignments for Kentucky. For $\epsilon_N = 0.01, 0.05$, group $g = 2$ is contained in the marginal confidence set for Kentucky (p-values 0.0662 and 0.0610). For $\epsilon = 0$, it is not (p-value 0.0372).

		baseline			SNS		
State	$\hat{g}_i$	p-val $\hat{g}_i$	card	CS	p-val $\hat{g}_i$	card	CS
Arkansas	2	0.027	1	2	0.041	1	2
Colorado	4	0.893	2	3, 4	1.000	2	3, 4
Connecticut	3	0.631	2	2, 3	0.715	2	2, 3
Florida	4	0.049	1	4	0.074	2	3, 4
Idaho	3	0.664	2	3, 4	0.852	2	3, 4
Indiana	3	0.024	1	3	0.038	1	3
Kansas	4	0.010	1	4	0.015	1	4
Kentucky	3	0.093	2	3, 4	0.151	3	2, 3, 4
Maine	2	0.015	1	2	0.023	1	2
New Hampshire	3	0.080	2	3, 4	0.130	2	3, 4
New Mexico	3	0.094	3	2, 3, 4	0.121	3	2, 3, 4
Oklahoma	3	0.168	2	2, 3	0.207	2	2, 3
Pennsylvania	3	0.021	1	3	0.030	1	3
West Virginia	3	0.014	1	3	0.041	1	3
Wisconsin	3	0.392	2	2, 3	0.455	2	2, 3

Table 2: Marginal confidence set at level $1-\alpha = 0.95$ for the states with group membership that are estimated with p-value between 1% and 90%. "p-val $\hat{g}_i$" is the p-value for the significance of the estimated group membership. "CS cardinality" is the cardinality of the marginal confidence set for the state. "CS" is the marginal confidence set. "Baseline" refers to the procedure with critical values defined in Section 3.4. "SNS" refers to the procedure with critical values defined in Section 5.1.

the state-wise marginal confidence sets is four for three states, three for seven states, two for twelve states, and one for twenty-nine states.

Based on our joint confidence set, we compute p-values for the significance of the estimated group memberships. We say that the estimated group membership $\hat{g}_i$ for state i is significant at level α if the marginal confidence set for state i with confidence level $1 - \alpha$ contains only $\hat{g}_i$. Up to a joint failure probability of at most α, significantly estimated group memberships reveal the true group membership and cannot be attributed to estimation error. The p-value for significance of $\hat{g}_i$ is the smallest value of α such that $\hat{g}_i$ is significant at level α. The p-values for a subset of states are reported in Table 2, and results for all states are reported in Table B.1 in the Supplemental Material.

Table 2 demonstrates that, for our sample, the SNS critical values from Section 5.1 yield slightly larger confidence sets than our baseline procedure. For example, the cardinality of the marginal confidence set for Florida increases from 1 to 2 (p-value increases from 0.049 to 0.074) when we use the SNS critical values.

Displaying a visual representation of the estimated clusters as we do in Figure 1 is standard practice even in applications where the clustering structure is a nuisance parameter and not of interest in its own right. Visual inspection of the clusters is meant to confirm their economic plausibility and serves as an informal test of model specification. A graphical representation of the clustering uncertainty detected by our confidence set, as illustrated in Figure 2, can complement such an informal analysis. The upper panel in Figure 2 represents clustering uncertainty by shading US states according to the p-values of their estimated group memberships. The lower panel shades states by the cardinalities of their marginal confidence sets ($1 - \alpha = 0.95$). The figure suggests a low degree of clustering uncertainty. It would suggest a large degree of clustering uncertainty if the maps were shaded mostly in a dark hue. Large clustering uncertainty indicates that the clustering algorithm may be overfitting on the sample, rather than picking up structural heterogeneity.

Our two-step procedure with parameter $\beta = \alpha/5 = 0.01$ eliminates only one unit in the first step and does not improve the final confidence set or p-values. In Supplemental Material B.2, we apply the two-step procedure under the implausible assumption of no serial correlation which eliminates more units and lowers the final p-values.

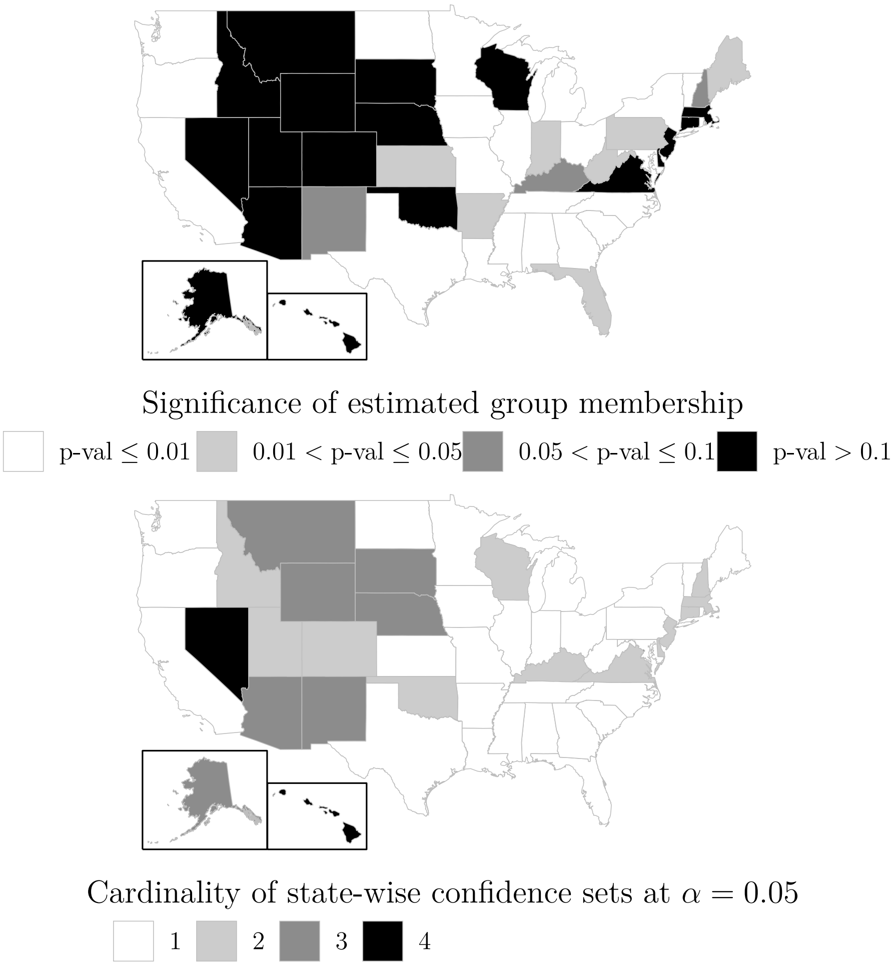
Significance of estimated group membership
p-val ≤ 0.01
0.01 < p-val ≤ 0.05
0.05 < p-val ≤ 0.1
p-val > 0.1
Cardinality of state-wise confidence sets at α = 0.05
1
2
3
4

Figure 2: Visual representation of clustering uncertainty. The upper panel shows p-values for estimated group memberships. The lower panel shows the cardinality of the marginal confidence sets with $1 - \alpha = 0.95$.

7 Simulations

Our simulation designs are based on our empirical application. The data-generating process is given by

$$\widetilde{\texttt{lemp}}_{it} = \theta_{g_i^0,1}\widetilde{\texttt{lmw}}_{it} + \theta_{g_i^0,2}\widetilde{\texttt{lpop}}_{it} + \theta_{g_i^0,3}\widetilde{\texttt{lemp}}_{it}^{\text{TOT}} + \sigma_i v_{it}$$

for $i = 1, \ldots, N$ and $t = 1, \ldots, T$. For $g = 1, \ldots, 4$, the coefficients $\theta_{g,1}$, $\theta_{g,2}$ and $\theta_{g,3}$ are set equal to the estimated coefficients in Table 1.

To generate $x_{it} = (\widetilde{\texttt{lmw}}_{it}, \widetilde{\texttt{lpop}}_{it}, \widetilde{\texttt{lemp}}_{it}^{\text{TOT}})$ that exhibit similar patterns of serial correlation as the time series in our empirical application, we estimate a VAR(1) model of x_{it} for each county c in our data sample and save the estimated coefficient $\hat{\Gamma}_c$ and the empirical residuals $\hat{e}_{ct}$. We then average over all county-wise coefficients to obtain a single VAR(1) coefficient matrix $\bar{\Gamma}$. To generate a time series $(x_{it})_{t=1}^T$, we sample the initial value x_{i0} from the covariates observed in our data and then iteratively generate $x_{i,t+1} = \bar{\Gamma}x_{it} + e_t$, where e_t is a randomly drawn empirical residual $\hat{e}_{ct}$.

Each unit i is assigned to one of the four groups with equal probability and assigned a heteroscedasticity parameter σ_i. For $\sigma = 0.1, 0.2$, we draw $\sigma_i = \sigma\chi^2(4)/4$, where $\chi^2(df)$ is a random draw from a χ^2-distribution with df degrees of freedom.

For $\rho = 0, 0.25, 0.5$, we set $v_{it} = \sqrt{1-\rho^2}\tilde{v}_{it}$ where $\tilde{v}_{it}$ follows an $AR(1)$ process with autoregressive parameter ρ and standard normal innovations and $\tilde{v}_{i0}$ is drawn from the stationary distribution of $(\tilde{v}_{it})_{t=1}^T$. To introduce relevant serial correlation, we require both v_{it} and x_{it} to be serially correlated (cf. Section 5.2). Therefore, setting $\rho = 0$ turns off the serial correlation (but not all temporal dependence) in the moment inequalities even though x_{it} is still serially correlated.

We simulate panels of size $N = 50, 100, 200$ and $T = 60, 120$. The lower values of these ranges are in the ballpark of the number of states (51 states) and time periods (66 quarters) observed in our application. We set the regularization parameter for covariance estimation to $\epsilon_N = 0.01$. In Supplemental Material C.1, we demonstrate that our results are robust to alternative specifications of ϵ_N. We simulate both the oracle confidence set, for which we take the true values of the group-specific coefficients as given, and the confidence set with group-specific coefficients estimated by the *k-means* estimator. We steer the *k-means* algorithm towards recovering the population labels of the groups by using the true group memberships as initial values. The simulation results are based on 500 replications and reported in Table 3.[15]

[15] Our Monte Carlo simulations were carried out on computing resources of the Swedish National

					coverage				cardinality			bandwidth	
					kmeans			oracle	kmeans		oracle	kmeans	oracle
ρ	σ	N	T	$\hat{\theta}-\theta$	hac	h	naive	hac	hac	h	hac	hac	hac
0.00	0.1	50	60	0.18	0.99	0.99	0.21	0.99	1.60	1.27	1.57	2.12	2.04
			120	0.13	1.00	1.00	0.83	1.00	1.07	1.02	1.07	2.27	2.27
		100	60	0.13	0.99	0.98	0.04	1.00	1.76	1.30	1.75	2.20	2.19
			120	0.09	1.00	1.00	0.80	1.00	1.12	1.02	1.11	2.32	2.36
		200	60	0.09	0.99	0.98	0.00	0.98	1.95	1.34	1.95	2.30	2.30
			120	0.07	1.00	0.99	0.60	1.00	1.17	1.03	1.18	2.39	2.38
	0.2	50	60	0.40	0.92	0.92	0.01	0.98	1.97	1.62	1.95	2.11	2.05
			120	0.26	1.00	0.99	0.65	0.99	1.16	1.08	1.14	2.23	2.23
		100	60	0.30	0.95	0.93	0.00	0.99	2.13	1.67	2.14	2.18	2.20
			120	0.18	1.00	0.99	0.36	0.98	1.21	1.08	1.20	2.35	2.34
		200	60	0.22	0.96	0.96	0.00	0.97	2.30	1.74	2.31	2.27	2.31
			120	0.13	0.98	1.00	0.14	1.00	1.28	1.09	1.27	2.36	2.39
0.25	0.1	50	60	0.22	0.94	0.85	0.10	0.97	2.40	1.26	2.38	3.73	3.71
			120	0.15	1.00	0.99	0.82	1.00	1.73	1.02	1.75	4.14	4.15
		100	60	0.15	0.96	0.87	0.01	0.97	2.68	1.30	2.71	3.85	3.87
			120	0.11	1.00	0.97	0.65	1.00	2.23	1.02	2.29	4.21	4.29
		200	60	0.11	0.94	0.84	0.00	0.96	3.03	1.33	3.05	3.99	4.02
			120	0.08	1.00	0.96	0.44	1.00	2.89	1.03	2.88	4.37	4.37
	0.2	50	60	0.51	0.84	0.64	0.00	0.94	2.63	1.60	2.69	3.59	3.67
			120	0.31	0.99	0.94	0.49	1.00	1.87	1.08	1.85	4.06	4.11
		100	60	0.36	0.88	0.64	0.00	0.95	2.96	1.67	2.97	3.78	3.82
			120	0.22	0.99	0.92	0.21	0.99	2.25	1.08	2.38	4.13	4.29
		200	60	0.27	0.89	0.64	0.00	0.91	3.24	1.71	3.24	4.00	3.98
			120	0.15	0.99	0.93	0.04	0.99	2.95	1.09	3.02	4.33	4.42
0.50	0.1	50	60	0.26	0.87	0.47	0.03	0.95	3.35	1.25	3.39	6.46	6.55
			120	0.17	0.99	0.90	0.68	0.99	3.71	1.02	3.71	7.21	7.21
		100	60	0.18	0.86	0.32	0.00	0.90	3.52	1.28	3.55	6.45	6.50
			120	0.13	0.99	0.85	0.50	1.00	3.80	1.02	3.81	7.42	7.46
		200	60	0.13	0.82	0.19	0.00	0.84	3.67	1.31	3.67	6.93	6.90
			120	0.09	1.00	0.77	0.26	1.00	3.84	1.03	3.84	7.68	7.70
	0.2	50	60	0.64	0.69	0.21	0.00	0.88	3.42	1.55	3.49	6.14	6.34
			120	0.36	0.98	0.74	0.30	0.98	3.72	1.07	3.74	7.16	7.19
		100	60	0.43	0.69	0.11	0.00	0.84	3.62	1.61	3.66	6.50	6.65
			120	0.26	0.98	0.62	0.09	0.99	3.84	1.08	3.84	7.35	7.48
		200	60	0.32	0.70	0.02	0.00	0.76	3.74	1.67	3.75	6.80	6.88
			120	0.18	0.97	0.48	0.00	0.98	3.87	1.09	3.87	7.62	7.66

Table 3: Simulation results. Nominal level $1-\alpha = 0.95$. "$\hat{\theta}-\theta$" is defined in the text. "hac" uses the data-driven bandwidth, "h" sets the bandwidth equal to zero, and "naïve" is the naïve confidence set defined in the text. "kmeans" uses the *k-means* algorithm to compute group-specific coefficients, and "oracle" uses the true coefficients. "coverage" is the empirical coverage probability of the joint confidence set. "cardinality" is the Monte Carlo average of the average (over all units) cardinality of the marginal unit-wise confidence sets. "bandwidth" is the Monte Carlo average of chosen bandwidth.

We first discuss the empirical coverage probability of our confidence set with *k-means* estimates and a data-driven bandwidth (coverage→kmeans→hac in Table 3). The joint confidence set has a coverage probability of at least $1 - \alpha = 95\%$ in most cases. This is in line with our theoretical result on the asymptotic validity of the confidence set. When $T = 60$, our confidence set under-covers for some designs. In these designs, we compute a large confidence set. This result indicates that our confidence set can detect that statistical uncertainty is large even when it is under-covering. In the designs with moderate serial correlation ($\rho = 0.25$), under-coverage can be alleviated by increasing the number of cross-sectional observations, keeping the number of time series observations constant. In the designs with large serial correlation ($\rho = 0.5$), improving coverage requires increasing the number of time-series observations. This result may be explained by the fact that our algorithm for bandwidth selection mechanically chooses a small bandwidth when the time series is short. With a small bandwidth, our procedure cannot control for temporal dependence over many time periods.

In some of our designs, our confidence set is conservative, with an empirical coverage probability of up to close to one. Since group membership is a discrete parameter, this is not necessarily a sign that the confidence set is underpowered. To see this, consider the naïve confidence set that contains only the vector of estimated group memberships. This is the smallest possible confidence set. As T increases, true group memberships are revealed with increasing probability, and even the naïve confidence set can become conservative eventually (see Section 5.3). The coverage probability of the naïve confidence set (coverage→kmeans→naïve in Table 3) gives the probability that data-driven clustering recovers the true group structure. In our designs, this probability varies between 0% and 83%.

We now turn to the power of our confidence set with *k-means* estimates and a data-driven bandwidth. The simulated average cardinality of a unit-wise marginal confidence set is reported in Table 3 under cardinality→kmeans→hac. Increasing σ or ρ makes the data more noisy as time-series shocks become larger or more persistent. As is expected, the power of our test decreases, and the size of our confidence set increases as the data become more noisy.

In the less noisy designs, the confidence set is highly informative. For example, if $\rho = 0$ and $\sigma = 0.1$, the confidence set rules out 2-3 out of 4 possible group assignments for most units. Our confidence set is only slightly larger than the naïve set (i.e., the set containing only the vector of estimated group memberships) in the designs where the

Infrastructure for Computing (SNIC), funded by Swedish Research Council grant 2018-05973.

naïve set performs well. The naïve set still under-covers, and some enlargement improves the coverage.

In the noisiest designs, our confidence set becomes quite uninformative and assigns the trivial marginal confidence set (all four possible groups) to most units.

Our asymptotic result is valid under a high-level assumption about the rate at which the estimator $\hat{\theta} = \{\hat{\theta}_g'\}'_{g\in\mathbb{G}}$ converges to the true group-specific coefficients $\theta = \{\theta_g'\}'_{g\in\mathbb{G}}$. This rate may be affected by misclassification. In this section, we study the effect of model estimation in finite samples, focusing on misclassification driven by the noise variance σ^2. Additional results for settings where model estimation is potentially affected by weak group separation are reported in Supplemental Material D.

The column "$\hat{\theta}-\theta$" in Table 3 quantifies estimation error in $\hat{\theta}$ and reports the simulated value of

$$\frac{||\hat{\theta}-\theta||}{||\theta||} = \sqrt{\sum_{g=1}^{4}\mathbb{E}||\hat{\theta}_g-\theta_g||_2^2}\Bigg/\sqrt{\sum_{g=1}^{4}||\theta_g||_2^2}\,,$$

where $||\cdot||_2$ is the L_2-norm. The simulated estimation error is not negligible, varying between 7% and 64% of the magnitude of the true coefficient vector. In all our designs, increasing T keeping N fixed or increasing N keeping T fixed decreases estimation error. This shows that, even though states are misclassified, the *k-means* estimator can exploit both time-series and cross-sectional variation. This is necessary for fulfilling the rate condition imposed in our asymptotic result.

As a more direct test of the effect of parameter estimation, we compare the confidence set using the true coefficients (oracle→hac in Table 3) to the confidence set using *k-means* estimates (kmeans→hac in Table 3). In many designs, both confidence sets are valid with a coverage probability of at least 95%. In most designs where the oracle confidence set has coverage of at least 95%, the confidence set with *k-means* estimates is valid or undercovers only slightly (1-3%). For all designs, increasing either N or T decreases the difference in coverage probability between the oracle confidence set and the confidence set using the *k-means* estimates. This simulation evidence aligns with our theoretical argument that the asymptotic effect of parameter estimation is not of first order.

Next, we compare our benchmark procedure based on the heteroskedasticity-and-autocorrelation-robust (HAC) variance estimator ("hac" in results table) and the alternative procedure discussed in Section 5.2 with variance estimation that is robust to heteroscedasticity but not auto-correlation ("h" in results table).

In the designs with no serial correlation ($\rho = 0$), the confidence set using the heteroscedasticity-robust variance estimator is valid. It covers the true group structure at least with probability $1 - \alpha = 95\%$ when the sample size is sufficiently large. In these designs, choosing the heteroscedasticity-robust estimator over the HAC estimator increases the power of the confidence set. If $T = 120$, the power gain is modest. If $T = 60$, the power gain can be quite large. For example, in the design with $\rho = 0$, $\sigma = 0.1$, $N = 200$, and $T = 60$, the average cardinality of a unit-wise confidence set is 1.95 for the HAC estimator and 1.34 for the heteroscedasticity-robust estimator. This power gain points to the inherent difficulty of estimating long-run variances for a high-dimensional number of relatively short time series.

For the settings with serial correlation ($\rho = 0.25, 0.5$), the confidence set using the variance estimator that is only robust to heteroscedasticity is not valid. In the designs with a lot of serial correlation ($\rho = 0.5$), it under-covers severely. For example, for $\rho = 0.5$, $\sigma = 0.2$, $N = 200$, $T = 120$, its coverage probability is only 48%, even though the sample size is fairly large. Our benchmark confidence set based on the HAC estimator set has correct coverage in this design.

An interpretation of our confidence set is that it distinguishes low-noise units for which the clustering algorithm reveals the true group memberships from noisy units for which group membership is uncertain. Our model and simulation designs parameterize the unit-specific noise-level by the latent parameter σ_i. Our confidence set picks up the latent heteroscedasticity and assigns, on average, a small marginal confidence set for a unit i with small σ_i and a large marginal confidence set for a unit i with large σ_i. This property is demonstrated in Table 4, where we report the average cardinality of the marginal confidence set for units with different magnitudes of σ_i. For example, in the first reported design ($\rho = 0$, $\sigma = 0.1$, $N = 50$, $T = 60$), the average cardinality for the units in the lowest quintile of the distribution of σ_i is 1.31. The average cardinality for the units in the highest quintile of the distribution of σ_i is 2.07.

8 Conclusion

We have constructed a confidence set for group membership for grouped panel models with time-invariant group-specific regression curves. Our confidence set can be easily tabulated and visualized as demonstrated in our empirical application. Empirical researchers can use our confidence set to quantify the statistical uncertainty about the true group structure in a grouped panel model.

				average cardinality					
ρ	σ	N	T	all	0-20perc	20-40perc	40-60perc	60-80perc	80-100perc
0.00	0.1	50	60	1.60	1.31	1.40	1.52	1.70	2.07
			120	1.07	1.03	1.04	1.05	1.07	1.15
		100	60	1.76	1.45	1.56	1.67	1.87	2.24
			120	1.12	1.08	1.08	1.10	1.13	1.21
		200	60	1.95	1.63	1.75	1.88	2.08	2.43
			120	1.17	1.12	1.12	1.16	1.18	1.28
	0.2	50	60	1.97	1.41	1.62	1.91	2.21	2.67
			120	1.16	1.05	1.08	1.13	1.18	1.37
		100	60	2.13	1.55	1.82	2.07	2.38	2.84
			120	1.21	1.07	1.11	1.17	1.25	1.45
		200	60	2.30	1.70	2.00	2.25	2.53	3.00
			120	1.28	1.14	1.18	1.23	1.32	1.55
0.25	0.1	50	60	2.40	2.08	2.26	2.34	2.50	2.80
			120	1.73	1.64	1.68	1.69	1.75	1.86
		100	60	2.68	2.40	2.53	2.64	2.79	3.05
			120	2.23	2.15	2.18	2.19	2.26	2.35
		200	60	3.03	2.81	2.90	3.00	3.13	3.32
			120	2.89	2.84	2.85	2.87	2.90	2.97
	0.2	50	60	2.63	2.11	2.36	2.59	2.85	3.24
			120	1.87	1.69	1.77	1.84	1.92	2.16
		100	60	2.96	2.52	2.74	2.93	3.15	3.44
			120	2.25	2.09	2.14	2.21	2.28	2.52
		200	60	3.24	2.89	3.08	3.23	3.40	3.61
			120	2.95	2.82	2.87	2.93	2.99	3.12
0.50	0.1	50	60	3.35	3.18	3.30	3.34	3.40	3.53
			120	3.71	3.68	3.70	3.71	3.72	3.74
		100	60	3.52	3.40	3.45	3.51	3.58	3.67
			120	3.80	3.76	3.79	3.80	3.81	3.86
		200	60	3.67	3.57	3.62	3.65	3.71	3.78
			120	3.84	3.81	3.82	3.82	3.85	3.88
	0.2	50	60	3.42	3.16	3.30	3.42	3.54	3.67
			120	3.72	3.67	3.69	3.73	3.74	3.79
		100	60	3.62	3.43	3.55	3.63	3.70	3.80
			120	3.84	3.79	3.81	3.84	3.86	3.90
		200	60	3.74	3.60	3.68	3.75	3.80	3.87
			120	3.87	3.81	3.84	3.86	3.89	3.93

Table 4: Simulated average cardinality of the marginal confidence sets by unit σ_i. Nominal level $1 - \alpha = 0.95$. 0-20perc refers to the units with a σ_i-value that lies between the 0 and 20 percentile of the distribution of $\sigma\chi^2(4)/4$. 20-40perc, 40-60perc, 60-80perc and 80-100perc are defined similarly.

Our method extends naturally to other models with a latent group structure. We construct our confidence set based on a test for the best fit in a least-squares sense. This idea can be adapted to a wide variety of settings with possibly different notions of what constitutes a best fit. For example, it may be possible to compute a confidence set for group membership in non-linear likelihood-based models (Liu et al. 2020; Wang and Su 2021) by testing group membership based on the fit measured by the log-likelihood function. This and other extensions of our method require new and non-trivial theoretical work and are interesting avenues for future research.

Our approach can be generalized to models with time-varying coefficients $\theta_g = \theta_{g,t}$, including models with time-varying intercepts as in Bonhomme and Manresa (2015). We studied this extension in a previous version of this paper (Dzemski and Okui 2018). Our present focus on time-invariant coefficients is motivated by the literature (see, e.g., Su, Shi, and Phillips 2016; Wang, Phillips, and Su 2018; Vogt and Linton 2017) and theoretical considerations. The conditions for establishing the validity of our method for models with time-varying coefficients are less transparent and substantially more restrictive. Intuitively, time-varying coefficients are identified purely from cross-sectional variation and are, therefore, estimated at a slower rate than time-invariant coefficients. Therefore, stronger rate conditions requiring at least $T \log N/N \to 0$ are needed to control estimation error in the time-varying coefficients.

Our confidence set is tailored to address research questions where unit identities are relevant. For example, in our application, we can identify (at a pre-specified confidence level) a set of US states that exhibit a positive effect of the minimum wage on employment. Identifying such states is relevant for conducting further research or implementing targeted policies. Our method cannot be directly applied to assess the effect of misclassification on statistics that average over units without regard to unit labels. One example of such a statistic is the estimator of the group-specific slope coefficients. Developing a theory tailored to averages is an interesting research agenda that is complementary to our work.

References

Ando, Tomohiro and Jushan Bai (2016). "Panel data models with grouped factor structure under unknown group membership". In: *Journal of Applied Econometrics* 31.1, pp. 163–191.

Andrews, Donald W. K. (1991). "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation". In: *Econometrica* 59.3, pp. 817–858.

Andrews, Donald W. K. and Panle Jia Barwick (2012). “Inference for parameters defined by moment inequalities: A recommended moment selection procedure”. In: *Econometrica* 80.6, pp. 2805–2826.

Andrews, Donald WK and Gustavo Soares (2010). “Inference for parameters defined by moment inequalities using generalized moment selection”. In: *Econometrica* 78.1, pp. 119–157.

Azzalini, Adelchi and Alan Genz (2016). *The R package* `mnormt`*: The multivariate normal and t distributions (version 1.5-5).* URL: `http://azzalini.stat.unipd.it/SW/Pkg-mnormt`.

Bonhomme, Stéphane and Elena Manresa (2015). “Grouped patterns of heterogeneity in panel data”. In: *Econometrica* 83.3, pp. 1147–1184.

Canay, Ivan and Azeem Shaikh (2017). “Practical and theoretical advances in inference for partially identified models”. In: *Advances in Economics and Econometrics: Eleventh World Congress.* Ed. by Bo Honoré et al. Vol. 2. Econometric Society Monographs. Cambridge University Press, pp. 271–306.

Chang, Jinyuan, Xiaohui Chen, and Mingcong Wu (2023). “Central limit theorems for high dimensional dependent data”. In: *Bernoulli* Forthcoming.

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato (2017). “Central limit theorems and bootstrap in high dimensions”. In: *The Annals of Probability* 45.4, pp. 2309–2352.

— (2019). “Inference on causal and structural parameters using many moment inequalities”. In: *Review of Economic Studies* 86, pp. 1867–1900.

Chetverikov, Denis and Elena Manresa (2022). “Spectral and post-spectral estimators for grouped panel data models”. In: *arXiv preprint arXiv:2212.13324*.

Dempster, Arthur, Nan Laird, and Donald Rubin (1977). “Maximum likelihood from incomplete data via the EM algorithm”. In: *Journal of the Royal Statistical Society. Series B (methodological)*, pp. 1–38.

Dube, Arindrajit, T. William Lester, and Michael Reich (2010). “Minimum wage effects across state borders: Estimates using contiguous counties”. In: *The Review of Economics and Statistics* 92.4, pp. 945–964.

Dzemski, Andreas and Ryo Okui (2018). *Confidence set for group membership.* arXiv: `1801.00332v3`.

— (2021). “Convergence rate of estimators of clustered panel models with misclassification”. In: *Economics Letters* 203, p. 109844.

Genz, Alan (1992). “Numerical computation of multivariate normal probabilities”. In: *Journal of Computational and Graphical Statistics* 1.2, pp. 141–149.

Grayling, Michael J and Adrian Mander (2016). “MVTNORM: Stata module to work with the multivariate normal and multivariate t distributions”. In: *Statistical Software Components*.

Gu, Jiaying and Stanislav Volgushev (2019). “Panel data quantile regression with grouped fixed effects”. In: *Journal of Econometrics* 213.1, pp. 68–91.

Hahn, Jinyong and Guido Kuersteiner (2002). “Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and T are large”. In: *Econometrica* 70.4, pp. 1639–1657.

Hahn, Jinyong and Hyungsik Roger Moon (2010). “Panel data models with finite number of multiple equilibria”. In: *Econometric Theory* 26.03, pp. 863–881.

Kudo, Akio (1963). “A multivariate analogue of the one-sided test”. In: *Biometrika* 50.3/4, pp. 403–418.

Li, Wenbo V and Qi-Man Shao (2002). “A normal comparison inequality and its applications”. In: *Probability Theory and Related Fields* 122.4, pp. 494–508.

Lin, Chang-Ching and Serena Ng (2012). “Estimation of panel data models with parameter heterogeneity when group membership is unknown”. In: *Journal of Econometric Methods* 1.1, pp. 42–55.

Liu, Ruiqi et al. (2020). “Identification and estimation in panel models with overspecified number of groups”. In: *Journal of Econometrics* 215.2, pp. 574–590.

Lu, Xun and Liangjun Su (2017). “Determining the number of groups in latent panel structures with an application to income and democracy”. In: *Quantitative Economics* 8.3, pp. 729–760.

Mammen, Enno, Ralf A. Wilke, and Kristina Zapp (2022). “Estimation of Group Structures in Panel Models with Individual Fixed Effects”. ZEW - Centre for European Economic Research Discussion Paper No. 22-023.

McLachlan, Geoffrey and David Peel (2004). *Finite mixture models*. John Wiley & Sons.

Mehrabani, Ali (2022). “Estimation and identification of latent group structures in panel data”. In: *Journal of Econometrics*.

Moon, Hyungsik Roger and Martin Weidner (2023). “Nuclear Norm Regularized Estimation of Panel Regression Models”. https://arxiv.org/abs/1810.10987.

Mugnier, Martin (2022). “A simple and computationally trivial estimator for grouped fixed effects models”. Working paper.

Mugnier, Martin (2023). “Unobserved clusters of time-varying heterogeneity in nonlinear panel data models”. Working paper.

Newey, Whitney K. and Kenneth D. West (1987). “A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix”. In: *Econometrica* 55.3, pp. 703–708.

Nickell, Stephen (1981). “Biases in dynamic models with fixed effects”. In: *Econometrica* 49.6, pp. 1417–1426.

Pearson, Karl (1896). “Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia”. In: *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character* 187, pp. 253–318.

Romano, Joseph P, Azeem M Shaikh, and Michael Wolf (2014). “A Practical Two-Step Method for Testing Moment Inequalities”. In: *Econometrica* 82.5, pp. 1979–2002.

Sarafidis, Vasilis and Neville Weber (2015). “A partially heterogeneous framework for analyzing panel data”. In: *Oxford Bulletin of Economics and Statistics* 77.2, pp. 274–296.

Su, Liangjun, Zhentao Shi, and Peter Phillips (2016). “Identifying latent structures in panel data”. In: *Econometrica* 84.6, pp. 2215–2264.

Vogt, Michael and Oliver Linton (2017). “Classification of nonparametric regression functions in heterogeneous panels”. In: *Journal of the Royal Statistical Society: Series B* 79 (1), pp. 5–27.

Vogt, Michael and Matthias Schmid (2021). “Clustering with statistical error control”. In: *Scandinavian Journal of Statistics* 48, pp. 729–760.

Wang, Wuyi, Peter C. B. Phillips, and Liangjun Su (2018). “Homogeneity pursuit in panel data models: theory and applications”. In: *Journal of Applied Econometrics* 33, pp. 797–815.

— (2019). “The heterogeneous effects of the minimum wage on employment across states”. In: *Economics Letters* 174, pp. 179–185.

Wang, Wuyi and Liangjun Su (2021). “Identifying latent group structures in nonlinear panels”. In: *Journal of Econometrics* 220.2, pp. 272–295.

Yu, Lu, Jiaying Gu, and Stanislav Volgushev (2023). “Spectral clustering with variance information for group structure estimation in panel data”. Working Paper.