Mammen, Enno; Wilke, Ralf A.; Zapp, Kristina Maria

Working Paper

Estimation of group structures in panel models with individual fixed effects

ZEW Discussion Papers, No. 22-023

Provided in Cooperation with:
ZEW - Leibniz Centre for European Economic Research

Suggested Citation: Mammen, Enno; Wilke, Ralf A.; Zapp, Kristina Maria (2022) : Estimation of group structures in panel models with individual fixed effects, ZEW Discussion Papers, No. 22-023, ZEW - Leibniz-Zentrum für Europäische Wirtschaftsforschung, Mannheim

This Version is available at:
https://hdl.handle.net/10419/261375

// NO.22-023 | 06/2022

DISCUSSION PAPER

// ENNO MAMMEN, RALF A. WILKE, AND KRISTINA ZAPP

Estimation of Group Structures in Panel Models With Individual Fixed Effects

Leibniz Association

ZEW

Estimation of Group Structures in Panel Models with Individual Fixed Effects

Enno Mammen*, Ralf A. Wilke†, Kristina Zapp‡

June 23, 2022

Abstract

The fixed effects (FE) panel model is one of the main econometric tools in empirical economic research. A major practical limitation is that the parameters on time-constant covariates are not identifiable. This paper presents a new approach to grouping FE in the linear panel model to reduce their dimensionality and ensure identifiability. By using unsupervised nonparametric density based clustering, cluster patterns including their location and number are not restricted. The approach works with large data structures (units and groups) and only clusters units that are sufficiently similar, while leaving others as unclustered atoms. Asymptotic theory and rates of convergence are presented. With the help of simulations and an application to economic data it is shown that the suggested method performs well and gives more insightful and efficient results than conventional panel models.

Keywords: Panel Data, Statistical Learning, Regularisation, Endogeneity
JEL: C14, C23, C38

*Heidelberg University, Institute for Applied Mathematics, E–mail: mammen@math.uni-heidelberg.de

†Copenhagen Business School, Department of Economics and ZEW Mannheim, E–mail: rw.eco@cbs.dk

‡ZEW Mannheim; E–mail: kristina.zapp@zew.de

We thank Christina Gathmann, François Laisney and Sebastian Siegloch for insightful feedback and participants at CFE-CMStatistics 2019 (London), 2nd Workshop "Machine Learning in Labor, Education, and Health Economics" (online) and the internal seminars at ZEW Mannheim and CBS Copenhagen for comments. We thank the ZEW Mannheim for support. We thank Sarah McNamara for proofreading and Dennis Hein for excellent research assistance.

1 Introduction

Panel data are characterised by high dimensionality due to having both cross-sectional (N units) and longitudinal (T time periods) dimensions. Panel analysis is appealing, because it gives consistent estimates under weaker restrictions on the correlation between observables and unobservables than cross-sectional analysis. The leading example is the so-called linear fixed effects (FE) model, where observables can be arbitrarily correlated with time-constant unobservables. A disadvantage of the FE model is that it is overparametrised in the presence of time-constant covariates. While parameters on time-varying covariates are identifiable and are consistently estimated by the classical FE estimator, or by Mundlak (1978) type estimators (compare Wooldridge, 2019), the parameters on the time-constant covariates are not identifiable in the FE model. Mundlak (1978) type models substantially restrict the correlation of the time-constant covariates with the FEs. Combining machine learning methods to regularise the space of fixed effects is tricky, as by increasing the number of units, the number of parameters in the model increases and therefore it is different from regularisation in a given parameter space. The overparametrisation also leads to multicollinearity, which causes problems for regularisation methods. Existing approaches therefore require additional restrictions on group numbers or the covariates structure.

This paper considers the conventional linear FE model with discrete time-constant covariates and suggests a new estimator that clusters the FEs. By adopting nonparametric density based clustering, the cluster structure, such as cluster locations and their number, are determined from the data without restrictions. The clustering is so adaptive that it does not force units into clusters if they are not similar enough. Therefore only similar units are clustered, while others remain atoms. Our approach is attractive to practitioners for the following reasons: It works with a large number of units and groups and gives estimates for parameters on discrete time-constant covariates that can only take on a small number of values such as dummy variables. We show that our estimator is consistent and converges at rate $O_p(1/\sqrt{NT})$.

Most closely related to our approach is the pioneering work by Bonhomme and Manresa (2015), which allows for time-constant covariates that are correlated with covariates. While in their model, these covariates must take on sufficiently many values to identify their parameters, our model contains any discrete time-constant covariates, such as dummy variables. While Bonhomme and Manresa (2015) use k-Means for the clustering, which requires a known number of clusters

and normally forces all units into clusters, our approach is distribution free by using the HDBSCAN algorithm. Other existing approaches that combine panel models with regularisation techniques either do not allow for time constant covariates, require them to be uncorrelated with fixed effects (Berger and Tutz, 2018; Bondell et al., 2010; Bonhomme et al., 2022; Fan and Li, 2012; Heinzl and Tutz, 2014; Li et al., 2018; Schelldorfer et al., 2011; Rohart et al., 2014; Su et al., 2016) or cannot be computed with large sample sizes (Tutz and Oelker, 2017; Tutz and Schauberger, 2015). We conduct a series of Monte Carlo simulations to provide evidence of our approach producing reliable estimates in a range of scenarios. Because it has a modular structure, the clustering algorithm can be easily changed by the researcher and we provide comparative results as a robustness check. We illustrate with the help of a wage equation from labour economics that our approach is practicable with a large number of units (77,500) and gives more insightful results than the classical FE and the Mundlak model.

The paper is structured as follows. Section 2 presents the model and the statistical approach. Section 3 presents simulation results to investigate finite sample performance, while Section 4 presents the results from an application to labour market data. The last section summarises the main findings and derives some recommendations.

2 The Model

We consider the linear FE panel model

$$\begin{aligned} y_{it} &= W_{it}\theta + v_i + u_{it} \\ &= X_{it}\beta + Z_i\gamma + v_i + u_{it}, \end{aligned} \tag{1}$$

where $i = 1, ..., N$ is the unit and $t = 1, ..., T$ is the time period. $W_{it}' = (X_{it}, Z_i)' \in \mathcal{W} \subset \mathbb{R}^K$ are observable covariates, where $X_{it}' \in \mathcal{X} \subset \mathbb{R}^{K_1}$ are time-varying covariates which may be continuous or discrete. $Z_i' \in \mathcal{Z} \subset \mathbb{R}^{K_2}$ are time-constant discrete covariates that can take on finitely many values from the finite set $\mathcal{Z}$. Only y_{it}, W_{it} are observed, while $\theta = (\beta, \gamma) \in \mathbb{R}^K$ is unknown, v_i is an unknown fixed effect and u_{it} is an unknown idiosyncratic error. The objective is to identify and estimate θ. Following general convention, we assume $E(u_{it}|W_i, v_i) = 0$, where $W_i = (W_{i1}', ..., W_{iT}')'$, i.e. W_{it} is strictly exogenous conditional to v_i. The fixed effects $v_i \in \mathbb{R}$ are not mean restricted, because we consider a variant of the model without a common intercept. The fixed effect is generated as mixture of

a continuous and a discrete distribution, where the latter has a finite support $\{1, \ldots, G_1\}$ with fixed deterministic G_1 where each value in the set corresponds to a cluster. Within a group g all observations have the same value $v_i = q_g$. The continuous part leads to isolated values of v_i which we call atoms or atomic groups. The random number of atoms is denoted by G_2. The classical FE panel model assumes $G_2 = N$ atomic groups and therefore cannot identify γ and v_i but only the sum $Z_i\gamma + v_i$. This is because the model is overparametrised, leading to multicollinearity between the time-constant Z_i and v_i. Therefore, though the model permits for general forms of endogeneities, the interpretability of the results is unclear as only $Z_i\gamma + v_i$ can be identified and not the role of the components Z_i. This is a severe limitation in applications, when the focus is on time-constant variables, such as geographic factors or gender. This paper suggests a new approach for identifying γ using a cluster structure of v_i.

2.1 Group FE Estimation

In this subsection we present our estimation approach. A summary of the approach can be found in Table 1. Further details about the procedure are given in Supplement S1.

Table 1: Steps of estimation procedure. See Supplement S1 for more details.

Step	Description
1a)	Estimate β in model (1) by the FE estimator and retrieve estimated fixed effects $\hat{a}_i = \bar{y}_i - \bar{X}_i\hat{\beta}$, where $\bar{y}_i = T^{-1}\sum_{t=1}^{T} y_{it}$, $\bar{X}_i = T^{-1}\sum_{t=1}^{T} X_{it}$. Note that $\hat{a}_i \approx Z_i\gamma + v_i + \bar{u}_i$ with $\bar{u}_i = T^{-1}\sum_{t=1}^{T} u_{it}$.
1b)	For each value $z \in \mathcal{Z}$: Cluster all units with $Z_i = z$ using the values of $\hat{a}_i$ for indices i with $Z_i = z$.
2a)	Define $\hat{G}_1$ as the maximum number of clusters in Step 1b) over $z \in \mathcal{Z}$.
2b)	Establish linkage between the clusters of Step 1b) for different values of z and assign to each atomic cluster a unit specific label between $\hat{G}_1 + 1$ and $\hat{G}_1 + \hat{G}_2$.
3)	Use OLS to estimate β, γ and α in a regression model with response y_{it} and covariates X_{it}, Z_i and $\hat{D}_i$ where the latter is a dummy variable indicating group membership, see equation (2) below.

Step 1: Clustering

a) There is some evidence that FE estimation of model (1) provides well behaved

estimates for $Z_i\gamma + v_i$ as N and T become large and if there is an underlying group structure, see Hahn and Moon (2010). In a classical individual level FE model, this term is the fixed effect. We denote the estimate of this individual level FE as $\hat{a}_i = Z_i\gamma + v_i + e_i$, where e_i is an estimation error in $\hat{a}_i$, see Step 1a) in Table 1. The estimation error diminishes by sample size and simulations show that convergence in T is quick and not many periods are required.

b) Consistent estimation of $\boldsymbol{\gamma}$ in the model for $\hat{a}_i$ by OLS is not possible because the relationship between Z_i and v_i is unrestricted and therefore Z_i is endogenous. We make use of a clustering approach to overcome this problem. We use a version of density-based clustering which is a slight modification of DBSCAN* (Ester et al. (1996)) and HDBSCAN (Campello et al. (2013, 2015)) which belong to the most well-known density-based clustering algorithms, see Supplement S3 for more details. In Assumption (A5) in the next subsection we will state the model for the distribution of v_i. We assume that with a positive probability v_i is generated from a continuous distribution. Further, with positive probability v_i takes a value from the finite set $\{q_g : g \in \{1, \ldots, G_1\}\}$, where q_g are some unknown real numbers. Thus we have a fraction of v_is spread over the real line and fractions of v_is equal to q_g for some $g \in \{1, \ldots, G_1\}$. We call the first v_is "atoms" and we call the index sets $\{i : v_i = q_g\}$ "clusters". In our asymptotic setting we allow that the probabilities of both fractions do not converge to zero. Then we will have that the number of atoms as well as the number of cluster points are of order N.

For the implementation of our density-based clustering algorithm we use the kernel density estimators

$$\hat{f}_b^z(x) = \frac{1}{N_z} \sum_{i=1}^{N} \mathbb{I}_[Z_i = z] \frac{1}{b} K\left(\frac{\hat{a}_i - x}{b}\right),$$

with $N_z = \#\{i : Z_i = z\}$ where K is a probability density function and b is a bandwidth with $b = c_1^b/\sqrt{T}$ for some $c_1^b > 0$. We consider high level sets of $\hat{f}_b^z$ and correct their boundaries

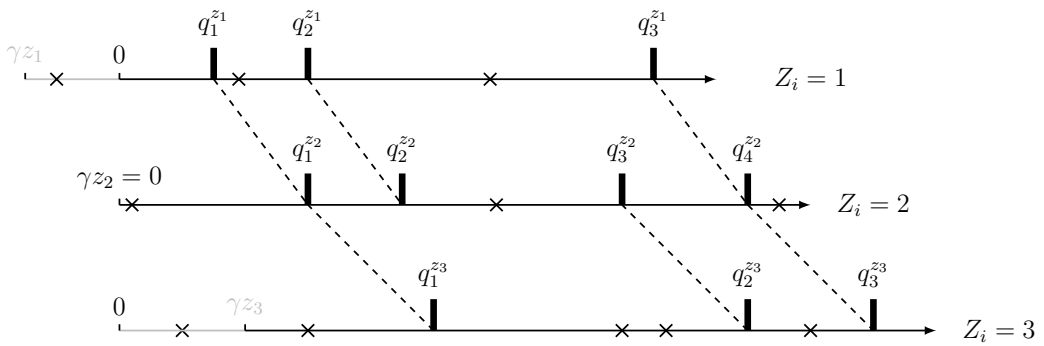
$$\begin{aligned} I_*^z &= \{x : \hat{f}_b^z(x) \geq c_2^b \frac{1}{b}\}, \\ I^z &= \{x : |x - w| \leq c_3^b b \quad \exists w \in I_*^z\} \text{ for some constants } c_2^b, c_3^b > 0. \end{aligned}$$

Here $c_1^b, ..., c_3^b$ are tuning parameters for the method.

We will show that I^z is a union of disjoint closed intervals

$$I^z = I_1^z \cup \ldots \cup I_{l(z)}^z \text{ with } l(z) \geq 1.$$

Figure 1: Population choice of the mappings $h_z(l)$



atom: x, cluster: ▌

$h_{z1}(1) = 1$, $h_{z1}(2) = 2$, $h_{z1}(3) = 4$
$h_{z2}(1) = 1$, $h_{z2}(2) = 2$, $h_{z2}(3) = 3$, $h_{z2}(4) = 4$
$h_{z3}(1) = 1$, $h_{z3}(2) = 3$, $h_{z3}(3) = 4$

We denote the midpoint of the interval I_l^z by $\hat{q}_l^z$. We assume that they are ordered in increasing values. These are the estimates of the cluster centers for $1 \leq l \leq l(z)$ of the subsample $\{i : Z_i = z\}$. In our simulations and data example we also applied the popular k-Means clustering algorithm for comparison with our approach. The k-Means algorithm tends to clusters all units with $\hat{G}_2 = 0$.

Step 2: Mapping of Cluster Membership Variables

a) The number G_1 of clusters is estimated by $\hat{G}_1 = \max_{z \in \mathcal{Z}} l(z)$.

b) To link the clusters for all values z of Z_i we define strictly monotone functions $h_z : \{1, ..., l(z)\} \to \{1, ..., \hat{G}_1\}$. In the fitted model all individuals i with the same value of $h_{Z_i}(l)$ with l chosen such that $\hat{a}_i \in I_l^{Z_i}$ belong to the same cluster. The number of this cluster is $\hat{g}(i) = h_{Z_i}(l)$. The linkage problem is illustrated in Figure 1 for three different values of Z_i and $G_1 = 4$. We now specify our algorithm for the choice of the functions h_z. We choose these functions by comparing the distances between neighbouring clusters for $z \in \mathcal{Z}$. We choose h_z for $z \in \mathcal{Z}$ by minimising

$$\sum \left| (\hat{q}_{l+1}^z - \hat{q}_l^z) - (\hat{q}_{k_2}^{z'} - \hat{q}_{k_1}^{z'}) \right|,$$

where the sum runs over all $z, z' \in \mathcal{Z}$, $1 \leq l \leq l(z) - 1$, $1 \leq k_1 < k_2 \leq l(z')$ with $h_z(l) = h_{z'}(k_1)$ and $h_z(l+1) = h_{z'}(k_2)$. For all atomic groups containing only individual i we choose $\hat{g}(i)$ equal to a unique value in $\{\hat{G}_1 + 1, ..., \hat{G}\}$ with $\hat{G} = \hat{G}_1 + \hat{G}_2$.

Step 3: Dummy Variable Regression for the Regularised Model

After the correspondence between the reference groups and the groups in all reparametrised models is established, a vector of $\hat{G}$ dummy variables $\hat{D}$ indicating estimated group membership is created and the following model is established:

$$y_{it} = X_{it}\beta + Z_i\gamma + \hat{D}_i\alpha + \tilde{u}_{it}, \tag{2}$$

where α is $\hat{G}$ dimensional. $\tilde{u}_{it} = u_{it}$ iff $\hat{D}_i\alpha = v_i$. This regression produces consistent estimates for previously unidentified components γ and α. The estimates for β are more precise than in the conventional FE model due to restricting the fixed effects to take on $\hat{G}$ values.

$\hat{D}_i$ may not be free of error in applications. That is it may not be the same as D_i, the vector of dummies based on the true groups. There are several types of possible errors. An individual that belongs to a cluster could be classified as a member of another cluster. In our asymptotic setting where the location of clusters is fixed and the distance between clusters is positive this happens with very small probability. In particular, we will see in our asymptotic analysis that the bias caused by this type of misclassification is asymptotically negligible. Secondly a cluster point could be classified as an atom. One can show that this does not contribute a bias term. But the individual will not be used in the estimation of γ. This increases the variance of the estimator of γ. The effect will not be negligible in our asymptotic setting because this misclassification may happen for a fraction of the sample that is bounded away from zero. The last type of error is a classification of an atom as cluster point. This misclassification adds a bias term in the estimation of the cluster centers which lead in our asymptotic setting to a bias term of order $T^{-3/2}$ in the estimation of γ. This bias term is only negligible if T converges to infinity fast enough, see the discussion in the next subsection and in Supplement S4.

When the algorithm in Step 1 creates too many groups, an additional supervised regularisation step could be implemented that will fuse groups and therefore remove inefficiencies at the cost of an increasing bias coming from misclassification of atoms as cluster points. This corresponds to finding whether $\hat{D}_i$ is of greater length than G. Given that the position and ordering of each subgroups are known from Step 2, the regularisation corresponds to a fused LASSO. The corresponding optimisation problem is:

$$\min_{\tilde{\lambda}\in\mathbb{R}^{K_1+K_2+\hat{G}}} \frac{1}{2}\|y - \tilde{W}\tilde{\lambda}\|_2^2 + \eta \sum_{g=1}^{\hat{G}_1-1} |\tilde{\lambda}_{g+1} - \tilde{\lambda}_g|, \tag{3}$$

where y is stacked $N * T \times 1$, $\eta \geq 0$ is a tuning parameter and $\tilde{W} = [D_1, W, D_2] \in \mathbb{R}^{(N*T)\times(K_1+K_2+\hat{G})}$. $\tilde{W}$ contains the stacked $\hat{D}$ and W matrices, arranged in a specific column order. The vectors in the $N * T \times \hat{G}_1$ matrix D_1 indicate the membership of individuals in the non-atomic groups and the $N * T \times \hat{G}_2$ matrix D_2 indicates the atomic groups respectively. Further, we order the groups (i.e. columns) in D_1 by the mean estimated fixed effect. To ensure comparability we compute the ordering only with the fixed effects of the units in the reference group with respect to Z, i.e. the reference level of Z which is assumed to contain all non-atomic groups. Using the same order as in $\tilde{W}$, $\tilde{\lambda} \in \mathbb{R}^{K_1+K_2+\hat{G}}$ is the rearranged λ. Problem (3) is a variant of the so-called fused LASSO, as only the coefficients of the non-atomic groups shrink towards each other. The coefficients on both time-variant and time-constant covariates are not regularised. The group coefficients are not regularised towards zero. As shown in Supplement S5 the problem in (3) can be transformed into a regular LASSO as in Tibshirani and Taylor (2011). The regular LASSO has computational advantages and its properties are well developed, see for example Tibshirani (1996) and Hastie et al. (2017). While our approach is a variant of the fused LASSO, an alternative shrinkage approach would be pairwise cross-smoothing as suggested by Heiler and Mareckova (2021).

2.2 Large Sample Properties

In our asymptotic approach we assume that N and T converge to infinity. More formally, we assume that $N \to \infty$ and that $T = T_N$ depends on N and fulfils $\lim_{N\to\infty} T_N = \infty$.

Furthermore, we will suppose that an estimator $\hat{\beta}$ of β is used in Step 1 which fulfils $\|\hat{\beta} - \beta\| = O_p((NT)^{-1/2})$, see Assumption (A1). For the sets $I_1^z, \ldots, I_{l(z)}^z$, introduced in the last subsection we will show that, with probability tending to one, each interval I_j^z contains $q_g + z\gamma$ for exactly one $1 \leq g \leq G_1$. We also write $I^{g,z}$ for this interval. If there exists no j with $q_g + z\gamma \in I^{g,z}$ we define $I^{g,z} = \varnothing$. Under our assumptions the mapping of cluster membership variables of Step 2 identifies the value of g with $I_j^z = I^{g,z}$ with probability tending to one, see Assumption (A7).

We now define the estimator $\hat{\gamma}$ of γ as the minimiser over γ of the least squares criterion

$$\sum_{i=1}^{N} (\hat{a}_i - \check{a}_{\hat{g}(i)} - Z_i\gamma)^2,$$

where

$$\check{a}_g = \sum_{i=1}^{N} \hat{a}_i \mathbb{I}_{\hat{g}(i)=g} / \sum_{i=1}^{N} \mathbb{I}_{\hat{g}(i)=g},$$

and where $\hat{g}(i)$ has been defined in the last subsection. It can be easily checked that

$$\hat{\gamma} = \left(\sum_{i=1}^{N} (Z_i - \check{Z}_{\hat{g}(i)})'(Z_i - \check{Z}_{\hat{g}(i)}) \right)^{-1} \sum_{i=1}^{N} (Z_i - \check{Z}_{\hat{g}(i)})'(\hat{a}_i - \check{a}_{\hat{g}(i)}),$$

where

$$\check{Z}_g = \sum_{i=1}^{N} Z_i \mathbb{I}_{\hat{g}(i)=g} / \sum_{i=1}^{N} \mathbb{I}_{\hat{g}(i)=g}.$$

Note that

$$\hat{\gamma} = \left(\sum_{i:1 \le \hat{g}(i) \le \hat{G}_1} (Z_i - \check{Z}_{\hat{g}(i)})'(Z_i - \check{Z}_{\hat{g}(i)}) \right)^{-1} \sum_{i:1 \le \hat{g}(i) \le \hat{G}_1} (Z_i - \check{Z}_{\hat{g}(i)})'(\hat{a}_i - \check{a}_{\hat{g}(i)}).$$

We now state an asymptotic stochastic expansion for the estimator $\hat{\gamma}$. The expansion implies that $\hat{\gamma}$ achieves a parametric rate of convergence of order $O_p(1/\sqrt{NT})$. We will discuss below how the stochastic expansion can be used to get the asymptotic normal distribution limit for $\hat{\gamma}$.

Theorem 1 *Make assumptions (A1)-(A7) stated below and assume that* $T^{-1} = o(1/\sqrt{N})$. *Then it holds that*

$$\begin{aligned}\hat{\gamma} - \gamma = {} & \Sigma_{Z,N}^{-1} \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{1 \le \hat{g}(i) \le \hat{G}_1} (Z_i - \check{Z}_{\hat{g}(i)})' \bar{u}_i \\ & - \Sigma_{Z,N}^{-1} \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{1 \le \hat{g}(i) \le \hat{G}_1} (Z_i - \check{Z}_{\hat{g}(i)})'(\bar{X}_i - \check{X}_{\hat{g}(i)})(\hat{\beta} - \beta) + o_P(1/\sqrt{NT}),\end{aligned}$$

where

$$\begin{aligned}\Sigma_{Z,N} &= \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{1 \le \hat{g}(i) \le \hat{G}_1} (Z_i - \check{Z}_{\hat{g}(i)})'(Z_i - \check{Z}_{\hat{g}(i)}), \\ \check{X}_g &= \sum_{i=1}^{N} \bar{X}_i \mathbb{I}_{\hat{g}(i)=g} / \sum_{i=1}^{N} \mathbb{I}_{\hat{g}(i)=g}.\end{aligned}$$

In particular we get that

$$\hat{\gamma} - \gamma = O_p(1/\sqrt{NT}).$$

The proof of Theorem 1 is given in the Appendix. For the statement that $\hat{\gamma} - \gamma = O_p(1/\sqrt{NT})$ it would suffice to assume that $T^{-1} = O(1/\sqrt{N})$. As discussed in the last subsection, there are two types of clustering errors that are not negligible in our asymptotic setting. First, atoms with a value of v_i in the neighbourhood of a cluster centre q_g are accidentally assigned to the cluster. Second, cluster elements with large enough $\bar{u}_i$, with $\bar{u}_i = T^{-1} \sum_{t=1}^{T} u_{it}$, are accidentally classified as atoms. The first error leads in particular to bias effects which are shown to be of second order in our proof of Theorem 1. The second type of errors leads to a loss of efficiency. Note that nevertheless our estimator of γ achieves the same rate $O_p(1/\sqrt{NT})$ as if all clusters were known. The assumption $T^{-1} = O(1/\sqrt{N})$ could be weakened if the relative number δ_N of atoms converges to 0 or, more explicitly if we replace α_0^z and α_g^z in Assumption (A5) by $\delta_N \alpha_0^z$ or $\alpha_g^z(1 - \delta_N \alpha_0^z)/(1 - \alpha_0^z)$, respectively. Then the assumption that $\delta_N T^{-1} = O(1/\sqrt{N})$ would suffice.

If T converges slower to $+\infty$ as $N^{-1/2}$ the asymptotic bias of $\hat{\gamma}$ is not negligible because it is of order $T^{-3/2}$, a rate which is not of order $O(1/\sqrt{NT})$. The bias term would vanish if the density of $\bar{u}_i$ is symmetric which in general may not be the case. We assume that the distribution of $\bar{u}_i$ differs from a symmetric distribution by a distance of order $T^{-1/2}$. Theoretically, it is possible to weaken the assumptions further by applying an approach that corrects for bias. In particular, one can construct an estimator of γ that achieves the $O_p(1/\sqrt{NT})$ rate under the assumption $T^{-3/2} = O(1/\sqrt{N})$ or $\delta_N T^{-3/2} = O(1/\sqrt{N})$, respectively. We do not report on this approach because its practical success would heavily depend on the finite sample accuracy of the then used higher order expansions which may be questioned at least for small and moderate sample sizes.

We now discuss how the theorem can be used to get the asymptotic normal limit for $\hat{\gamma}$. We suppose that also a stochastic expansion for $\hat{\beta}$ is given:

$$\hat{\beta} - \beta = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} A_{i,t} u_{i,t}$$

for some vectors $A_{i,t}$ depending on the covariates X_i, t and Z_i. An example is the difference estimator for which such an expansion has been used for an asymptotic analysis. Then we get with the theorem that there exist vectors $B_{i,t}$ depending on the covariates $X_{i,t}$ and Z_i such that

$$\begin{pmatrix} \hat{\gamma} - \gamma \\ \hat{\beta} - \beta \end{pmatrix} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \begin{pmatrix} B_{i,t} \\ A_{i,t} \end{pmatrix} u_{i,t} + o_P(1/\sqrt{NT}).$$

Asymptotic normality of the joint distribution of $\hat{\gamma} - \gamma$ and $\hat{\beta} - \beta$ follows under suitable conditions by application of an appropriate central limit theorem. Our theory does not cover a two-step estimator where in a first step the above procedure is used for the cluster allocations $\hat{g}(i)$ and where in a second step the cluster effects, and the estimates of β and γ are updated by minimizing a least squares criterion. One may expect that one can achieve $1/\sqrt{NT}$ convergence and normal limits if the values of X_{it} lie in a finite set and similar conditions are made for the conditional distribution of v_i, given $X_i t$ and Z_i as above for the conditional distribution given Z_i. But now $(X_i t : 1 \leq t \leq T)$ are vectors of increasing dimension which complicates the situation. The major problem are atoms that erroneously categorized as belonging to a cluster. The number of these atoms is of order $NT^{-1/2}$. The error in the estimation of their individual effect is of order $NT^{-1/2}$. Thus a crude bound for the error in estimation of β is of order $N^{-1}NT^{-1/2}T^{-1/2} = T^{-1}$ which is in general not of order $N^{-1/2}T^{-1/2}$. Thus finer arguments are needed to check under which conditions the two-step estimator achieves $1/\sqrt{NT}$ convergence and normal limits.

Assumptions

Assumption A1 *There exists an estimator $\hat{\beta}$ of β with $\|\hat{\beta} - \beta\| = O_p((NT)^{-1/2})$. The values of X_{it} lie in a bounded set: $\|X_{it}\| \leq C^x$ a.s. for some $C^x > 0$.*

With this assumption there is no need to discuss estimation of β in Step 1. A possible choice for an estimator that fulfils this condition is the fixed effects estimator.

Assumption A2 *The tuples $(Z_i, v_i, \bar{u}_i)$ are i.i.d. It holds that for $1 \leq i \leq N$ that $Eu_{it} = 0$ and that*

$$sup_{u \in \mathbb{R}} |F_N(u) - \Phi(u/\sigma)| \leq C_F T^{-1/2}$$

for some $C_F > 0$ with $\sigma^2 = Eu_{it}^2$. Here, F_N is the distribution function of $\sqrt{T}\bar{u}_i$ and Φ is the distribution function of a standard normal distribution.

For the case that $u_{i1}, ..., u_{iT}$ are i.i.d. with $E|u_{it}|^3 < \infty$ the bound on the distribution function F_N of $\sqrt{T}\bar{u}_i$ follows directly by an application of the Berry-Esseen bound (Feller, 1971). We will exploit below that the density of Φ is symmetric. This will be essential when estimating the location of the centers of the clusters.

We now describe the distribution of $(Z_i, v_i, \bar{u}_i)$.

Assumption A3 *$\bar{u}_i$ is independent of (Z_i, v_i).*

Assumption A4 *The Z_i's have a finite support $\mathcal{Z} \subset \mathbb{R}^{d_Z}$. The linear span of $\mathcal{Z}$ is equal to $\mathbb{R}^{d_Z}$. For $z \in \mathcal{Z}$ we put $p(z) = P(Z_i = z)$. We suppose that p does not depend on N.*

We now describe the conditional distribution of v_i given Z_i.

Assumption A5 *The conditional distribution of v_i given $Z_i = z \in \mathcal{Z}$ is equal to*

$$\alpha_0^z S^z + \sum_{g=1}^{G_1} \alpha_g^z \delta_{q_g},$$

where δ_q denotes a mass point in q, where $q_1 > \ldots > q_{G_1}$ are points in [0,1] and S^z are probability measures on [0,1] with densities s^z that allow for a continuous derivative. Furthermore, α_g^z $(z \in \mathcal{Z}, 0 \leq g \leq G_1)$ are real weights with $\alpha_g^z \geq 0$, $\sum_{g=0}^{G_1} \alpha_g^z = 1$ for all $z \in \mathcal{Z}$, $\sup_{z \in \mathcal{Z}} \alpha_g^z > 0$ $\forall g$.

In (A5) we allow that $\alpha_g^z = 0$ for some $(g, z) \in \{0, \ldots, G_1\} \times \mathcal{Z}$. In (A5) we assume that the order of the number of atom units is the same as that for units in clusters. One could change and model the conditional distribution of v_i given $Z_i = z$ as

$$\delta_N \alpha_0 S^z + (1 - \delta_N) \sum_{g=1}^{G_1} \alpha_g^z \delta_{q_g},$$

where δ_N is a sequence with $\delta_N \to 0$. By doing so we can replace the assumption $T^{-1} = O(1/\sqrt{N})$ by $\delta_N T^{-1} = O(1/\sqrt{N})$, or in case we do assume that the density f_N of $\bar{u}_i$ is symmetric, see comment after the statement of (A2), by $\delta_N T^{-3/2} = O(1/\sqrt{N})$.

For the kernel K we assume:

Assumption A6 *The kernel K is a strongly unimodal symmetric density function and differentiable with derivative absolutely bounded by c_1^K. For the bandwidth b it holds $b = c_1^b/\sqrt{T}$ for some c_1^b. The constant c_2^b in the definition of I_*^z is chosen small enough.*

Note that a density is strongly unimodal if its convolution with a unimodal density is always unimodal. This also implies that the density of the convolution of two strongly unimodal densities is also strongly unimodal. For a density strong unimodality is equivalent to log-concavity. In particular, normal densities are strongly unimodal. Below we will make use of the fact that the convolution of the

kernel K with a normal density is strongly unimodal and thus log-concave. For a discussion of strongly unimodal densities see Ibragimov (1956).

As explained in Subsection 2.1, we make a simplifying assumption to ease identification of clusters and their centres in Step 2.

Assumption A7 *There exists a* $z_* \in \mathcal{Z}$ *with* $\alpha_g^{z_*} > 0$ *for all* $1 \leq g \leq G_1$*. For all* $z \in \mathcal{Z}$ *we assume that there exist* $g_1, g_2 \in \{1, \ldots, G_1\}$*,* $g_1 \neq g_2$*, depending on* z *with* $\alpha_{g_1}^z, \alpha_{g_2}^z > 0$*. Furthermore, we suppose that the values of* $q_{g_1} - q_{g_2}$ *are pairwise different for* $1 \leq g_1 < g_2 \leq G_1$*.*

We conjecture that Assumption (A7) could be weakened but this would require more refined statistical methods and the application of more technical arguments in the mathematical analysis. Note that we identify clusters for each value of Z_i separately without making use of the link $(Z_{i_1} - Z_{i_2})\gamma$. Including this information may motivate more effective approaches if $\mathcal{Z}$ contains more than 2 elements.

2.3 Comparison of Approaches

We compare how classical approaches such as the FE model and Mundlak model compare to our approach in terms of restrictions. We also consider how the choice of the clustering algorithm induces different restrictions.

There are no restrictions on the correlation between v_i and Z_i in the FE model. Moreover, v_i can take on N values. This corresponds to that all units are allowed to be atoms. γ is not identified without restricting G or the relationship between the fixed effect and the observables. The model by Mundlak (1978) assumes that the fixed effect is a function of the time average of time-varying covariates ($\bar{X}_i$) and that its residual variation is not correlated with Z_i. The model produces inconsistent estimates if there is anything in v_i that is related to Z_i conditional on $\bar{X}_i$. The model does not restrict the marginal distribution of v_i , therefore it only restricts the correlation structure of the observables with v_i.

Our suggested approach does not restrict the correlation structure between observables and v_i but for identifiability $G < N - K_2$, the number of atomic units. In the case of k-Means clustering it is assumed that G is a known small number. In the case of density-based clustering, G is unknown and allowed to be large. For identifiability, for each realisation (or more precisely for each distinct realised combination of the components) of Z_i at least two non-atomic clusters must be present in the dataset. However, our theory explicitly allows that groups are not present in some realisations of Z_i (compare Assumption (A5)). Assumption (A7)

ensures that there is one realisation of Z that contains all non-atomic groups. We conjecture that this latter assumption could be relaxed both in the theory as well as in the practical estimation. In regard to theory, this would require both more refined statistical methods and the application of more technical argumentation in the mathematical analysis. The different group numbers across realisations of Z_i are determined from the data by density-based clustering. k-Means in contrast clusters the observations into the same known number of groups for all levels of Z, unless the researcher has additional knowledge about group numbers to differ across levels. It is therefore possible to characterise the various models. When using density-based clustering, no specific assumption is made on the number of groups and groups are allocated by a mechanism that bases on a nonparametric density estimation. In contrast, when using k-Means clustering, the model is more restrictive as the number of groups is known and identification is through their means only. In Section 3 we show with simulations that as long as the restrictions are satisfied, the more restrictive models, including Mundlak are more efficient, while they are biased when these restrictions are violated. This is the usual trade-off one faces in terms of bias and efficiency. On the grounds of these considerations it would be of interest to formulate inference approaches that test for the validity of restrictions. For example, a Hausman type test could be done to test for the validity of the additional restrictions of the k-Means clustering in comparison to density-based clustering.

2.4 Remarks and Extensions

The following two remarks should be of use for practitioners who apply our approach:

- $P(Z_i = z|g_i = g)$ can become low for some values of Z in the case of strong correlation between v_i and Z_i. In this case a large data set may be required for the algorithm to detect a cluster.

- While from a theoretical point of view, Z_i can be high dimensional, there are practical constraints as the clustering step 1 has to be done conditional for all values of Z_i. The applied researcher is advised to include only low dimensional Z_i of key interest. The remaining time-constant variables will be simply absorbed by v_i.

There are several practically relevant extensions to our model that we omitted to focus on the main idea of our approach:

Multi-level models. Linear multi level models are routinely applied in a wide range of applications. Our model can be extended to multi-level fixed effects, e.g. $v_i + f_j$ in the case of two levels. They comprise of, for example, a regional or firm component f_j in addition to v_i. Higher dimensional density clustering methods can be used for regularisation in Step 1.

Continuous Z_i. In the case Z_i contains one or multiple continuous covariates, we suggest a pragmatic approximation by specifying the partial relationship of the continuous time-constant covariates and y_{it} as piecewise constant model (interval dummies).

Continuous v. The fixed effects could be continuously distributed with unknown distribution. This is likely the case in many empirical applications. Forcing them into groups, will lead to an approximation error. The problem is similar to that considered in Bonhomme et al. (2022) who show that incorrect grouping of similar units will not or will only slightly bias estimates. Our simulation results in Section 3 confirm that incorrect grouping of similar units will only lead to small inconsistencies in coefficients of interest.

Further regularisation step to group atomic units. The clustering in Step 2 of our procedure typically produces atoms in applications. These are units that are not clustered with any other unit. In addition to the supervised regularisation to combine groups as outlined in Supplement S5, it is possible to test whether atoms are different from groups or other atoms. In this case they can be combined or merged into existing groups to further reduce the dimensionality of the model. The starting point is that a dummy variable regression model as in equation (2) after the generalised LASSO, will give estimated fixed effects for groups and atoms. On the grounds of these estimates it is possible to determine the nearest neighbours for each atom. Using a Wald test or a t-test based on a reparametrised model it is possible to test the null that the FE of the atom and its nearest neighbour are the same. If the null is not rejected, the two are to be merged into one group.

Inference. For honest bootstrap inference it would be important to take the uncertainty of the regularisation steps into account. Chatterjee and Lahiri (2011) and Chatterjee and Lahiri (2013) suggest residual based bootstrap methods for high dimensional linear regression models that are valid for sparse estimators. Our estimation procedure additionally involves an unsupervised clustering step, but it

would be of interest to develop a residual based bootstrap procedure that produces valid standard errors and p-values.

3 Simulations

We conduct a series of Monte Carlo simulations to investigate the numerical performance of our approach in finite samples and compare it to OLS and the Mundlak approach. Table 2 summarises the 4 simulation designs, that we mainly consider in this section. Designs M1-M4 differ in terms group structures of fixed effects and correlation structures between observables and the FEs. Design M1 is characterised by a high correlation between the fixed effect and the time-varying regressor, design M2 by a high correlation of the fixed effect and the time-constant regressor, design M3 models a large number of atoms, where half of the population is not part of a cluster. Finally design M4 illustrates the effect of different group sizes and different distances between the group intercepts.

We choose the designs such that they possess similarities to the related literature (Tutz and Schauberger, 2015; Berger and Tutz, 2018; Bonhomme et al., 2022). We adapted them to make them more aligned to the theoretical model of Section 2. By doing so the existing approaches become incompatible and can therefore not be included in the comparison.

While the main features of the 4 designs are listed in Table 2, we discuss some of them in more detail in what follows. In terms of Assumption (A5) M1 is characterised by $\alpha_0^z = 0$ for all z and $\alpha_g^z = 1/G = 1/G_1$ for all G clusters and all z levels. In M2 α_g^z varies both across groups and z. M3 includes atoms, i.e. $\alpha_0^z \neq 0$. In design M4 $\alpha_0^z = 0$ and α_g^z varies in g but not in z. The distributions of v_i are as follows: In M1/M2 each v_i is a realisation of a $N(1,2)/N(1,10)$ random variable, that is subsequently discretised into 5 groups. First observations are binned into 5 quantiles and the quantile means are used as final group intercepts. In design M3 the clusters are modelled as in M1/M2, the atomic v_i are independent random draws from a $N(0,1)$ distribution. The latter is aligned to the simulation design in Bonhomme et al. (2022, Supplementary Appendix S3). The mixture of clear groups and atoms creates intervals with different densities, there HDBSCAN is known to have advantages. In design M4 the increased distance between the group intercepts increases the bias induced by choosing an incorrect group number in k-means.

In design M4, for four groups the share of the entire population is drawn from a uniform distribution in the interval [0.1,0.25]. The fifth group is formed by

the residual share. The group intercepts are drawn from five different uniform distributions to ensure more spacing between the groups.

There is one bivariate time-constant covariate and one continuous time-varying covariate in all designs, which are drawn from a binomial or standard normal distribution, respectively. The correlation structures between the covariates and v_i differ across designs (compare row **Correlation Structure**). In M2, the probability of $Z_i = 1$ depends on v_i and differs across groups. The Mundlak estimator is expected to perform worse in this setting compared to the settings M1 and M3.

We use T=20 as in Bonhomme et al. (2022, Supplement S3).

Table 2: Simulation Designs

Design	**M1**	**M2**	**M3**	**M4**
Group Structure adapted from	B&T(2018) & T&O(2017)	B&T(2018) T&O(2017)	Mixed	
G	5	5	$G_1 = 5$ $G_2 = N/2$	5 varying sizes
N	500	500	1000	1000
T	20	20	20	20
Fixed Effect v_i drawn from	$N(1,2)$	$N(1,10)$	$N/2 \sim N(1,2)$ $N/2 \sim N(0,1)$	$U(G_5)$
discretised	5 quantile means	5 quantile means	N/2: 5 q. means $N/2$: none	yes
Time-constant covariate Z_i	$B(0.5)$	$P(Z_i = 1\|v_i) = P_2$	$B(0.5)$	$B(0.5)$
Time-varying x_{it} **covariate**	$0.4v_i + 0.6N(0,1)$	$N(0,1)$	$0.4v_i + 0.6N(0,1)$	$N(0,1)$
β, γ	2,2	2,2	2,2	2,2
Correlation structure	$cor(v_i, x_{it})$ ≈ 0.8	$cor(v_i, Z_i)$ > 0	$cor(v_i, x_{it})$ ≈ 0.8	none
Error term u_{it}	$N(0,3)$	$N(0,3)$	$N(0,3)$	$N(0,3)$

Notes: B&T(2018): Berger and Tutz (2018),T&O(2017): Tutz and Oelker (2017), B,L&M (2022): Bonhomme et al. (2022). Vector $G_5 = [[-15,-14],[-2,-1.5],[1.5,2.5],[6,8.5],[13.5,14.5]]$, Vector $P_2 = (0.35, 0.45, 0.55, 0.55, 0.65)$

We apply different variants of our estimation approach in order to compare their performances. As clustering techniques we use HDBSCAN with and without the optional LASSO step and k-means. In the LASSO step we choose the tuning parameter with BIC, Cross Validation and General Cross Validation. We also compute Post LASSO after Cross Validation but do not report the results here for reasons of brevity, in most settings Cross Validation performs better. HDBSCAN is computed using the R package dbscan described in Hahsler et al. (2019). $MinPts$ (compare Supplement S3) is set to 7 in M1,M3 and 10 in M2,M4. k-Means is

Table 3: Simulation Results

	β			γ		
	Bias	MAD	MSE	Bias	MAD	MSE
M1						
POLS	1.5196	1.5196	2.3117	-0.0010	0.0718	0.0080
Mundlak	0.0014	0.0407	0.0025	-0.0000	0.0505	0.0041
k-Means						
k-Means, 3	0.3427	0.3427	0.1199	-0.0100	0.1538	0.0393
k-Means, 5	0.1206	0.1207	0.0171	-0.0024	0.2426	0.0915
k-Means, 10	0.0400	0.0517	0.0041	-0.0197	0.2549	0.0983
HDBSCAN	0.0360	0.0535	0.0055	-0.0227	0.3484	0.2019
HDBSCAN with LASSO						
Cross Validation	-0.0076	0.0462	0.0041	-0.0168	0.2828	0.1291
Gen Cross Val	0.0339	0.0526	0.0053	-0.0225	0.3434	0.1960
BIC	0.0335	0.0524	0.0053	-0.0224	0.3431	0.1957
M2						
POLS	0.0022	0.0735	0.0086	3.7064	3.7064	14.3185
Mundlak	0.0013	0.0235	0.0009	3.7041	3.7041	14.3027
k-Means						
k-Means, 3	0.0032	0.0339	0.0018	4.6190	4.6235	23.3694
k-Means, 5	0.0013	0.0224	0.0008	-0.0011	0.0474	0.0035
k-Means, 10	0.0013	0.0237	0.0009	0.3208	0.3887	0.8949
HDBSCAN	0.0013	0.0224	0.0008	0.0175	0.0661	0.0903
HDBSCAN with LASSO						
Cross Validation	-0.2037	0.2037	0.0429	0.1251	0.1281	0.1015
Gen Cross Val	-0.0368	0.0399	0.0022	0.0376	0.0686	0.0906
BIC	-0.0368	0.0399	0.0022	0.0376	0.0686	0.0906
M3						
POLS	1.5981	1.5981	2.5552	-0.0001	0.0493	0.0038
Mundlak	-0.0021	0.0282	0.0013	0.0013	0.0347	0.0019
k-Means						
k-Means, 3	0.4994	0.4994	0.2508	0.0106	0.1580	0.0390
k-Means, 5	0.2263	0.2263	0.0525	-0.0230	0.2522	0.0987
k-Means, 10	0.0738	0.0744	0.0068	-0.0010	0.3168	0.1588
HDBSCAN	0.0163	0.0330	0.0017	-0.0014	0.3023	0.1450
HDBSCAN with LASSO						
Cross Validation	0.0012	0.0312	0.0016	-0.0032	0.2618	0.1089
Gen Cross Val	0.0162	0.0330	0.0017	-0.0016	0.2985	0.1412
BIC	0.0161	0.0330	0.0017	-0.0018	0.2982	0.1409
M4						
POLS	0.0070	0.0570	0.0052	0.0012	0.5056	0.3958
Mundlak	0.0007	0.0171	0.0005	0.0017	0.5056	0.3956
k-Means						
k-Means, 3	0.0025	0.0215	0.0007	0.0201	0.2617	0.3107
k-Means, 5	0.0005	0.0168	0.0004	0.0008	0.0333	0.0017
k-Means, 10	0.0006	0.0171	0.0005	0.0513	0.4351	0.6393
HDBSCAN	0.0005	0.0168	0.0005	0.0147	0.1152	0.4071
HDBSCAN with LASSO						
Cross Validation	-0.1770	0.1770	0.0323	0.0131	0.1099	0.3546
Gen Cross Val	-0.0394	0.0402	0.0020	0.0144	0.1139	0.3997
BIC	-0.0395	0.0403	0.0020	0.0145	0.1139	0.3997

Notes: Means of 500 simulations. Simulation designs are defined in Table 2.

computed with different choices of k, including too small, too large and the correct number of groups to investigate how results are affected by misspecification. All k-Means computations apply 100 iterations and 1000 random starting values. As a baseline, we compare our estimators to the Mundlak estimator and a pooled OLS regression.

We simulate the 500 samples for each design and report bias, MAD and MSE for the various approaches in Table 3. The results confirm our suggested approach performs well. Whether HDBSCAN or k-Means clustering give superior results depends on the design and the chosen k. Given that G is normally unknown in applications, there is always the risk of assuming the wrong k. There is no clear pattern for the MSE, whether G is assumed to be too little or too great. The group intercepts have a larger distance in settings M2 and M4. Choosing an incorrect number increases the error by a larger factor than for example in setting M1. Specifying a too small k leads to worse performances for coefficients on both the time-varying and the time-constant covariates. Setting k too large leads to a larger MSE for coefficient on the time-constant covariate. Due to its nonparametric nature the MSE for HDBSCAN tends to be larger than for k-Means if there are any sizeable differences. The LASSO step improves the results with the HDBSCAN clustering, although not always. Cross validation outperforms BIC and general cross validation in most settings.

Further simulations are presented in Supplement S6. They include variants of design M2 with varying combinations of fixed effects and error terms. A larger variance of fixed effects and a smaller error variance both improve estimation results with HDBSCAN and HDBSCAN with LASSO. We explain this by a clearer and more distinct group structure and more precise estimation of fixed effects. We also provide results for $T = 5$. While the errors are larger than for $T = 20$, as expected, our approach is shown to work reasonably well in very short panels. We also provide results for a continuously distributed v_i without mass points. This is in the spirit of Bonhomme et al. (2022), who consider the problem of discretising unobserved heterogeneity. Although, this scenario is not compatible with our modelling, our approach shows a reasonable performance. We also provide a graphical representation of the clustering step in Supplement S7.

4 Application

We apply the proposed methods to labour market data and estimate the gender wage gap. Thereby we demonstrate the applicability to large scale data structures

that are commonly used for empirical economic research. We extract a sample from the Sample of the Integrated Employment Biographies 1975-2014 (SIAB) of the Institute for Employment Research (IAB), Germany. These data contain information from various linked administrative social security registers. SIAB is a 2% random sample of the workforce in Germany that contributed to social insurance in the period 1975–2014. Among other things the SIAB contains daily information about periods of dependent employment and wages with basic information about the individual (such as gender, age and education) and the employing firm (such as business sector). SIAB is available as a scientific use file for independent research. For more information on the data see Ganzer et al. (2017). We extract a yearly panel of wages on the 30th of June for the years 2006-2013. Our sample contains employees aged 16-65, that are subject to social insurance contributions, including those in vocational training. If an employee has a part-time and a full-time job we only consider the full-time job. Further, we only consider the job with the highest salary. In addition to the provided variables, we compute others based on the individual employment history to include tenure (time with the current employer) and additional labour market experience (in addition to current tenure). After some data cleansing, we are left with a balanced panel of 241,076 individuals with 1,928,608 person-year observations. In our model we use one time-constant covariate ($female$), 14 time-varying covariates, 7 year dummies and an intercept, whenever adequate. The analysis of the partial effect of gender and education on wages is popular in empirical economic research and is the reference example in leading econometrics textbooks (e.g. Wooldridge, 2010).

We compare results of the following models:

- Pooled OLS model, where X_{it} and Z_i are contemporaneously exogenous and therefore not allowed to be correlated with v_i.
- Mundlak model, which allows for arbitrary correlation between v_i and X_{is} and some correlation between v_i and Z_i if it is through $\bar{X}_i$, the within time average of the time-varying covariates.
- Our regularisation approach with HDBSCAN as clustering step.
- Our regularisation approach with k-Means as clustering step. We work with 5 clusters and with 55 clusters to illustrate the role of the number of clusters.
- Our regularisation approach with HDBSCAN, followed by a LASSO to regularise group membership further as outlined in Appendix S5.

We use R V4.0.2 for the analysis on Windows Server 2019 with 96GB RAM. Our suggested clustering methods run quickly and give results within several hours, though we encounter memory problems in the clustering steps and when running the grouped fixed effects regression (2) of step 3 when groups are created by HDBSCAN. Because of the large number of atoms (individuals that are not assigned to any group), this regression easily contains 10,000s of dummy variables. Despite the use of big data packages such as biglm (Lumley, 2020) we must restrict the analysis to a randomly chosen 77,500 individuals. The final LASSO step to reduce the groups numbers turned out to require even more memory. For this reason, this last step is only estimated on a smaller sample of 7,500 individuals. This gives some insight as to how much the last supervised regularisation step contributes to dimension reduction. In practice, the final LASSO step is only applicable to large scale data when high performance computing facilities are available. For this step, R requests more than 2800GB of RAM in the case of 77,500 individuals. The running time for the large sample is approximately 3 days to obtain the results in Table 4, where the HDBSCAN based model takes around 2 of these days. The results for the smaller sample are obtained within a couple of hours. We report cluster robust standard errors if not otherwise stated using lm.cluster (Robitzsch et al., 2020), where clustering is done at the individual level. For our suggested approaches we report post-clustering standard errors. For the fused LASSO, we only report point estimates to avoid further computational challenges. It would of course be possible with little difficulty to compute post LASSO standard errors.

The estimation results for the various models are displayed in Table 4 and in Table 5 for the smaller sample. The results in Table 4 show that using Mundlak regression instead of POLS leads to considerable changes in many coefficients, including the work history, part-time, certain business sectors and education. Such an observation is frequent in empirical work as POLS is only consistent if the regressors are not correlated with any component in the error term, while the Mundlak model allows for such correlation via the means of the time-varying covariates. The application of our method with HDBSCAN and k-Means with a larger number of groups gives often similar results as already seen in the simulations. k-Means with a small number of groups is also similar, although there are some economically meaningful differences for several variables such as part-time, several business sectors and higher education. Similar to those findings in the simulations, this can be interpreted as evidence suggesting that an insufficient number of groups has been selected. When comparing the Mundlak results with the results of our methods, we see that the estimated effect of the time-constant variable *female* in

particular is quite different when Mundlak is used. Even though the estimates with our methods are not identical they are much more similar and negative. Our method with HDBSCAN clustering suggests a gender wage gap of 32%, while it is only 18% when the Mundlak model is used. Interestingly, while the Mundlak model suggest that POLS is downward biased for this variable, the results with our methods suggest that the direction of the bias is actually in the opposite direction. This illustrates that the Mundlak model can lead to incorrect conclusions when the correlation of the observables with the fixed effects is not only through the means of the time-varying observables. However, most of the coefficients on the time-varying variables do not differ economically between our methods and Mundlak with the exception of age and part-time. In conjunction with the simulation results, Table 5 confirms that the additional LASSO step leads only to small changes in results. In our application it is because only a small number of group FE are being regularised (6 after HDBSCAN and 1 after k-Means). The main benefit of the LASSO step therefore seems to be that the resulting estimates have statistical optimality properties. Thus, it can be used to check whether the clustering method is working well.

Our example here shows that the application of statistical learning methods in panel analysis is possible for larger data sets. Our results demonstrate that our suggested methods produce sizeably different estimates than the classical panel models under stronger restrictions. This is particularly true in the case of the coefficient on the time-constant covariate that benefits most from the weaker restrictions of our methodology. Our application has also shown that an analysis with 620,000 person year observations is possible on a computer with 96GB RAM, although the last regularising LASSO step requires too much memory. Note that our application cannot definitively answer the question of the size of the gender wage gap. This is because the dependent variable is daily and not hourly wages. The variable *part-time* provides some information about the number of hours worked, but only represents an indicator for reduced working time without precisely controlling for hours worked. Further, the reported variable might be incomplete. To conclude, our estimates point to considerably lower daily wages for females.

Table 4: Estimated coefficients of wage regression model.

	POLS	**Mundlak**	**HDBSCAN**	**k-Means(55)**	**k-Means(5)**
Z_i					
$female$	-0.2115***	-0.1829***	-0.3190***	-0.3647***	-0.3347***
	(0.0034)	(0.0035)	(0.0005)	(0.0007)	(0.0016)
X_{it}					
tenure	0.0216***	-0.0213***	-0.0199***	-0.0197***	-0.0145***
	(0.0003)	(0.0032)	(0.0000)	(0.0000)	(0.0001)
additional experience	0.0187***	-0.0207***	-0.0194***	-0.0193***	-0.0144***
	(0.0003)	(0.0032)	(0.0000)	(0.0000)	(0.0001)
age	-0.0087***	-0.0089***	0.0498***	0.0497***	0.0415***
	(0.0002)	(0.0002)	(0.0000)	(0.0000)	(0.0001)
$part-time$	-0.4289***	-0.1270***	-0.1569***	-0.1748***	-0.2101***
	(0.0042)	(0.0032)	(0.0006)	(0.0010)	(0.00019)
trainee	-1.0791***	-1.0254***	-1.0204***	-1.0163***	-1.0225***
	(0.0095)	(0.0074)	(0.0019)	(0.0057)	(0.0066)
business sector (ref: production)					
agriculture	-0.2787***	-0.1205***	-0.1154***	-0.1070***	-0.1290***
	(0.0117)	(0.0155)	(0.0017)	(0.0022)	(0.0053)
gastronomy	-0.4663***	-0.2264***	-0.2275***	-0.2290***	-0.2681***
	(0.0112)	(0.0194)	(0.0016)	(0.0026)	(0.0053)
construction	-0.2291***	-0.0612***	-0.0540***	-0.0490***	-0.0752***
	(0.0051)	(0.0081)	(0.0009)	(0.0010)	(0.0026)
trade	-0.1221***	-0.0561***	-0.0561***	-0.0563***	-0.0681***
	(0.0042)	(0.0058)	(0.0006)	(0.0008)	(0.0018)
services	-0.0266***	-0.1200***	-0.1125***	-0.1105***	-0.0986***
	(0.0035)	(0.0050)	(0.0005)	(0.0007)	(0.0016)
education/social/health	-0.0220***	-0.0860***	-0.0795***	-0.0748***	-0.0739***
	(0.0043)	(0.0112)	(0.0007)	(0.0001)	(0.0020)
public institutions	0.0302***	-0.0580***	-0.0466***	-0.0407***	-0.0382***
	(0.0045)	(0.0133)	(0.0008)	(0.0010)	(0.0023)
education (ref: none)					
higher education	0.5727***	0.0331***	0.0375***	0.0393***	0.1084***
	(0.0038)	(0.0028)	(0.0007)	(0.0010)	(0.0020)
vocational education	0.1062***	0.0136***	0.0139***	0.0151***	0.0296***
	(0.0027)	(0.0011)	(0.0004)	(0.0006)	(0.0012)
$N = 77,500$, $T = 8$					
Clustering					
individuals with $Z = 0$: 45,974, $Z = 1$: 31,526					
cluster (0/1)			131/134	55/55	5/5
atoms (0/1)			14,172/9,769	0/0	0/0

Notes: *p<0.1; **p<0.05; ***p<0.01. Cluster robust standard errors in parentheses. Non-robust for HDBSCAN. Post-clustering standard errors for HDBSCAN and k-Means. Intercept and year dummies not reported. Averages of $\mathbf{x}_{it}$ not reported (Mundlak).

Table 5: Estimated coefficients of wage regression model (smaller sample).

	Mundlak	**HDBSCAN**	**HDBSCAN +fLASSO**	**k-Means**	**k-Means +fLASSO**
Z_i					
$female$	-0.1664***	-0.3881***	-0.3746	-0.2564***	-0.2529
	(0.0114)	(0.0025)		(0.0030)	
X_{it}					
tenure	-0.0020	-0.0010***	0.0000	-0.0016***	0.0000
	(0.0098)	(0.0002)		(0.0002)	
additional experience	-0.0021	-0.0011***	-0.0001	-0.0016***	-0.0003
	(0.0098)	(0.0001)		(0.0002)	
age	-0.0089***	0.0316***	0.0292	0.0316***	0.0294
	(0.0007)	(0.0002)		(0.0002)	
$part-time$	-0.1349***	-0.1678***	-0.1829	-0.1862***	-0.1977
	(0.0106)	(0.0037)		(0.0040)	
trainee	-1.0206***	-1.0176***	-1.0213	-1.0156***	-1.0215
	(0.0229)	(0.0190)		(0.0182)	
business sector (ref: production)					
agriculture	-0.0931*	-0.0830***	-0.0947	-0.0709***	-0.0770
	(0.0562)	(0.0077)		(0.0082)	
gastronomy	-0.2559***	-0.2664***	-0.2793	-0.2581***	-0.2662
	(0.0733)	(0.0106)		(0.0101)	
construction	-0.0693**	-0.0587***	-0.0731	-0.0576***	-0.0666
	(0.0318)	(0.0042)		(0.0046)	
trade	-0.0467***	-0.0476***	-0.0536	-0.0473***	-0.0503
	(0.0180)	(0.0028)		(0.0031)	
services	-0.0951***	-0.0905***	-0.0856	-0.0850***	-0.0801
	(0.0166)	(0.0026)		(0.0028)	
education/social/health	-0.0630*	-0.0561***	-0.0558	-0.0531***	-0.0520
	(0.0348)	(0.0030)		(0.0035)	
public institutions	-0.0117	-0.0010	-0.0005	0.0060*	0.0065
	(0.0430)	(0.0032)		(0.0035)	
education (ref: none)					
higher education	0.0445***	0.0407***	0.0743	0.0443***	0.0715
	(0.0113)	(0.0039)		(0.0041)	
vocational education	0.0190***	0.0193***	0.0251	0.0209***	0.0236
	(0.0037)	(0.0022)		(0.0024)	
$N = 7,500$, $T = 8$					
Clustering					
number of individuals with $Z = 0 : 4,359$, $Z = 1 : 3,141$					
cluster (0/1)		57/56		55/55	
atoms (0/1)		1,314/884		0/0	
group FE		2,255	2,249	55	54

Notes: *p<0.1; **p<0.05; ***p<0.01. Cluster robust standard errors in parentheses. Post-clustering standard errors for HDBSCAN and k-Means. Intercept and year dummies not reported. Averages of $\mathbf{x}_{it}$ not reported (Mundlak).

5 Summary

We introduce a new approach that incorporates unsupervised learning for the estimation of the linear FE panel model. Our method gives consistent estimates for the parameters on both time-constant and time-varying covariates. It complements existing approaches to estimation of panel models by means of statistical regularisation techniques by using nonparametric clustering that does not restrict the number and location of groups and allows for atoms. Moreover, it gives coefficients on discrete covariates that take on only a small number of values. We provide asymptotic theory for the estimator of the parameters on the time-constant covariates and show that it converges in probability at rate $\sqrt{NT}$. Our simulations confirm that our method works as expected and yields low MSE and bias. Our application to the estimation of the gender wage gap confirms that it works with large samples sizes and group numbers and gives practically relevant different estimates compared to the Mundlak model.

References

Berger, M. and Tutz, G. (2018). Tree-structured clustering in fixed effects models. *Journal of Computational and Graphical Statistics*, 27(2):380–392.

Bondell, H. D., Krishna, A., and Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66(4):1069–1077.

Bonhomme, S., Lamadon, T., and Manresa, E. (2022). Discretizing unobserved heterogeneity. *Econometrica*, 90(2):625–643.

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

Campello, R. J. G. B., Kröger, P., Sander, J., and Zimek, A. (2020). Density-based clustering. *WIREs Data Mining and Knowledge Discovery*, 10(2):e1343.

Campello, R. J. G. B., Moulavi, D., and Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In Pei, J., Tseng, V., Cao, L., Motoda, H., and Xu, G., editors, *Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science*, volume 7819, pages 160–172. Springer.

Campello, R. J. G. B., Moulavi, D., Zimek, A., and Sander, J. (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(1):1–51.

Chatterjee, A. and Lahiri, S. N. (2011). Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106(494):608–625.

Chatterjee, A. and Lahiri, S. N. (2013). Rates of convergence of the adaptive lasso estimators to the oracle distribution and higher order refinements by the bootstrap. *The Annals of Statistics*, 41(3):1232–1259.

Chiquet, J., Rigaill, G., Sundqvist, M., and Dervieux, V. (2020). *aricode: Efficient computations of standard clustering comparison measures*. R package.

Croissant, Y. and Millo, G. (2008). Panel data econometrics in R: The plm package. *Journal of Statistical Software*, 27(2):1–43.

Cuadrado, J. L. (2020). *VeryLargeIntegers: Store and operate with arbitrarily large integers*. R package.

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., and Swinton, J. (2019). *xtable: Export tables to LaTeX or HTML*. R package.

Ester, M. (2014). Density-based clustering. In Aggarwal, C. C. and Reddy, C. K., editors, *Data Clustering. Algorithms and Applications*, pages 111–124. CRC Press.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, page 226–231. AAAI Press.

Fan, Y. and Li, R. (2012). Variable selection in linear mixed effects models. *The Annals of Statistics*, 40(4):2043–2068.

Feller, W. (1971). *An introduction to probability theory and its applications*. Wiley.

Fox, J. and Weisberg, S. (2019). *An R companion to applied regression*. Sage, third edition.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Ganzer, A., Schmucker, A., vom Berge, P., and Wurdack, A. (2017). Sample of integrated labour market biographies - regional file 1975-2014 : (siab-r 7514). *FDZ Datenreport. Documentation on Labour Market Data 201701 en*.

Hahn, J. and Moon, H. R. (2010). Panel data models with finite number of multiple equilibria. *Econometric Theory*, 26(3):863–881.

Hahsler, M., Piekenbrock, M., and Doran, D. (2019). dbscan: Fast density-based clustering with R. *Journal of Statistical Software*, 91(1):1–30.

Hastie, T., Tibshirani, R., and Friedman, J. (2017). *The elements of statistical learning : data mining, inference, and prediction*. Springer, second edition.

Heiler, P. and Mareckova, J. (2021). Shrinkage for categorical regressors. *Journal of Econometrics*, 223(1):161–189.

Heinzl, F. and Tutz, G. (2014). Clustering in linear-mixed models with a group fused lasso penalty. *Biometrical Journal*, 56(1):44–68.

Ibragimov, I. (1956). On the composition of unimodal distributions. *Theory of Probability and its Applications*, 1(2):255–260.

Lai, R. (2020). *arrangements: Fast generators and iterators for permutations, combinations, integer partitions and compositions*. R package.

Li, Y., Wang, S., Song, P. X.-K., Wang, N., Zhou, L., and Zhu, J. (2018). Doubly regularized estimation and selection in linear mixed-effects models for high-dimensional longitudinal data. *Statistics and Its Interface*, 11(4):721–737.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

Lumley, T. (2020). *biglm: Bounded memory linear and generalized linear models*. R package.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Le Cam, L. M. and Neyman, J., editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 5, pages 281–297. University of California Press, Berkeley.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2021). *cluster: Cluster analysis basics and extensions*. R package.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, 46(1):69–85.

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robitzsch, A., Grund, S., and Henke, T. (2020). *miceadds: Some additional multiple imputation functions, especially for 'mice'*. R package.

Rohart, F., San Cristobal, M., and Laurent, B. (2014). Selection of fixed effects in high dimensional linear mixed models using a multicycle ECM algorithm. *Computational Statistics and Data Analysis*, 80(C):209–222.

Schelldorfer, J., Bühlmann, P., and De Geer, S. V. (2011). Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. *Scandinavian Journal of Statistics*, 38(2):197–214.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. *Annals of Statistics*, 39(3):1335–1371.

Tutz, G. and Oelker, M. (2017). Modelling clustered heterogeneity: Fixed effects, random effects and mixtures. *International Statistical Review*, 85(2):204–227.

Tutz, G. and Schauberger, G. (2015). Extended ordered paired comparison models with application to football data from German Bundesliga. *AStA Advances in Statistical Analysis*, 99(2):209–227.

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.

Wickham, H., François, R., Henry, L., and Müller, K. (2021). *dplyr: A grammar of data manipulation*. R package.

Wickham, H. and Miller, E. (2021). *haven: Import and export 'SPSS', 'Stata' and 'SAS' files*. R package.

Wilke, C. O. (2020). *cowplot: Streamlined plot theme and plot annotations for 'ggplot2'*. R package.

Wooldridge, J. M., editor (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, second edition.

Wooldridge, J. M. (2019). Correlated random effects models with unbalanced panels. *Journal of Econometrics*, 211(1):137–150.

Appendix

A.I Proof of Theorem 1

We start by showing that, with probability tending to one,

(a) I^z is a union of disjoint closed intervals $I^z = I_1^z \cup \cdots \cup I_{l(z)}^z$ with $l(z) \leq \hat{G}_1$.

Furthermore we will show that, with probability tending to one,

(b) each interval I_j^z contains $q_g + z\gamma$ for exactly one $1 \leq g \leq \hat{G}_1$. As said, we also write $I^{g,z}$ for this interval.

For a proof of these claims define:

$$\tilde{f}_b^z(x) = \frac{1}{N_z} \sum_{i=1}^{N} \mathbb{1}(Z_i = z) \frac{1}{b} K\left(\frac{Z_i\gamma + v_i + \bar{u}_i - x}{b}\right).$$

It can be easily checked that

$$(c) \qquad \sup_{z \in \mathcal{Z}, x} |\tilde{f}_b^z(x) - \hat{f}_b^z(x)| = O_p\left(\frac{\sqrt{T}}{\sqrt{N}}\right).$$

For a proof of this statement one makes use of

$$\sup_{z \in \mathcal{Z}, x} \frac{1}{N_z} \sum_{i=1}^{N} \mathbb{1}(Z_i = z) \mathbb{1}(|v_i + \bar{u}_i - x|) \leq Cb) = O_p(b\sqrt{T}) = O_p(1)$$

for $C > 0$,

$$\sup_{1 \leq i \leq N} \left| K\left(\frac{Z_i\gamma + v_i + \bar{u}_i - x}{b}\right) - K\left(\frac{a_i - x}{b}\right) \right| = O_p\left(\frac{1}{\sqrt{N}}\right).$$

For a proof of (a) choose $\delta > 0$ with $q_g' - q_g > 2\delta$ for all $g' \neq g$. Because of (A2), (A5) and (c), with probability tending to one it holds that $\hat{f}_b^z(x) = O_p(1)$, uniformly for x not in an interval

$$I_{g,\delta}^z = [q_g + z\gamma - \delta, q_g + z\gamma + \delta], (1 \leq g \leq \hat{G}_1, z \in \mathcal{Z}).$$

Choose $g_0 \in \{1, \ldots, \hat{G}_1\}, z_0 \in \mathcal{Z}$ with $\alpha_{g_0}^{z_0} > 0$. We now show that:

$$(d) \qquad I^{z_0} \cap I_{g_0,\delta}^{z_0} \text{ is a closed interval.}$$

Note that (d) implies (a) and (b). To simplify notation we assume that $q_{g_0} + z_0\gamma = 0$ and that $c_1^b = 1$. Then we have that $b = 1/\sqrt{T}$. For the proof of (d) we define independent random variables

$$V^z(i)\ (z \in \mathcal{Z}, 1 \le i \le N)$$

with

$$P(V^z(i) = g) = \alpha_g^z, (g = 0, \ldots, \hat{G}_1).$$

Given $Z_i = z, V^z(i) = g$, put $v_i^\# = q_g$ if $1 \le g \le \hat{G}_1$ and $v_i^\#$ conditionally distributed according to S^z if $V^z(i) = 0$.

Note that, given $Z_i, \bar{u}_i$, the variable $v_i^\#$ has the same conditional distribution as v_i. W.l.o.g. we assume $v_i^\# = v_i$. For $x \in I_{g,\delta}^z$ we have with probability tending to one,

$$\hat{f}_b^z(x) = \hat{f}_{b,0}^z(x) + \hat{f}_{b,g}^z(x)$$

with

$$\hat{f}_{b,v}^z(x) = \frac{1}{N_z}\sum_{i=1}^N \mathbb{1}(Z_i = z, V^z(i) = v)\frac{1}{b}K\left(\frac{\hat{a}_i - x}{b}\right)$$

for $0 \le v \le \hat{G}_1$. Put

$$\tilde{f}_{b,v}^z = \frac{1}{N_z}\sum_{i=1}^N \mathbb{1}(Z_i = z, V^z(i) = v)\frac{1}{b}K\left(\frac{z\gamma + v_i + \bar{u}_i - x}{b}\right).$$

Uniformly for $x \in I_{g,\delta}^z$ it holds that

$$\begin{aligned} &\text{(e)} \qquad \hat{f}_{b,0}^z(x) - \tilde{f}_{b,0}^z(x) = O_p(1/\sqrt{N}), \\ &\text{(f)} \qquad \hat{f}_{b,g}^z(x) - \tilde{f}_{b,g}^z(x) = O_p(\sqrt{T}/\sqrt{N}). \end{aligned}$$

Expansions (e), (f) follow similarly as (c). With $x_* = x\sqrt{T}$, $\bar{u}_{*,i} = \bar{u}_i\sqrt{T}$ we get for all constants $C > 0$ uniformly for $|x_*| \le C$ that

$$\begin{aligned} \tilde{f}_{b,g_0}^{z_0}(x) &= \tilde{f}_{b,g}^{z_0}(x_*/\sqrt{T}) \\ &= \frac{1}{Nb}\sum_{i=1}^N \mathbb{1}(Z_i = z_0, V^{z_0}(i) = g_0)K\left(\frac{\bar{u}_i - x}{b}\right) \\ &= \frac{\sqrt{T}}{N}\sum_{i=1}^N \mathbb{1}(Z_i = z_0, V^{z_0}(i) = g_0)K\left(\bar{u}_{*,i} - x_*\right) \\ &= \sqrt{T}\left(\Delta_N^1(x_*, z_0) + \Delta_{N,T}^2(x_*, z_0) + O_p(1/\sqrt{N})\right), \end{aligned}$$

where with the standard normal density φ

$$\Delta_N^1(x_*, z_0) = p(z_0)\alpha_{g_0}^{z_0} \int K(u - x_*)\frac{1}{\sigma}\varphi(u/\sigma)\,\mathrm{d}u = O(1),$$
$$\Delta_{N,T}^2(x_*, z_0) = p(z_0)\alpha_{g_0}^{z_0} \int K(u - x_*)\left(\frac{1}{\sigma}\varphi(u/\sigma)\,\mathrm{d}u - F_N(\,\mathrm{d}u)\right) = O(T^{-1/2}),$$

where in the second statement we used (A2), (A6) and

$$\int K(u - x_*)\left(\frac{1}{\sigma}\varphi(u/\sigma)\,du - F_N(\,\mathrm{d}u)\right) = -\int K'(u - x_*)\left(\Phi(u/\sigma) - F_N(u)\right)\,\mathrm{d}u.$$

Furthermore, we get for all constants $C > 0$ uniformly for $|x_*| \le C$ that

$$\begin{aligned}
&\tilde{f}_{b,0}^{z_0}(x) + \tilde{f}_{b,0}^{z_0}(-x) \\
&\quad = \frac{1}{Nb}\sum_{i=1}^{N} \mathbb{1}(Z_i = z_0, V^{z_0}(i) = 0) \\
&\qquad \times \left\{K\left(\frac{v_i - q_{g_0} + \bar{u}_{*,i}/\sqrt{T} - x}{b}\right) + K\left(\frac{v_i - q_{g_0} + \bar{u}_{*,i}/\sqrt{T} + x}{b}\right)\right\} \\
&\quad = p(z_0)\alpha_0^{z_0} \int \left\{K\left(\frac{v - q_{g_0} + v_i/\sqrt{T} - x}{b}\right) + K\left(\frac{v - q_{g_0} + v_i/\sqrt{T} + x}{b}\right)\right\} \\
&\qquad \times \frac{1}{b}F_N(\mathrm{d}u)s^{z_0}(v)\mathrm{d}v + O_p(1/\sqrt{Nb}) \\
= \;& \Delta_{N,T}^3(x_*, z_0) + O_p(N^{-1/2}T^{1/4}),
\end{aligned}$$

where

$$\begin{aligned}
\Delta_{N,T}^3(x_*, z_0) &= p(z_0)\alpha_0^{z_0} \int K(w)(s^{z_0}(q_{g_0} + bw - u/\sqrt{T} + x_*/\sqrt{T}) \\
&\qquad + s^{z_0}(q_{g_0} + bw - u/\sqrt{T} - x_*/\sqrt{T}))F_N(\,\mathrm{d}u)\,\mathrm{d}w \\
&= O(1).
\end{aligned}$$

Finally, we get for all constants $C > 0$ uniformly for $|x_*| \le C$ that

$$\tilde{f}_{b,0}^{z_0}(x) - \tilde{f}_{b,0}^{z_0}(-x) \;=\; \Delta_{N,T}^4(x_*, z_0) + O_p(N^{-1/2}T^{1/4}),$$

where

$$\begin{aligned}\Delta^4_{N,T}(x_*, z_0) &= p(z_0)\alpha_0^{z_0} \int K(w)(s^{z_0}(q_{g_*} + bw - u/\sqrt{T} + x_*/\sqrt{T}) \\ &\qquad - s^{z_0}(q_{g_*} + bw - u/\sqrt{T} - x_*/\sqrt{T}))\, F_N(\mathrm{d}u)\, \mathrm{d}w \\ &= p(z_0)\alpha_0^{z_0} T^{-1/2} \partial s^{z_0}(q_g) 2x_* + o(T^{-1/2}) \\ &= O(T^{-1/2}).\end{aligned}$$

Here ∂s^{z_0} denotes the derivative of s^{z_0}. We now consider $x_{*,-} < 0 < x_{*,+}$, where these values are solutions of the equations

$$\hat{f}_b(x_{*,-}/\sqrt{T}) = c_2^b \frac{1}{b} = \hat{f}_b(x_{*,+}/\sqrt{T}).$$

Note that $x_{*,-}$ and $x_{*,+}$ may not be uniquely defined by the equations. But one can check that the following considerations apply for all choices of $x_{*,-}$ and $x_{*,+}$. For $x_{*,\pm} \in \{x_{*,-}, x_{*,+}\}$ we get that

$$\begin{aligned} c_2^b = \frac{1}{\sqrt{T}} \hat{f}_b\left(x_{*,\pm}\right) &= H(\sqrt{T} x_{*,\pm}) + O\left(\frac{1}{T}\right) + O_p\left(\frac{1}{\sqrt{N}}\right), \text{ where} \\ H_{N,T,z_0}(x_*) = H(x_*) &= \Delta^1_N(x_*, z_0) + \Delta^2_{N,T}(x_*, z_0) + \frac{1}{2\sqrt{T}} \Delta^3_{N,T}(x_*, z_0).\end{aligned}$$

We compare $x_{*,+}$ and $x_{*,-}$ with $x^j_{*,+} > 0 > x^j_{*,-}$ $(1 \le j \le 3)$, where $x^j_{*,\pm} \in \{x^j_{*,-}, x^j_{*,+}\}$ solves

$$\begin{aligned} &\Delta^1_N(x^1_{*,\pm}, z_0) = c_2^b, \\ &\Delta^1_N(x^2_{*,\pm}, z_0) + \Delta^2_{N,T}(x^2_{*,\pm}, z_0) = c_2^b, \\ &\Delta^1_N(x^3_{*,\pm}, z_0) + \Delta^2_{N,T}(x^3_{*,\pm}, z_0) + \frac{1}{2\sqrt{T}} \Delta^3_{N,T}(x^3_{*,\pm}, z_0) = c_2^b. \end{aligned}$$

For a study of $x^1_{*,\pm}$ note that $x_* \to J(x_*) = \int K(v - x_*) \frac{1}{\sigma} \varphi(v/\sigma)\, dv$ is a log-concave function. At this point we assume that $p(z_0)\alpha_g^{z_0} J(0) > c_2^b$. For this reason we assume in Assumption (A6) that c_2^b is small enough. We now use that $\log J$ is concave. This gives for $\delta > 0$ small enough that for $0 < x_1 < x_2$ with

$$\log c_2^{*,b} + \delta \ge \log J(x_1) > \log J(x_2) \ge \log c_2^{*,b} - \delta$$

for $c_2^{*,b} = c_2^b/(p(z_0)\alpha_g^{Z_*}J(0))$ it holds that

$$x_2 - x_1 \leq \frac{\log J(x_1) - \log J(x_2)}{\log J(0) - \log c_2^{*,b} - \delta}\ x_\delta,$$

where x_δ is the solution of

$$\log J(x_\delta) = \log c_2^{*,b} + \delta.$$

From this inequality we conclude that

$$\begin{aligned} x_{*,\pm}^j - x_{*,\pm}^3 &= O(1/\sqrt{T}) + O_p(1/\sqrt{N}),\ \text{for } j \in \{1,2\}, \\ x_{*,\pm}^3 - x_{*,\pm} &= O(1/T) + O_p(1/\sqrt{N}). \end{aligned}$$

Note also that, because of

$$\Delta_N^1(x_*, z_0) = \Delta_N^1(-x_*, z_0)$$

we have that $x_{*,-}^1 = -x_{*,+}^1$. We conclude that I^{g,z_0} is equal to $\left[\frac{x_{*,-}}{\sqrt{T}} - c_3^b b, \frac{x_{*,+}}{\sqrt{T}} + c_3^b b\right]$. Note that the centre of this interval $\frac{1}{2\sqrt{T}}(x_{*,+}+x_{*,-})$ is of order $O(T^{-1})+O_p(1/\sqrt{NT})$ and that its length is of order $\frac{1}{\sqrt{T}}(x_{*,+} - x_{*,-}) = O(1/\sqrt{T}) + O_p(1/\sqrt{NT})$.

Remind that for simplifying notation we have assumed that $z_0\gamma + q_g = 0$. For general z, g we have that $I^{q,z}$ is an interval with midpoint $z\gamma + q_g + O(T^{-1}) + O_p(1/\sqrt{NT})$ and length $O(1/\sqrt{T}) + O_p(1/\sqrt{NT})$, which shows (d) and thus also (a) and (b).

At this point we would like to mention that the term $O(T^{-1})$ for the rate of the midpoint of the intervals is caused by the term $\Delta_{N,T}^2(x_*, z_0)$. In principal one could apply a bias correction of this term based on an estimate of the skewness of the errors u_{it}. Because of symmetry of the third term $\Delta_{N,T}^3(x_*, z_0) = \Delta_{N,T}^3(-x_*, z_0)$ this would result in an error of order $O(T^{-3/2}) + O_p(1/\sqrt{NT})$ for the midpoint of the intervals. We do not pursue this idea here and we do not construct bias corrected estimates of the midpoints of the intervals because their success heavily depends on the finite sample accuracy of Edgeworth expansions which may be doubted. Furthermore it requires that the error variables have the same skewness which may not be true in many applications. We only mention shortly below the resulting order of convergence for the estimator of γ.

We now make use of our considerations to discuss the rate of convergence of the estimator $\hat{\gamma}$. The essential point here is to show that atoms that are classified as

belonging to a cluster are asymptoically negligible for the distribution of $\hat{\gamma}$. Using the results from above we will now show that

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(Z_i = z, \hat{g}(i) = g, V^z(i) = 0)(v_i + \bar{u}_i - q_g) \\ = o_P\left(T^{-1/2}N^{-1/2}\right).\end{aligned} \tag{4}$$

For getting this bound we note first that for constants $c > 0$ and for $1 \leq g \leq \hat{G}_1$ one gets that the number of atom points (i.e. $V^z(i) = 0$) in the interval $[q_g - c/\sqrt{T}, q_g + c/\sqrt{T}]$ is of order $O_P(N/\sqrt{T})$. Furthermore, conditionally given $Z_i = z, V^z(i) = 0$ and that $v_i + \bar{u}_i \in [q_g - c/\sqrt{T}, q_g + c/\sqrt{T}]$, the random variables $v_i + \bar{u}_i - q_g$ have a conditional expectation of order $O(1/T)$ and a conditional standard deviation of order $O(1/\sqrt{T})$. This gives that

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(Z_i = z, V^z(i) = 0)(v_i + \bar{u}_i - q_g)\mathbb{1}_{v_i+\bar{u}_i \in [q_g - c/\sqrt{T}, q_g + c/\sqrt{T}]} \\ &\quad = O_P\left(N^{-1}(N/\sqrt{T})T^{-1} + N^{-1}\sqrt{N/\sqrt{T}}(1/\sqrt{T}\right) \\ &\quad = O_P\left(T^{-3/2} + N^{-1/2}T^{-3/4}\right) \\ &\quad = o_P\left(T^{-1/2}N^{-1/2}\right),\end{aligned}$$

where the condition $T^{-1} = o(N^{-1/2})$ has been used. Under the above discussed bias correction we expect that at this point as at other points of the proof the much weaker condition $T^{-3/2} = O(N^{-1/2})$ would suffice. Now, $\hat{g}(i) = g$ is equivalent to the condition that $v_i + \bar{u}_i \in [q_g - c/\sqrt{T} + \Delta_1, q_g + c/\sqrt{T} + \Delta_2]$, where c is an appropriately chosen constant and where Δ_1 and Δ_2 are random variables of order $O_P(T^{-1/2}N^{-1/2} + T^{-1})$. Thus we have $O_P(N(T^{-1/2}N^{-1/2} + T^{-1}))$ values of $v_i + \bar{u}_i$ between $q_g - c/\sqrt{T}$ and $q_g - c/\sqrt{T} + \Delta_1$ or between $q_g - c/\sqrt{T}$ and $q_g - c/\sqrt{T} + \Delta_2$. This gives

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(Z_i = z, \hat{g}(i) = g, V^z(i) = 0)(v_i + \bar{u}_i - q_g) \\ &\quad = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(Z_i = z, V^z(i) = 0)(v_i + \bar{u}_i - q_g)\mathbb{1}_i \\ &\quad = O_P\left(N^{-1}T^{-1/2}(N(T^{-1/2}N^{-1/2} + T^{-1}))\right) \\ &\quad = o_P\left(T^{-1/2}N^{-1/2}\right),\end{aligned}$$

where $\mathbb{1}_i$ is the indicator function of the event that $v_i + \bar{u}_i$ lies between $q_g - c/\sqrt{T}$

and $q_g - c/\sqrt{T} + \Delta_1$ or between $q_g - c/\sqrt{T}$ and $q_g - c/\sqrt{T} + \Delta_2$. This shows (4).

We now come to a proof of the stochastic expansion of $\hat{\gamma}$ stated in the theorem. By definition of $\hat{\gamma}$ we have that

$$\begin{aligned}
&\Sigma_{Z,N}(\hat{\gamma} - \gamma) \\
&= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{1\leq \hat{g}(i)\leq \hat{G}_1}(Z_i - \check{Z}_{\hat{g}(i)})'(\hat{a}_i - \check{a}_{\hat{g}(i)} - (Z_i - \check{Z}_{\hat{g}(i)})\gamma).
\end{aligned}$$

This implies that

$$\begin{aligned}
W_N &= \Sigma_{Z,N}(\hat{\gamma} - \gamma) + \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_{1\leq \hat{g}(i)\leq \hat{G}_1}(Z_i - \check{Z}_{\hat{g}(i)})'(\bar{X}_i - \check{X}_{\hat{g}(i)})(\hat{\beta} - \beta) \\
&= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{1\leq \hat{g}(i)\leq \hat{G}_1}(Z_i - \check{Z}_{\hat{g}(i)})'(v_i + \bar{u}_i - \check{v}_{\hat{g}(i)} - \check{u}_{\hat{g}(i)}),
\end{aligned}$$

where

$$\begin{aligned}
\check{u}_g &= \frac{\sum_{i=1}^{N} \bar{u}_i \mathbb{1}(\hat{g}(i) = g)}{\sum_{i=1}^{N} \mathbb{1}(\hat{g}(i) = g)}, \\
\check{v}_g &= \frac{\sum_{i=1}^{N} v_i \mathbb{1}(\hat{g}(i) = g)}{\sum_{i=1}^{N} \mathbb{1}(\hat{g}(i) = g)}.
\end{aligned}$$

We now apply (4) and

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{1\leq \hat{g}(i)\leq \hat{G}_1}(Z_i - \check{Z}_{\hat{g}(i)})'(q_{\hat{g}(i)} - \check{v}_{\hat{g}(i)} - \check{u}_{\hat{g}(i)}) = 0.$$

This implies that

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{1\leq \hat{g}(i)\leq \hat{G}_1, V^{Z_i}(i)=0}(Z_i - \check{Z}_{\hat{g}(i)})'(v_i + \bar{u}_i - \check{v}_{\hat{g}(i)} - \check{u}_{\hat{g}(i)}) \\
&= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{1\leq \hat{g}(i)\leq \hat{G}_1, V^{Z_i}(i)=0}(Z_i - \check{Z}_{\hat{g}(i)})'(q_{\hat{g}(i)} - \check{v}_{\hat{g}(i)} - \check{u}_{\hat{g}(i)}) + o_P(N^{-1/2}T^{-1/2}) \\
&= -\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{1\leq \hat{g}(i)\leq \hat{G}_1, V^{Z_i}(i)\neq 0}(Z_i - \check{Z}_{\hat{g}(i)})'(q_{\hat{g}(i)} - \check{v}_{\hat{g}(i)} - \check{u}_{\hat{g}(i)}) + o_P(N^{-1/2}T^{-1/2}).
\end{aligned}$$

We conclude that

$$W_N = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{1\leq \hat{g}(i)\leq \hat{G}_1, V^{Z_i}(i)\neq 0}(Z_i - \check{Z}_{\hat{g}(i)})'(v_i + \bar{u}_i - \check{v}_{\hat{g}(i)} - \check{u}_{\hat{g}(i)} - (q_{\hat{g}(i)} - \check{v}_{\hat{g}(i)} - \check{u}_{\hat{g}(i)}))$$
$$+ o_P(N^{-1/2}T^{-1/2})$$
$$= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{1\leq \hat{g}(i)\leq \hat{G}_1, V^{Z_i}(i)\neq 0}(Z_i - \check{Z}_{\hat{g}(i)})'\bar{u}_i + o_P(N^{-1/2}T^{-1/2}).$$

One can easily verify that the right hand side of the last equation is equal to

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{1\leq \hat{g}(i)\leq \hat{G}_1}(Z_i - \check{Z}_{\hat{g}(i)})'\bar{u}_i + o_P(N^{-1/2}T^{-1/2}).$$

This shows the stochastic expansion stated in the theorem. It remains to show that $\hat{\gamma} - \gamma = O_P(1/\sqrt{NT})$. For this statement it suffices to show that the smallest eigen value of $\Sigma_{Z,N}$ is bounded away from 0. This can be done by choosing $g_z \in \{1, ..., \hat{G}_1\}$ with $\alpha^z_{g_z} > 0$. One can show that with probability tending to one for $\delta > 0$ small enough

$$\begin{aligned}\Sigma_N &\geq \frac{1}{N}\sum_{i=1}^{N}\sum_{z\in\mathcal{Z}} \mathbb{1}_{Z_i=z, V^z(i)=g_z, \hat{g}(i)=g_z}(Z_i - \check{Z}_{\hat{g}(i)})'(Z_i - \check{Z}_{\hat{g}(i)}) \\ &\geq \delta\mathbb{E}\left[(Z_i - \mathbb{E}\left[Z_i\right])'(Z_i - \mathbb{E}\left[Z_i\right])\right] + o_P(1),\end{aligned}$$

where $A \leq B$ for two quadratic matrices means that $B - A$ is positive semidefinite. One can now use Assumption (A4) to bound the smallest eigenvalue of this matrix from below. This concludes the proof of the theorem. □

Estimation of Group Structures in Panel Models with Individual Fixed Effects SUPPLEMENTARY MATERIAL

Enno Mammen,[1] Ralf A. Wilke,[2] Kristina Zapp[3]

S1 Estimation Approach

This supplement describes step by step the estimation procedure. Notation is either introduced or taken from the main text. The approach is implemented in R. See Supplement S2 for the list of required R packages.

1a) Use the FE estimator to estimate β in the model:

$$y_{it} = X_{it}\beta + Z_i\boldsymbol{\gamma} + v_i + \epsilon_{it} \tag{5}$$

and retrieve the estimated fixed effects $\hat{a}_i = Z_i\boldsymbol{\gamma} + v_i + e_i$.

1b) For each distinct combination of values of $z \in \mathcal{Z}$: Cluster all units with $Z_i = z$ using their corresponding values for $\hat{a}_i$. See Supplement S3 for further details on the clustering algorithms HDBSCAN and k-means. HDBSCAN is computed using R package dbscan (Hahsler et al., 2019), k-Means with base R (R Core Team, 2021).

2a) Let C_{z_*} be the assigned cluster membership variable for all units with $Z_i = z_*$. We define the reference level z_0 of the variable Z_i as $z_0 \in \mathcal{Z} : max(C_{z0}) > max(C_{\tilde{z}}) \forall \tilde{z} \in \mathcal{Z}, \tilde{z} \neq z_0 \in \mathcal{Z}$. This means the reference level is the level of Z_i for which the maximum number of clusters was estimated. The cluster membership variable contains labels corresponding to all distinct clusters, starting at 1 and counting upwards. Therefore the maximum of the vector

[1]Heidelberg University, Institute for Applied Mathematics, E–mail: mammen@math.uni-heidelberg.de

[2]Copenhagen Business School, Department of Economics and ZEW Mannheim, E–mail: rw.eco@cbs.dk

[3]ZEW Mannheim, E–mail: kristina.zapp@zew.de

corresponds to the number of distinct non-atomic clusters. Non-clustered "atomic" units are labelled as zero and do not enter the estimated number of clusters. The number of identified clusters in the reference level is denoted $\hat{G}_1$, the estimated value for G_1.

2b) We use the clustering in the subsamples to create a uniform cluster variable in the whole sample. First, the clusters identified in the reference level Z_0 are sorted and relabelled in ascending order of mean($\hat{a}_i$). Let m_0 denote the vector of sorted means, where each element in the vector corresponds to a distinct cluster.

Then, for each $z \in \mathcal{Z}$:

a $\forall c \in \mathcal{C} = \{C_{Zi} | Z_i = z\}$, i.e. $\mathcal{C}$ is the set of all cluster labels corresponding to units with $Z_i = z$, : compute the corresponding mean of the estimated fixed effects $m_c = \text{mean}(\hat{a}_i) : Z_i = Z \,\&\, C_{Zi=c}$. Store the computed means in the first column of a matrix with the corresponding cluster labels in a second column. Order the matrix rows in ascending order of the means, denote the resulting matrix as M_z.

b Compute the set D of all possible draws (combinations) of $|\mathcal{C}|$ elements out of $\hat{G}_1$ elements, where $|.|$ denotes the cardinality of a set. For each $d \in D$: compute steps 1-3.

1. Compute the subvector m_{0d} of m_0 containing all elements indexed with elements contained in d.

2. Compute the vectors $diff_{0d}$ with $diff_{0d(i)} = m_{0d(i)} - m_{0d(i+1)}$, $diff_z$ with $diff_{z(i)} = M_{z(,1)(i)} - M_{z(,1)(i+1)}$ and

3. $f_d = \sum |diff_z - diff_{0d}|$, where $M_{z(,1)(i)}$ denotes the i-th element of vector $M_{z(,1)}$ and $M_{z(,1)}$ the first column of the matrix M_z and $\sum |x|$ the sum of all absolute values of elements in a vector x.

Choose $\hat{d} \in D$ such that $f_{\hat{d}} < f_d \quad \forall d \in D, d \neq \hat{d}$. This combination out of all combinations is chosen as the clusters of the reference level into which the clusters of the level z are grouped into.

Relabel the cluster assignment of z: the cluster label stored in $M_{z(i,2)}$ is relabelled as the i-th element of $\hat{d}$.

The number of combinations can take on very large values, when the number of estimated groups (substantially) differ between the realisations of Z_i. Therefore, we make use of an iterative strategy: only one combination is computed at a time (using R package arrangements Lai

(2020)) and a difference f_d is saved only if it is smaller than all previous differences.

2b) (continued) Assign to each atomic cluster a unit specific label. Starting at $\hat{G}_1 + 1$ up to $\hat{G}_1 + \hat{G}_2$.

3) Regularise the model with a generalised LASSO:

a Set up matrix $\tilde{Q} = [Q', A']'$ and matrix $\tilde{W} = [D_1, W, D_2]$, where the matrices Q and A are defined in Appendix A.IV.

b Compute $\tilde{W}\tilde{Q}^{-1}$, set up $\tilde{W}_1 = \tilde{W}[1 : N * T, 1 : \hat{G}_1]$, $\tilde{W}_2 = \tilde{W}[1 : N * T, 1 + \hat{G}_1 : \hat{G}_1 + \hat{G}_2 + K_1 + K_2]$. This means that $\tilde{W}_1$ contains the first $\hat{G}_1$ columns of $\tilde{W}$ and $\tilde{W}_2$ the remainder of the columns of $\tilde{W}$.

c compute $P = \tilde{W}_2(\tilde{W}_2'\tilde{W}_2)^{-1}\tilde{W}_2'$ and $y_p = (I - P)y$, $\tilde{W}_{1p} = (I - P)\tilde{W}_1$, where I denotes the identity matrix.

d Compute the LASSO path with y_p as response vector and $\tilde{W}_{1p}$ as input matrix.

e Choose the optimal tuning parameter for the LASSO estimator:
We apply three different criteria: 10-fold cross validation (CV), generalised cross validation (GCV) and Bayesian Information Criterion (BIC). In the cross validation the 10 random subsets of the data are created using the dimension N of individuals only such that all T observations of one specific individual are in the same subset. As optimal coefficient vector the most regularised model is chosen such that the CV error conditional on the coefficient vector is within one standard error of the minimum. Regarding BIC and GCV we implement the expressions defined in Hastie et al. (2017), see p. 244, formula 7.52 for GCV and p.233, formula 7.36 for BIC. This leads to an optimal parameter vector $\boldsymbol{\varphi_1}$.

f Transform the parameter vector back to match the response y and input matrix $\tilde{W}$.
Compute $\hat{\varphi}_2 = (\tilde{W}_2'\tilde{W}_2)^{-1}\tilde{W}_2'(y - \tilde{W}_1)\hat{\varphi}_1$ and $\tilde{\lambda} = \tilde{D} * (\hat{\varphi}_1, \hat{\varphi}_2)$

3: Option 1) Different Option: Without the LASSO step directly after step **2b)** estimate the linear model:

$$y_{it} = X_{it}\beta + Z_i\gamma + v_{\hat{g}(i)} + u_{it}, \tag{6}$$

where $\hat{g}(i)$ denotes the estimated cluster for unit i.

3: Option 2) We compute step 3 using cross validation. This leads to a shrunken vector $v_{\hat{g}(i)cv}$. Then we estimate by OLS:

$$y_{it} = X_{it}\beta + Z_i\gamma + v_{\hat{g}(i)cv} + u_{it}. \tag{7}$$

S2 Computation

We use the following R packages in the computation: dbscan (Hahsler et al., 2019), glmnet (Friedman et al., 2010), biglm (Lumley, 2020), plyr (Wickham, 2011), dplyr (Wickham et al., 2021), arrangements (Lai, 2020) , plm (Croissant and Millo, 2008), aricode (Chiquet et al., 2020), miceadds (Robitzsch et al., 2020), haven (Wickham and Miller, 2021), car (Fox and Weisberg, 2019), cluster (Maechler et al., 2021), VeryLargeIntegers (Cuadrado, 2020). For plots and tables we further use ggplot2 (Wickham, 2016), cowplot (Wilke, 2020), xtable (Dahl et al., 2019).

S3 Clustering

After estimated fixed effects have been retrieved as described in Section 2, the aim is to detect latent patterns of heterogeneity by means of a clustering algorithm. These are generally applicable in our context as fixed effects are real valued, can be ordered and are unlabelled, i.e. the group membership is unknown. The clustering algorithm assigns units into groups ("clusters") such that clustered units are more similar than those across clusters. Clustering algorithms require a notion of similarity and dissimilarity, i.e. specifying a distance measure, in our case the Euclidean distance is the natural choice. Units that are not similar enough to other units are not clustered and are called atoms. In the context of our model it is important to allow for a data driven approach, where the number of clusters is not exogenously set but determined on the grounds of a distribution free nonparametric density estimate. For this reason, we use density based clustering methods, where clusters are defined as high-density regions (Campello et al., 2020) without restrictions on the shape of cluster patterns (Ester, 2014, p.111). Compare also the supplementary material S.S7 for a numerical illustration of density-based clustering. Campello et al. (2013, 2015) introduce the HDBSCAN algorithm. It bases on DBSCAN*, which is a small refinement of DBSCAN by Ester et al. (1996), one of the most well-known density-based clustering algorithms. Whether a region

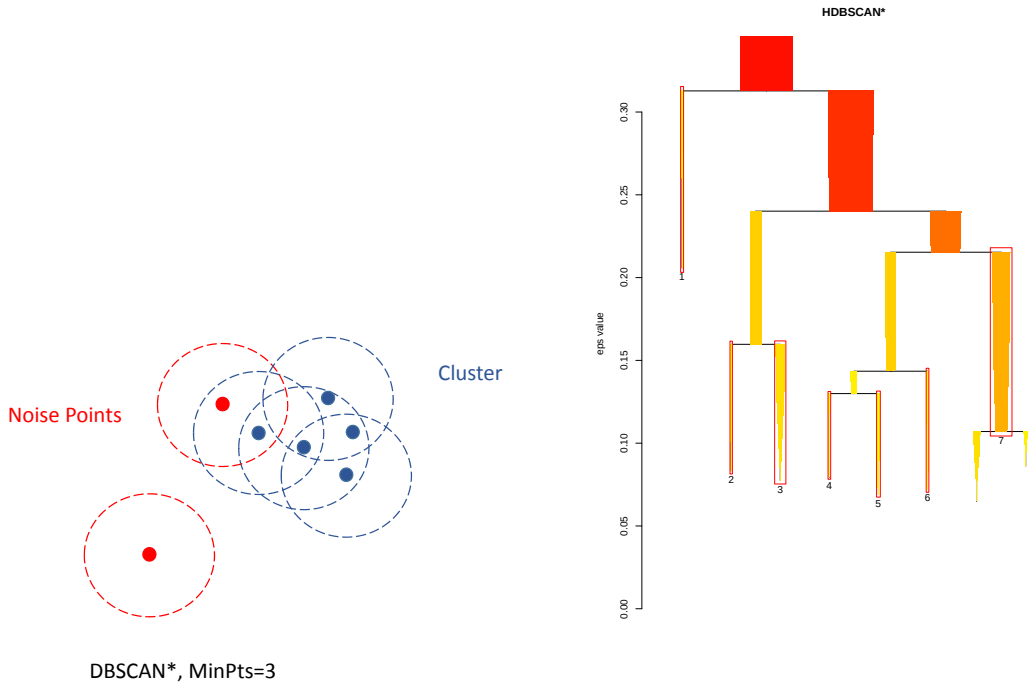
in the data is high-density according to DBSCAN*, and intuitively speaking defines a cluster, is defined by a minimum distance parameter ϵ and a minimum number of points parameter $MinPts$. Points in high-density regions, so-called core points, are surrounded by at least $MinPts$ points within a distance of ϵ. Noise points, all points that are not core points, are not part of a cluster and considered as "atomic" points. Core points lie in the same cluster if they are connected by a chain of core points with each distance being smaller than ϵ. Let $N_\epsilon(x) = \{y \in X | d(x,y) < \epsilon\}$ and $|.|$ denotes the cardinality of a set on which the clustering is computed. The formal definition is given by Campello et al. (2013, p. 162):

x in a dataset X is a **core point** w.r.t ϵ and $MinPts \Leftrightarrow |N_\epsilon(x)| \geq MinPts$.
y in a dataset X is a noise point $\Leftrightarrow |N_\epsilon(y)| < MinPts$.
Two core objects x and $y \in X$ are **ϵ-reachable** if $x \in N_\epsilon(y)$ and $y \in N_\epsilon(x)$.
A **cluster** C w.r.t. ϵ and $MinPts$ is a non-empty maximal subset of X
such that every pair of objects in C is density-connected.

Campello et al. (2013) develop the algorithm going back to Ester et al. (1996) further to HDBSCAN by embedding it in a hierarchical clustering structure. Thereby, they also allow for different density thresholds, i.e. ϵ can vary across clusters within the dataset.This also implies that no ϵ parameter must be predefined by the researcher: The algorithm computes the different clustering outcomes for all possible ϵ values. For $\epsilon \to 0$ all data points will be atoms. For $\epsilon \to \infty$ all data points will be put into one large cluster. Between those extremes lies a nested clustering hierarchy, a tree structure. HDBSCAN identifies all ϵ values where changes in the clustering occur and spans the whole hierarchical clustering tree. Then a simplified tree is built by identifying the ϵ thresholds where "significant clustering changes" occur. These are defined as a split of one cluster into two non-atomic clusters or the disappearance of a non-atomic cluster. Finally out of this simplified clustering hierarchy a final clustering outcome is chosen. This is the result of an optimization that finds the most stable clusters with respect to changes in ϵ , i.e. clusters that are present in the hierarchy over the longest interval of ϵ, with the additional condition that each data point is in exactly one cluster or a noise point.

Because of its popularity and use in related literature (Bonhomme and Manresa, 2015; Bonhomme et al., 2022) we also apply the k-Means algorithm for a comparison in our numerical analysis. The k-Means algorithm, dating back to MacQueen (1967) and Lloyd (1982), assumes that the data can be partitioned into a number

Figure S2: DBSCAN* and HDBSCAN



Notes: Left Picture: own illustration based on illustrations in Ester et al. (1996). Stylised Illustration of DBSCAN* in $\mathbb{R}^2$ with MinPts=3. Right Picture: own illustration created with R package dbscan. Simplified Tree of HDBSCAN in Simulation Setting M2, MinPts=10. The vertical axis plots different ϵ values.

k of convex clusters, where k is exogenously set by the researcher. Put into a statistical perspective, k-Means can be interpreted as the estimation of the means of k underlying Gaussian distributions (Campello et al., 2020). Under these restrictions, the algorithm enjoys greater computational efficiency than the density based algorithms with measurable shorter computation times. The disadvantages are that it does not base on a nonparametric density and that cluster numbers have to be known. In practice, it tends to cluster all units and does not give atoms.

S4 Consequences of Incorrect Subgrouping

We provide large sample results in Section 2.2 and show consistency of our estimator for γ. Using density based clustering, the estimator reaches the same convergence rate as if group membership was known ex ante. Given that any data set is finite, it is of importance to study possible errors that occur in the clustering step. Given a finite dataset a_i will contain an estimation error. In this Supplement we provide a non-technical discussion to compare different estimators. For a more technical discussion with respect to density based clustering see Section 2.2. The clustering algorithm makes two types of errors: atoms with values of v_i in close neighbourhood of a non-atomic cluster may be grouped into this cluster. Cluster points with corresponding large average error terms $\bar{u}_i = T^{-1} \sum_{t=1}^{T} u_{it}$ can be considered as atoms. The latter will increase the number of estimated parameters in the final model and decrease efficiency. The former error leads to a bias, because there are units with different v_i that are assigned to the same cluster. In the limit both errors do still exist. Nevertheless the estimator converges with $O_P(1/\sqrt{NT})$ i.e. with the same rate as if the true group structure would be known. The main result of our proof bases on the assumption that T apporaches infinity with rate $N^{-1/2}$. In practice it is important that the estimator works well in short to medium panels. Typically the number of observations in a dataset is much larger than the number of available time periods. In our simulations we show that the estimator produces reliable estimates in finite samples, we also provide simulation results for very short panels ($T = 5$). In general, the requirements for T can be relaxed if we are willing to make stricter assumptions regarding the error term: i.e. a symmetric density of $\bar{u}_i$ (see Section 2.2 for more details and specifically Assumptions (A2) and (A3) for the proposed more general assumptions on $\bar{u}_i$). Importantly there is also a relation between the requirements for T and the existence of atoms. The rate of T approaching infinity can be relaxed if we assume that the relative number of atoms approaches 0 as N approaches infinity. This corresponds to a stricter

or less general Assumption (A5). In the limit we will not group two non-atomic clusters into one cluster. This is however true for density based clustering but not necessarily for other clustering approaches. In the k-Means clustering approach the number of clusters has to be specified ex ante. If it is unknown to the researcher two types of errors are possible: If G is specified too small clusters will be grouped together although the corresponding observations have different values of v_i. This will lead to biased estimation. G can also be specified too large. If clusters are formed by splitting up true clusters this will only affect efficiency. If additional clusters are formed "between" two existing clusters by combining observations both, this will also lead to a bias. The presence of atoms is not incorporated in the k-Means approach: atoms will be grouped into one of the clusters. We test the finite sample performance of both density based clustering and k-Means in our simulations in Section 3. In Appendix S7 we provide graphs that illustrate the cluster assignment in settings for density based clustering an k-Means with and without atoms and for different values of G in k-Means.

S5 Regularisation of Redundant Groups

In this supplement, we first show that the fused LASSO in Problem (3) is a generalised LASSO. Then we show that it can be transformed into a regular LASSO.

We define a matrix $Q \in \mathbb{R}^{(\hat{G}_1-1)*(K_1+K_2+\hat{G})}$ as:

$$Q = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & & \\ 0 & 0 & \dots & 0 & -1 & 1 & 0 & \dots \end{bmatrix} \tag{8}$$

where only the first $\hat{G}_1$ columns of Q contain non-zero elements. By using Q it is possible to see that Equation (3) is equivalent to:

$$\min_{\tilde{\lambda} \in \mathbb{R}^{K_1+K_2+\hat{G}}} \frac{1}{2}\|y - \tilde{W}\tilde{\lambda}\|_2^2 + \eta\|Q\tilde{\lambda}\|_1, \tag{9}$$

which defines a generalised LASSO problem as discussed by Tibshirani and Taylor (2011).

While the LARS algorithm can be used to find the solution for the regular

LASSO, the fused or generalised LASSO is computationally demanding, in particular if there are many groups. It is unfortunately not straightforward to transfer results for a regular LASSO to a generalised LASSO including the computation of degrees of freedom, choice of optimal tuning parameters and p-values. Tibshirani and Taylor (2011) show, however, that generalised LASSO problems can be written as a regular LASSO problem under a mild restriction by applying a known transformation. We adopt their approach such that more efficient software implementations can be used and to simplify the problem.

The condition that the link to a regular LASSO exists is satisfied in our context because the matrix Q in Problem (9) has full row rank. Following Tibshirani and Taylor (2011) we extend the matrix Q to $\tilde{Q} = [Q', A']'$, where A is $K_1 + K_2 + 1 \times K_1 + K_2 + \hat{G}$ and comprises of $\hat{G}_1$ column vectors of zeros and a block diagonal matrix plus the last row of the matrix being a vector of $\hat{G}_1$ 1s and $K_1 + K_2 + \hat{G}_2$ zeros:

$$A = \begin{bmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & \dots & 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \dots & \vdots & \vdots & \dots & \ddots & \dots & \vdots \\ 0 & \dots & 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 & 1 \\ 1 & \dots & 1 & 0 & 0 & \dots & 0 & 0 \end{bmatrix} \tag{10}$$

$\tilde{Q}$ is $K_1 + K_2 + \hat{G} \times K_1 + K_2 + \hat{G}$ invertible and the rows in A are orthogonal to Q. Therefore, the conditions on A defined in Tibshirani and Taylor (2011) are satisfied. By applying a transformation to the Problem in (9), we obtain

$$\min_{\varphi \in \mathbb{R}^{K_1+K_2+\hat{G}}} \frac{1}{2}\|y - \tilde{W}\tilde{Q}^{-1}\varphi\|_2^2 + \eta\|Q\varphi\|_1, \tag{11}$$

with $\varphi = \tilde{Q}\tilde{\lambda} = (\varphi_1', \varphi_2')'$, where φ_1 contains the first $\hat{G}_1 - 1$ elements of φ. Problem (11) is a regular LASSO with the exception that the penalty shrinks differences in a subset of parameters. Using an orthogonalisation Tibshirani and Taylor (2011) show that there is actually equivalence to a regular LASSO. Let $\tilde{W}\tilde{Q}^{-1}\varphi = \tilde{W}_1\varphi_1 + \tilde{W}_2\varphi_2$, where $\tilde{W}_1$ contains the first $\hat{G}_1 - 1$ columns of $\tilde{W}\tilde{Q}^{-1}$. Problem (11) corresponds then to

$$\min_{\boldsymbol{\varphi_1} \in \mathbb{R}^{\hat{G}_1 - 1}} \frac{1}{2}\|(I - P)y - (I - P)\tilde{W}_1\varphi_1\|_2^2 + \eta\|\varphi_1\|_1, \tag{12}$$

with $P = \tilde{W}_2(\tilde{W}_2'\tilde{W}_2)^{-1}\tilde{W}_2'$, the projection onto the column space of $\tilde{W}_2$ and I the identity matrix. The LARS algorithm can be applied to Problem (12) for estimating

λ, which is just a differently ordered $\tilde{\lambda}$. This is achieved by back-transforming the estimated coefficients through pre-multiplication with the matrix $\tilde{Q}^{-1}$, i.e. $\tilde{Q}^{-1}\hat{\varphi}$. $\hat{\varphi}_2$ is obtained by a linear regression of $y - \tilde{W}_1\hat{\varphi}_1$ on $\tilde{W}_2$.

S6 Additional Simulation Results

Additional Designs

The following table S6 shows variants of simulation design M2 from the main text with different distribution of the group intercepts v_i and idiosyncratic errors u_{it}. HDBSCAN with and without LASSO performs best, when the group differences are larger. In these settings it leads to large errors when k-Means is computed with a too small k. Bonhomme et al. (2022) suggest that a too small k is leading to omitted variable bias. Too large k is also leading to errors, but by a much smaller magnitude. Both Mundlak and Pooled OLS lead to biased results, especially in the settings with larger group intercepts and differences. HDBSCAN performs worse in the setting M2A where errors are larger and group intercepts relatively small. This might indicate that it is sensitive to biased estimation of fixed effects rather than to an included correlation structure.

Further, we simulate an additional design M5, where all individuals are modelled as atoms. This is similar to the model in Bonhomme et al. (2022). We note that this is not in line with the assumptions in our model but we consider it as an insightful special case. For this simulation design we use larger k values in k-Means (compare table S9) and set *MinPts* to 5 in the HDBSCAN algorithm. The simulation design is defined in table S8, the results in table S9.

Table S6: Simulation Designs M2

Design	Group Structure adapted from	G	N	T	Fixed Effect v_i drawn from	Fixed Effect v_i discretised	error u_{it}
M2A	B&T(2018) & T&O(2017)	5	500	20	$N(1,2)$	5 quantile means	$N(0,3)$
M2	B&T(2018) & T&O(2017)	5	500	20	$N(1,10)$	5 quantile means	$N(0,3)$
M2B	B&T(2018) & T&O(2017)	5	500	20	$N(1,10)$	5 quantile means	$N(0,1)$
M2C	B&T(2018) & T&O(2017)	5	500	20	$N(1,2)$	5 quantile means	$N(0,1)$

Notes: B&T(2018): Berger and Tutz (2018),T&O(2017): Tutz and Oelker (2017), $P_2 = (0.35, 0.45, 0.55, 0.55, 0.65)$, x_{it},Z_i,γ,β defined as in Table 2, Design M2.

Table S7: Simulation results M2

	β			γ		
	Bias	MAD	MSE	Bias	MAD	MSE
M2A						
POLS	0.0015	0.0272	0.0011	0.7406	0.7406	0.5755
Mundlak	0.0013	0.0235	0.0009	0.7402	0.7402	0.5749
k-Means						
k-Means, 3	0.0007	0.0240	0.0009	0.3412	0.3557	0.1755
k-Means, 5	0.0013	0.0235	0.0009	0.3590	0.4035	0.2343
k-Means, 10	0.0013	0.0234	0.0009	0.4250	0.4452	0.2740
HDBSCAN	0.0015	0.0236	0.0009	0.3085	0.7462	0.9394
HDBSCAN with LASSO						
Cross Validation	-0.1951	0.1951	0.0397	0.4277	0.6624	0.6870
Gen Cross Val	-0.0080	0.0247	0.0009	0.3143	0.7414	0.9237
BIC	-0.0081	0.0247	0.0009	0.3144	0.7415	0.9237
Cross Val Post Lasso	0.0019	0.0238	0.0009	0.3071	0.7478	0.9406
M2						
POLS	0.0022	0.0735	0.0086	3.7064	3.7064	14.3185
Mundlak	0.0013	0.0235	0.0009	3.7041	3.7041	14.3027
k-Means						
k-Means, 3	0.0032	0.0339	0.0018	4.6190	4.6235	23.3694
k-Means, 5	0.0013	0.0224	0.0008	-0.0011	0.0474	0.0035
k-Means, 10	0.0013	0.0237	0.0009	0.3208	0.3887	0.8949
HDBSCAN	0.0013	0.0224	0.0008	0.0175	0.0661	0.0903
HDBSCAN with LASSO						
Cross Validation	-0.2037	0.2037	0.0429	0.1251	0.1281	0.1015
Gen Cross Val	-0.0368	0.0399	0.0022	0.0376	0.0686	0.0906
BIC	-0.0368	0.0399	0.0022	0.0376	0.0686	0.0906
Cross Val Post Lasso	0.0013	0.0224	0.0008	0.0176	0.0663	0.0903
M2B						
POLS	0.0014	0.0692	0.0078	3.7070	3.7070	14.3176
Mundlak	0.0004	0.0078	0.0001	3.7046	3.7046	14.3016
k-Means						
k-Means, 3	0.0024	0.0266	0.0011	4.6768	4.6822	23.7525
k-Means, 5	0.0004	0.0075	0.0001	-0.0004	0.0158	0.0004
k-Means, 10	0.0006	0.0088	0.0001	0.3224	0.3462	0.8951
HDBSCAN	0.0005	0.0075	0.0001	0.0080	0.0240	0.0350
HDBSCAN with LASSO						
Cross Validation	-0.0764	0.0764	0.0060	0.0484	0.0490	0.0364
Gen Cross Val	-0.0376	0.0376	0.0015	0.0281	0.0318	0.0354
BIC	-0.0376	0.0376	0.0015	0.0281	0.0318	0.0354
Cross Val Post Lasso	0.0005	0.0075	0.0001	0.0080	0.0241	0.0350
M2C						
POLS	0.0006	0.0161	0.0004	0.7412	0.7412	0.5729
Mundlak	0.0004	0.0078	0.0001	0.7407	0.7407	0.5723
k-Means						
k-Means, 3	0.0007	0.0091	0.0001	0.8453	0.8461	0.8162
k-Means, 5	0.0004	0.0075	0.0001	0.0004	0.0179	0.0005
k-Means, 10	0.0004	0.0078	0.0001	0.0738	0.0988	0.0375
HDBSCAN	0.0004	0.0076	0.0001	0.0033	0.0252	0.0046
HDBSCAN with LASSO						
Cross Validation	-0.0679	0.0679	0.0048	0.0407	0.0436	0.0059
Gen Cross Val	-0.0105	0.0119	0.0002	0.0093	0.0256	0.0046
BIC	-0.0105	0.0119	0.0002	0.0093	0.0256	0.0046
Cross Val Post Lasso	0.0004	0.0076	0.0001	0.0035	0.0252	0.0046

Notes: Simulation Designs are defined in Table S6. Means of 500 simulations.

Table S8: Simulation Designs

Design	**M5**
Group Structure	
adapted from	B,L&M(2022)
G	N
N	500
T	20
Fixed Effect v_i	
drawn from	$N(0,1)$
discretised	none
Time-constant covariate $\boldsymbol{Z}_i$	$B(0.5)$
Time-varying covariate	$N(0,1)+v_i$
β, γ	1,1
Correlation structure	$cor(v_i, x_{it})$ ≈ 0.7
Error term u_{it}	$N(0,1)$

Notes: B,L&M (2022): Bonhomme et al. (2022).

Table S9: Simulation Results

	β			γ		
	Bias	MAD	MSE	Bias	MAD	MSE
M3						
POLS	0.4976	0.4976	0.2479	0.0033	0.0404	0.0025
Mundlak	-0.0001	0.0082	0.0001	0.0011	0.0217	0.0007
k-Means						
k-Means, 5	0.0688	0.0688	0.0049	0.0126	0.1854	0.0531
k-Means, 20	0.0092	0.0116	0.0002	-0.0030	0.1628	0.0396
k-Means, 100	0.0046	0.0093	0.0001	0.0093	0.1004	0.0163
HDBSCAN	0.0068	0.0103	0.0002	-0.0051	0.1664	0.0457
HDBSCAN with LASSO						
Cross Validation	-0.0044	0.0101	0.0002	-0.0032	0.1537	0.0387
Gen Cross Val	0.0056	0.0097	0.0002	-0.0048	0.1648	0.0448
BIC	0.0056	0.0097	0.0002	-0.0048	0.1648	0.0448

Notes: Simulation Design is defined in Table S8. Means of 500 simulations.

Smaller Time Dimension

Table S10 displays the results for Monte Carlo simulations with T=5, all other parameters are kept as in Table 2.

Clustering Evaluation

The effect of grouping individuals from different groups into the same cluster on the estimation error will depend on the difference of their true underlying group intercepts. Therefore we compute the difference between an individuals true intercepts and the mean true intercept of all individuals in the same cluster.

Table S10: Simulation results, T=5

	β			γ		
	Bias	MAD	MSE	Bias	MAD	MSE
M1						
POLS	1.5208	1.5208	2.3185	0.0011	0.1185	0.0210
Mundlak	-0.0028	0.0929	0.0137	0.0001	0.1091	0.0182
k-Means						
k-Means, 3	0.3164	0.3165	0.1101	-0.0087	0.2660	0.1124
k-Means, 5	0.1290	0.1412	0.0287	-0.0093	0.3682	0.2157
k-Means, 10	0.0363	0.0954	0.0146	-0.0257	0.3878	0.2308
HDBSCAN	0.0261	0.0970	0.0171	0.0197	0.4604	0.3490
HDBSCAN with LASSO						
Cross Validation	-0.0933	0.1258	0.0238	0.0256	0.3930	0.2536
Gen Cross Val	0.0210	0.0963	0.0168	0.0200	0.4566	0.3432
BIC	0.0200	0.0962	0.0167	0.0204	0.4555	0.3418
Cross Val Post Lasso	0.0835	0.1476	0.1063	0.0403	0.4533	0.3438
M3						
POLS	1.5987	1.5987	2.5585	-0.0040	0.0854	0.0110
Mundlak	0.0073	0.0608	0.0059	-0.0050	0.0794	0.0095
k-Means						
k-Means, 3	0.4330	0.4330	0.1916	-0.0132	0.2184	0.0775
k-Means, 5	0.1974	0.1974	0.0439	-0.0137	0.3203	0.1597
k-Means, 10	0.0664	0.0804	0.0099	-0.0393	0.3463	0.1931
HDBSCAN	0.0223	0.0632	0.0064	0.0118	0.3397	0.1841
HDBSCAN with LASSO						
Cross Validation	-0.0437	0.0715	0.0078	0.0103	0.2957	0.1410
Gen Cross Val	0.0192	0.0623	0.0062	0.0117	0.3366	0.1811
BIC	0.0180	0.0619	0.0062	0.0111	0.3354	0.1797
Cross Val Post Lasso	0.2710	0.2990	0.5435	0.0400	0.3178	0.1666
HDBSCAN	0.0223	0.0632	0.0064	0.0118	0.3397	0.1841
M4						
POLS	-0.0095	0.1132	0.0207	0.0363	0.4888	0.3698
Mundlak	-0.0006	0.0405	0.0025	0.0363	0.4898	0.3714
k-Means						
k-Means, 3	-0.0028	0.0447	0.0031	0.0012	0.2574	0.1368
k-Means, 5	-0.0008	0.0386	0.0022	0.0084	0.1250	0.0653
k-Means, 10	0.0000	0.0404	0.0025	0.0248	0.5234	0.5902
HDBSCAN	-0.0014	0.0400	0.0025	-0.1102	0.6857	2.5903
HDBSCAN with LASSO						
Cross Validation	-0.2703	0.2703	0.0780	-0.1033	0.6447	2.2937
Gen Cross Val	-0.0404	0.0528	0.0041	-0.1097	0.6815	2.5582
BIC	-0.0409	0.0531	0.0042	-0.1100	0.6815	2.5563
Cross Val Post Lasso	0.0301	0.0706	0.3161	-0.0689	0.7182	2.7161

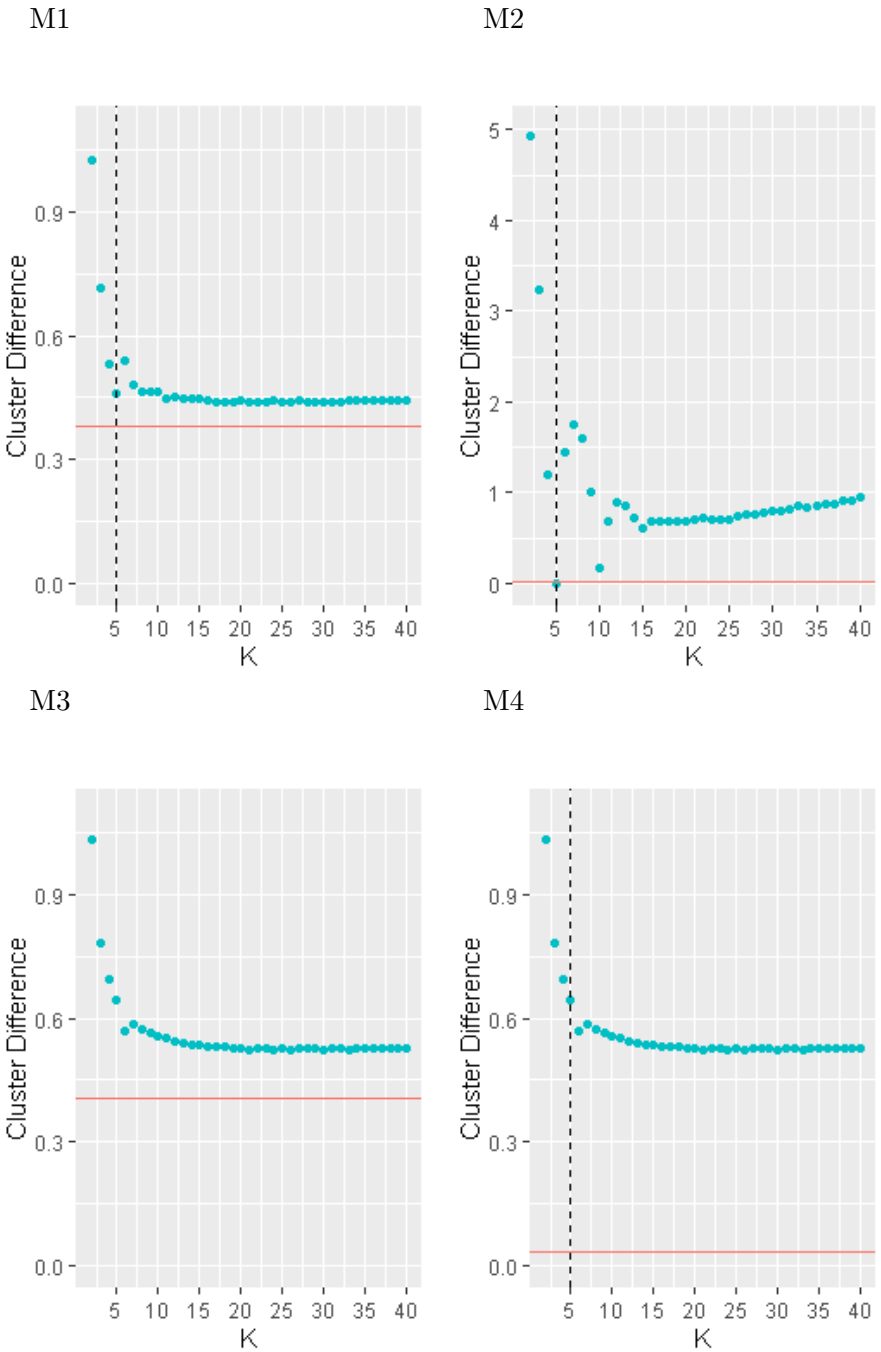
Notes: Simulations as defined in Table 2 with T=5. Means of 500 simulations.

Specifically, let $c_i \in 1, ..., C$ be the cluster where individual i was grouped into, c_I the set of all individuals grouped into this cluster and v_i i's true group intercept. Then we compute

$$CD = \frac{1}{N}(\sum_{c \in C}(\sum_{i \in c_I} |v_i - \frac{1}{|c_I| - 1} \sum_{j \in c_I, j \neq i} v_j|)). \tag{13}$$

Table S11 displays this measure for different clustering methods and across the different settings defined in Table 2. Further values of k are plotted in figure S3. Table S12 displays additional information regarding the estimated group structures in the HDBSCAN step and the LASSO step after HDBSCAN.

Figure S3: Within Cluster Differences



Notes: The blue scatterplot displays CD as defined in equation 13 for different values of k in k-Means across the simulation settings as defined in Table 2. For each setting the value for HDBSCAN with cross validation as described in Section 3 is plotted as the orange line, compare Table S11. The dashed line denotes the true groups in the settings with small number of true groups. Data source: simulations.

Table S11: Clustering Bias

	M1	M2	M3	M4
HDBSCAN with LASSO				
Cross Validation	0.3801	0.0114	0.4047	0.0338
Gen Cross Val	0.3795	0.0114	0.4035	0.0338
BIC	0.3795	0.0114	0.4035	0.0338
HDBSCAN	0.3795	0.0114	0.4035	0.0338
k-Means				
k-Means, 3	0.7156	3.2451	0.7868	1.7834
k-Means, 5	0.4634	0.0004	0.6394	0.0091
k-Means, 10	0.4639	0.1549	0.5560	0.1908

Notes: Displays the measure CD defined in equation (13). Means of 500 simulations. Simulation designs are defined in Table 2.

Table S12: Estimated Group Structure

	No Groups		No Atoms		No Groups Regularized		
	$Z=0$	$Z=1$	$Z=0$	$Z=1$	CV	GCV	BIC
M1	12.168	12.122	57.572	56.620	1.144	0.068	0.116
M2	5.058	5.042	0.808	0.624	0.008	0	0
M3	24.742	24.700	118.048	117.154	2.290	0.540	0.592
M4	5.448	5.482	9.336	9.808	0.148	0.010	0.016

Notes: Estimated Group structures by HDBSCAN and HDBSCAN with LASSO. Means across 500 simulations. Simulation designs are defined in Table 2.

S7 Illustration of Clustering Algorithms

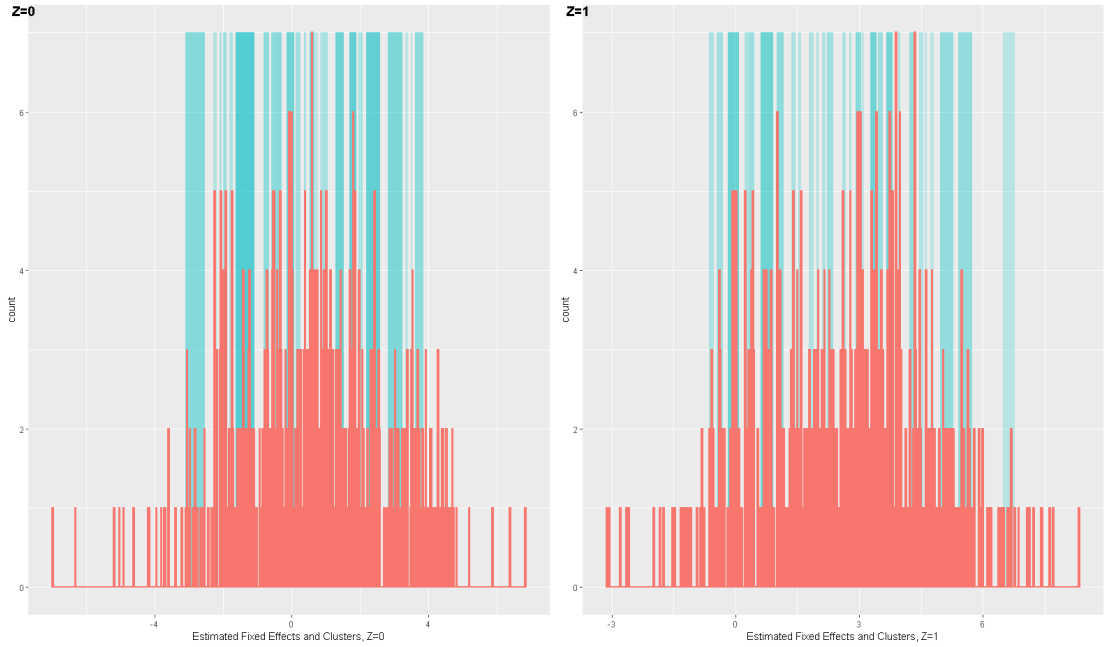
Illustration of HDBSCAN

Figure S4 illustrates the clusters computed by the HDBSCAN algorithm for simulation design M3, the first of 500 iterations is used as an example. The histogram displays the distribution of the computed fixed effects for two realisations of Z. The blue intervals display the regions of non-atomic clusters. Observations outside of these regions are labelled as atoms. Figure S5 displays the analogous picture for the first Monte Carlo realisation of Design M4.

Illustration of k-Means

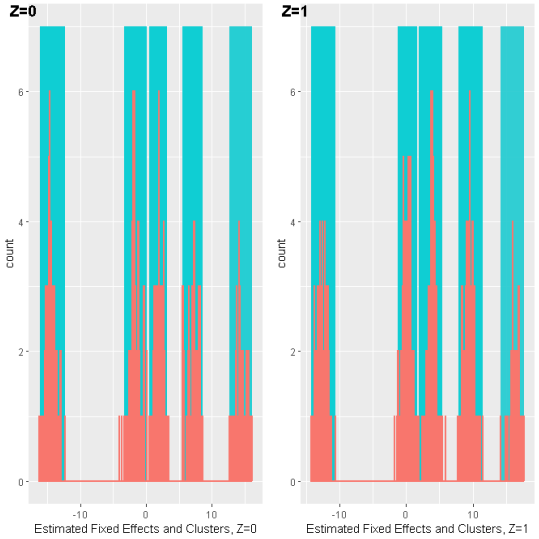
Figure S6 illustrates the clusters computed by the k-Means algorithm for simulation design M3, the first of 500 Monte Carlo iterations is used as an example. The histogram displays the distribution of the computed fixed effects for two realisations of Z. The coloured intervals display the regions of all k clusters, each cluster is illustrated with a different colour. Figure S7 displays the analogous picture for M4.

Figure S4: Estimated Fixed Effects and HDBSCAN Clusters Simulation M3



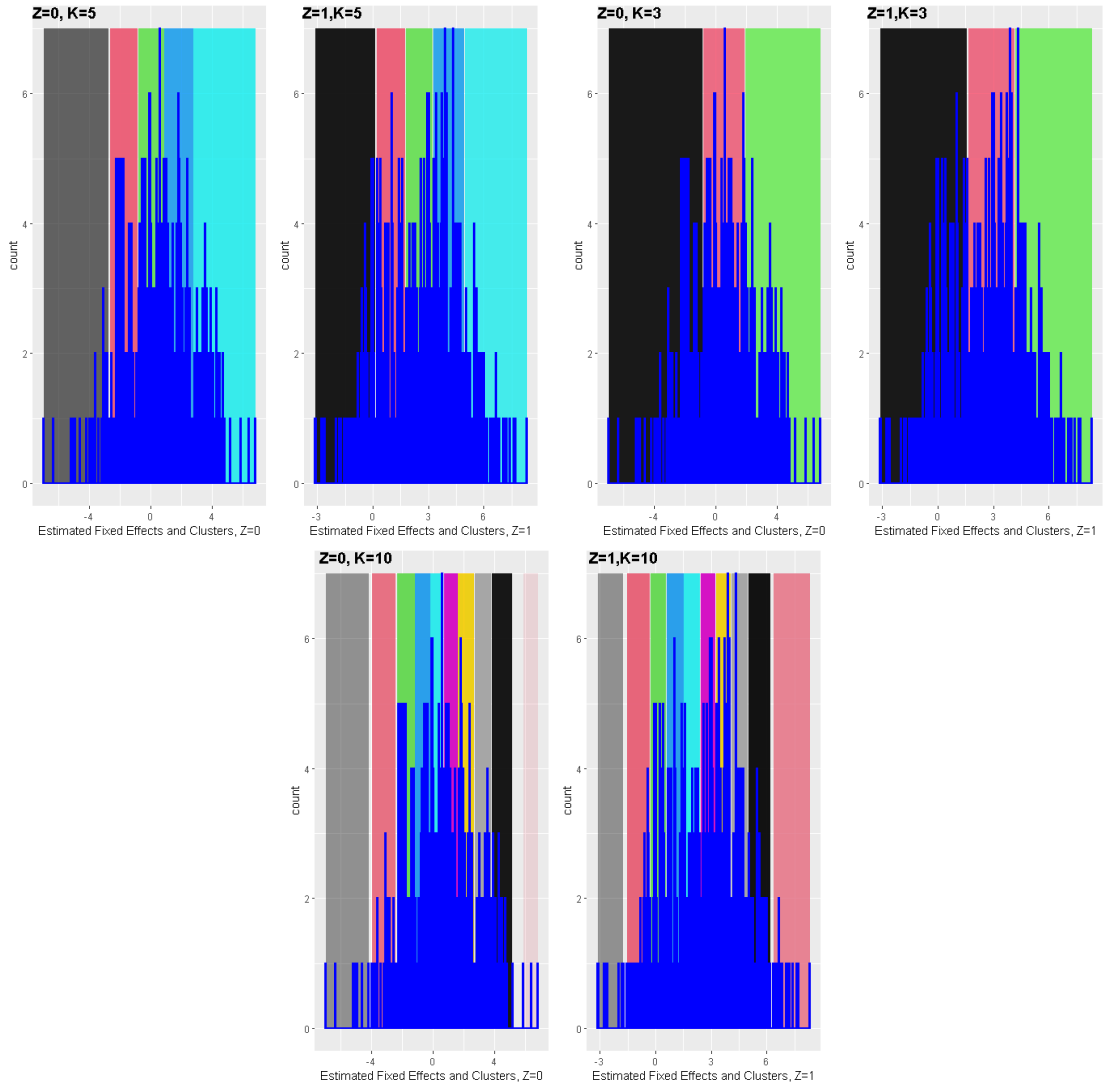
Notes: Histograms of estimated fixed effects. The blue regions indicate the intervals of non-atomic clusters computed by HDBSCAN. Dataset: first Monte Carlo realisation of Simulation Design M3 as defined in Table 2.

Figure S5: Estimated Fixed Effects and HDBSCAN Clusters Simulation M4



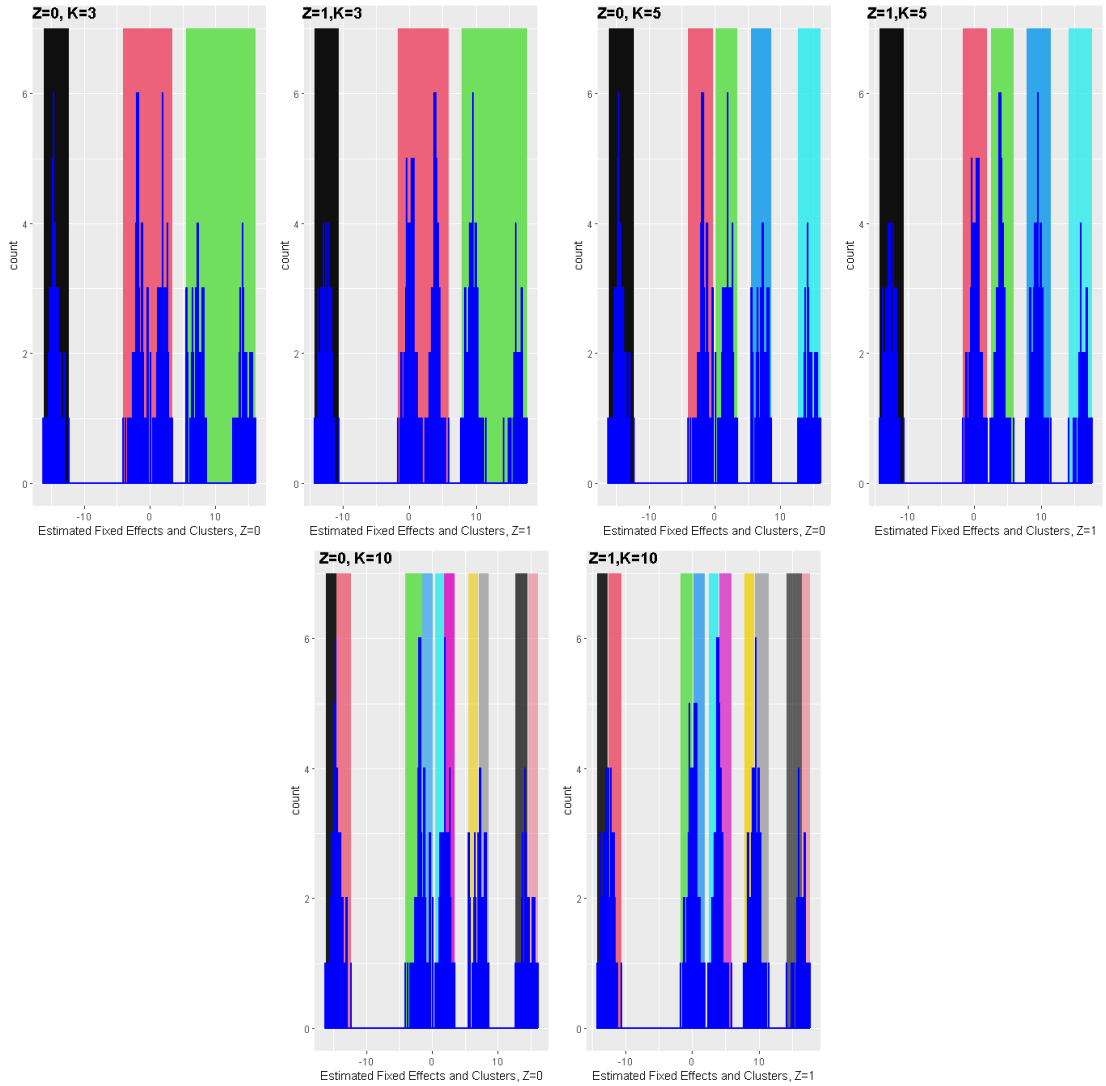
Notes: Histograms of estimated fixed effects. The blue regions indicate the intervals of non-atomic clusters computed by HDBSCAN. Dataset: first Monte Carlo realisation of Simulation Design M4 as defined in Table 2.

Figure S6: Estimated Fixed Effects and k-Means Clusters Simulation M3



Notes: Histograms of estimated fixed effects. The coloured regions indicate the intervals of k different clusters computed by k-Means. Dataset: first Monte Carlo realisation of Simulation Design M3 as defined in Table 2.

Figure S7: Estimated Fixed Effects and k-Means Clusters Simulation M4



Notes: Histograms of estimated fixed effects. The coloured regions indicate the intervals of k different clusters computed by k-Means. Dataset: first Monte Carlo realisation of Simulation Design M4 as defined in Table 2.

Download ZEW Discussion Papers:

https://www.zew.de/en/publications/zew-discussion-papers

or see:

https://www.ssrn.com/link/ZEW-Ctr-Euro-Econ-Research.html
https://ideas.repec.org/s/zbw/zewdip.html

IMPRINT

ZEW – Leibniz-Zentrum für Europäische Wirtschaftsforschung GmbH Mannheim
ZEW – Leibniz Centre for European Economic Research

L 7,1 · 68161 Mannheim · Germany
Phone +49 621 1235-01
info@zew.de · zew.de