Estimating Multi-Dimensional Group Structures in Panel Data Models with an Application to Asset Pricing*

Yousef Kaddoura†
University of Liverpool

Joakim Westerlund
Lund University
and
Deakin University

February 28, 2025

Abstract

A major concern when estimating linear regression models with panel data is that coefficients are not always poolable. A common solution to this problem is to split the sample into groups of cross-sectional units that are in some sense "similar" and to pool within those groups. However, researchers' sense of similarity is likely imperfect, and hence results likely misleading. An arguably better approach is to estimate the unknown group structure. However, similarly to manual sample splitting, most existing group estimators are "one-dimensional" in the sense that they assume that there is a single grouping of the cross-sectional units that accounts for the heterogeneity of all slope coefficients, which is unlikely to be the case in practice. To address this limitation, the present paper introduces a new estimation approach that is suitable in general when the dimensionality of the group structure is unknown. The asymptotic validity of the new approach is established and verified in small samples using Monte Carlo simulation. The empirical usefulness of the approach is illustrated through an application to asset pricing.

Keywords: Panel Data; Multi-Dimensional Group Structures; Coefficient Heterogeneity; LASSO; Asset Pricing; Stock Returns.

JEL Classification: C33; C38; C51; G12.

*Westerlund would like to thank the Knut and Alice Wallenberg Foundation for financial support through a Wallenberg Academy Fellowship.

†Corresponding author: Department of Economics, University of Liverpool Management School, Chatham St, Liverpool, L69 7ZH, UK. E-mail address: yousef.kaddoura@liverpool.ac.uk.

1 Introduction

Pooled fixed effects estimation has always been attractive to empirical economic researchers working with panel data, as it enables simple and accurate estimation of common slope coefficients while controlling for some form of unobserved heterogeneity. In practice, however, there are often reasons to suspect that the estimated coefficients are not common, making pooling potentially misleading.[1] This is reflected in the empirical literature, where researchers often split their samples and fit separate models to each part. The underlying assumption is that slope coefficients are not entirely heterogeneous, but that there are groups with common slopes that differ from those of other groups. Examples are pervasive, appearing in almost every corner of applied economics. For example, of the 95 papers published in *The American Economic Review* in 2023, 48 involve some form of panel data estimation. Of these, 36 employ splits along the lines just described, using a variety of splitting criteria deemed relevant to the specific research question.

There are several problematic aspects of the above-mentioned practice. One issue is that researchers often suspect that coefficients may vary in more than one way; that is, there may be heterogeneity in more than one "dimension".[2] Researchers therefore tend to perform many splits. However, they usually report only a subset of those results, which means that the uncertainty regarding the specification of the model is likely understated (see Athey and Imbens, 2015, and Lu and White, 2014, for discussions). Another issue is that even when the search procedure is well documented and all results are reported, the grouping of cross-sectional units is treated as if it was known. This is risky, as misspecified groups can be just as problematic as ignoring the group structure altogether. In our empirical application to asset pricing, the dependent variable is stock returns and the regressors represent different sources of risk, such as "size" and "value" (see Fama and French, 1993).[3] If we take the

[1] If coefficients are heterogeneous and this is ignored, pooled estimators will estimate a linear combination of the true coefficients, which is arguably not very informative.

[2] To take a classical example, in the productivity literature studies have emphasized the importance of allowing technology heterogeneity across countries. In the agriculture sector, for instance, this heterogeneity may reflect differences in agro-climatic environment, agricultural output mix, and level of development and commercialization (see Eberhardt and Teal, 2013, for a thorough discussion).

[3] "Size" refers to the capitalization of firms. It is based on the observation that smaller firms tend to out-

stock returns of two firms – one small and the other large – but their value is about the same, we would expect their size coefficients to differ but not their value coefficients. Thus, while in terms of value both firms belong to the same group, in terms of size they do not. Of course, this observation has not gone unnoticed in the empirical literature. Researchers have responded by splitting their samples into groups, often called "portfolios", based on certain predetermined percentiles of both size and value, and reporting estimation results for each point on the resulting grid. However, the choice of which percentiles to consider is arbitrary. This arbitrariness is problematic from a model misspecification perspective but also because it lends itself to misuse (see, for example, Ang et al., 2010, and Berk, 2000).

Observations like those just described have recently motivated researchers in econometrics to develop procedures that enable the estimation of unknown group structures. Prominent examples include Bonhomme and Manresa (2015), Lin and Ng (2012), Miao et al. (2020), Mehrabani (2023), Sarafidis and Weber (2015), Su et al. (2016), and Wang et al. (2018). Most of these studies are based on modifications of the classical k-means unsupervised learning method for clustering data points, adapted to the panel data regression context.[4] However, the k-means optimization problem is known to be "NP-hard", meaning that procedures based on it can be both slow and difficult to get to converge, and even if they do converge it may not be to the global optimum (see Chetverikov and Manresa, 2022). All studies that we are aware of require that the number of cross-sectional units, N, is large, which is typically not the case in the type of macroeconomic and financial applications that we have in mind.

The main limitation, however, which to the best of our knowledge applies to all existing

perform larger ones over time in terms of returns. This is known as the "size effect". However, while small firms may have higher returns, they also carry higher risk, because they are often more vulnerable to economic downturns and may have less resources to fight financial difficulties. By contrast, "value" refers to the book-to-market ratio of firms. It is based on the observation that firms with high book-to-market ratios, often referred to as "value firms", tend to outperform those with low ratios, often referred to as "growth firms". This is known as the "value effect". Value firms are considered undervalued by the market. These are firms that have strong fundamentals such as earnings, dividends and sales, but their market prices are low given their intrinsic value. While these firms can offer higher returns, they can also be riskier than growth firms, as they are often associated with cyclical industries.

[4] Wang et al. (2018) propose a clustering algorithm for regression via data-driven segmentation, or "CARDS". However, this approach assumes that the slope coefficients are pre-ordered, which in practice requires sequencing of the data. This is the case with a single regressor. If there are multiple regressors, it is not clear how to proceed.

studies except for Cheng et al. (2023), Cytrynbaum (2020), and Leng et al. (2023), is that each cross-sectional unit can belong to only one group. This is true not only in our asset pricing application, but in general since it seems hard – if not impossible – to a priory rule out the possibility that there may be more than one reason for why coefficients differ. As we explain in Section 4, while multi-dimensional group structures can always be represented in one dimension, the number of one-dimensional groups grows, and hence the number of members of each group decreases, with the dimensionality of the true structure. The one-dimensional representation will therefore be relatively heavily parameterized with fewer observations available for estimating each parameter. Equally as important, the economic information contained in the true group structure, which is often an object of great interest, will be lost if the dimensionality is not preserved in the estimation. There is therefore a need to move beyond the conventional one-dimensional environment and to consider group structures as potentially multi-dimensional.[5]

Cheng et al. (2023), Leng et al. (2023), and Cytrynbaum (2020) recognize the importance of allowing for multi-dimensional group structures; however, while certainly better than existing one-dimensional alternatives, the structures considered still lack generality. In particular, while the first two studies consider two group structures – one for the fixed effects and one for all the slope coefficients, the third assumes that the slopes can be divided into blocks within which the grouping is the same. All three studies assume that both N and T are large, and employ versions of the k-means algorithm, making them subject to the same critique presented earlier.

Partly inspired by one-dimensional studies such as Mehrabani (2023) and Su et al. (2016), in the present paper we view the determination of the unknown group structure as a shrinkage problem, and estimate both the regression coefficients, and the groups and their dimensionality by applying a version of the least absolute shrinkage and selection operator

[5]The concept of multi-dimensional group structures should not be mistaken for "soft" (or "fuzzy") clustering in which each data point has a probability distribution over multiple clusters. In our paper, each cross-sectional unit can belong to different coefficient clusters. But the group assignment is based on a vector of coefficients where for each individual coefficient every cross-sectional unit can only belong to one group. The clustering problem is therefore not of the soft but of the "hard" type.

(LASSO), henceforth referred to as "LASSO–MD". Unlike k-means, the LASSO solution is convex and can be easily calculated using standard algorithms. The particular objective function that we consider penalizes small differences in individual slope coefficients across all pairs of cross-sectional units. This penalty design allows for entirely different group structures for different slopes. It is therefore possible to accommodate situations like the one described at the beginning of this section, where there are multiple groups for some slopes, while only one group exists for others. We do not require that the unobserved heterogeneity is made up of fixed effects but allow for more general interactive effects in which the effects may or may not be grouped. We also do not impose any restrictions on the size of the groups, which is noteworthy, as previous literature typically assumes large groups (see, for example, Bonhomme and Manresa, 2015, Lin and Ng, 2012, Sarafidis and Weber, 2015, and Su et al., 2016). Another advantage of the new approach is that it remains asymptotically valid even if N is small and only T is large.

The rest of the paper is organized as follows: In Section 2, we present the model and the LASSO–MD procedure that we will use to estimate it. Sections 3 and 4 are concerned with the asymptotic and small-sample properties of the procedure, respectively. Section 5 focuses on our empirical asset pricing application. Section 6 concludes. All proofs are provided in an online appendix.

2 Model and procedure

Consider a scalar panel variable $y_{i,t}$, observed across $t = 1, \dots, T$ time periods and $i = 1, \dots, N$ cross-section units. The data generating process (DGP) that we consider for this variable is similar to those considered in the bulk of the previous literature (see, for example, Mehrabani, 2023, and Su et al., 2016), and is given by

$$y_{i,t} = \alpha_i + \mathbf{x}_{i,t}'\boldsymbol{\beta}_i + u_{i,t}, \tag{2.1}$$

where $\mathbf{x}_{i,t} := [x_{i,t,1}, \dots, x_{i,t,P}]'$ is a $P \times 1$ vector of known regressors with $a := b$ signifying that a is defined by b, α_i is a cross-section specific constant or "fixed effect" that may be

correlated with $\mathbf{x}_{i,t}$, and $u_{i,t}$ is an error term. We assume that $u_{i,t}$ has zero mean, which is without loss of generality since the model includes a constant, and that it is at most weakly serially and cross-sectionally correlated. In particular, since there is no guarantee that fixed effects are enough to capture all of the unobserved heterogeneity in $y_{i,t}$, similarly to Pesaran and Yamagata (2024) (see also Su and Ju, 2018), we assume that

$$u_{i,t} = \boldsymbol{\gamma}_i' \mathbf{f}_t + \varepsilon_{i,t}, \tag{2.2}$$

where $\mathbf{f}_t$ and $\boldsymbol{\lambda}_i$ are $r \times 1$ vectors of unobserved common factors and loadings, respectively, and $\varepsilon_{i,t}$ is an idiosyncratic error term. The interactive effects are here given by $\boldsymbol{\lambda}_i' \mathbf{f}_t$.

The main object of interest in this paper is the $p \times 1$ vector $\boldsymbol{\beta}_i := [\beta_{i,1}, \ldots, \beta_{i,p}]'$ of unknown slope coefficients. The cross-sectional heterogeneity of this slope vector is typically assumed to be either completely unrestricted or absent altogether so that $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_N$. However, there are as already mentioned often good reasons to believe that some cross-sectional units react similarly to changes in the regressors, and the set of similar units is likely different depending on the regressor being considered.

In the context of the Fama and French (1993) three-factor model considered in the empirical application of Section 5, $y_{i,t}$ typically represents the excess return on asset i during time period t. The regressors in $\mathbf{x}_{i,t}$ consist of a market excess-return (MKT) factor, a size factor, dubbed "small minus big (SMB)", which captures the return spread between small and large firms, and a value factor, dubbed "high minus low (HML)", which captures the return spread between high and low book-to-market firms. A fundamental proposition in asset pricing is that systematically riskier assets should earn higher expected returns. In the Fama–French three-factor model, the degree of systematic risk is measured by the slope coefficients of the factors, often referred to as "betas". Since the factors do not vary by asset, any cross-sectional variation in expected returns must come from the betas. Therefore, if risk is correctly priced, there should be a systematic relationship between betas and factors; specifically, the beta of SMB (HML) should be larger for relatively small (high book-to-market) firms (see Chordia et al., 2015, for a discussion). Thus, in terms of the notation of (2.1), we expect the elements of $\boldsymbol{\beta}_i$ to group differently.

Consider the p-th element $\beta_{i,p}$ of $\boldsymbol{\beta}_i$. The following specification of $\beta_{i,p}$ allows for the type of multi-dimensional group structures just discussed:

$$\beta_{i,p} := \begin{cases} \theta_{1,p} & \text{if} \quad i \in \mathbb{G}_{1,p} \\ \vdots & \\ \theta_{G_p,p} & \text{if} \quad i \in \mathbb{G}_{G_p,p} \end{cases}, \tag{2.3}$$

where $\theta_{g,p}$ is the distinct coefficient for group $g = 1, \ldots, G_p$ with $G_p \in \{1, \ldots, N\}$ being the total number of groups based on slope p. The set of cross-sectional units that are members of group g is here denoted $\mathbb{G}_{g,p} \subseteq \{1, \ldots, N\}$. We naturally assume that $\mathbb{G}_{g,p} \cap \mathbb{G}_{g',p} = \emptyset$ for all $g \neq g'$, so that if the i-th cross-sectional unit is assigned to group g, it cannot be assigned to any other group g'. If cross-sectional units $i \in \mathbb{G}_{g,p}$ and $j \in \mathbb{G}_{g',p}$ are assigned to different groups $g \neq g'$ based on slope coefficient p, we do not exclude the possibility that $i, j \in \mathbb{G}_{g,p'}$ for $p \neq p'$, so that the two units are assigned to the same group g based on slope $p' \neq p$.[6]

We do not impose any restrictions on the size of the groups. Hence, there might just be one group ($G_p = 1$) containing all N cross-sectional units, so that $\beta_{1,p} = \cdots = \beta_{N,p}$, but there can also be N groups ($G_p = N$) with just one cross-sectional unit in each, so that $\beta_{i,p} = \beta_{j,p}$ for all $i \neq j$. The fact that in the present paper there are no restrictions in this regard is noteworthy because in the existing literature it is standardly assumed that the size of the groups is expanding with N (see, for example, Bonhomme and Manresa, 2015, Lin and Ng, 2012, Sarafidis and Weber, 2015, and Su et al., 2016). Not requiring large groups is a great advantage in practice because estimated groups can sometimes be very small, as we illustrate in Section 5.

Denote by $\mathbb{G}_p := \{\mathbb{G}_{1,p}, \ldots, \mathbb{G}_{G_p,p}\}$ the set consisting of all the group sets for slope p and let $\boldsymbol{\theta}_p := [\theta_{1,p} \ldots, \theta_{G_p,p}]'$ be the corresponding $G_p \times 1$ vector of distinct group-specific coefficients. The set of all group sets for all slopes is given by $\mathbb{G} := \{\mathbb{G}_1, \ldots, \mathbb{G}_P\}$ with $\boldsymbol{\theta} := [\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_P']'$ being the $\sum_{p=1}^{p} G_p \times 1$ vector containing all distinct coefficients. These are the

[6] Here now is an example that illustrates our notation: Suppose that $P = 3$ and $N = 5$. Suppose also that the vectors of slope coefficients are given by $\boldsymbol{\beta}_1 = [0,1,0]'$, $\boldsymbol{\beta}_2 = [1,2,0]'$, $\boldsymbol{\beta}_3 = [0,2,0]'$, $\boldsymbol{\beta}_4 = [0,1,0]'$ and $\boldsymbol{\beta}_5 = [1,3,0]'$. The first element of these vectors takes on two distinct values, 0 and 1, which means that $G_1 = 2$. The groups are given by $\mathbb{G}_{1,1} = \{1,3,4\}$ and $\mathbb{G}_{2,1} = \{2,5\}$ for which $\beta_{i,1} = \theta_{1,1} = 0$ and $\beta_{i,1} = \theta_{2,1} = 1$, respectively. The second slope element takes on three distinct values, 1, 2 and 3, and therefore $G_2 = 3$. The groups are given by $\mathbb{G}_{1,2} = \{1,4\}$, $\mathbb{G}_{2,2} = \{2,3\}$ and $\mathbb{G}_{2,3} = 3$. The third and final element is always equal to 0, and therefore $G_3 = 1$ and $\mathbb{G}_{1,3} = \{1, \ldots, 5\}$.

objects of interest. Let us therefore denote by $\mathbb{G}^0 := \{\mathbb{G}_1^0, \ldots, \mathbb{G}_P^0\}$ and $\boldsymbol{\theta}^0 := [\boldsymbol{\theta}_1^{0\prime}, \ldots, \boldsymbol{\theta}_P^{0\prime}]'$ the true values of $\mathbb{G}$ and $\boldsymbol{\theta}$, respectively, where $\mathbb{G}_p^0 := \{\mathbb{G}_{1,p}^0, \ldots, \mathbb{G}_{G_p^0,p}^0\}$ and $\boldsymbol{\theta}_p^0 := [\theta_{1,p}^0 \ldots, \theta_{G_p^0,p}^0]'$ with G_p^0 being the true value of G_p. The true value of $\boldsymbol{\beta}_i$ is henceforth denoted $\boldsymbol{\beta}_i^0 := [\beta_{i,1}^0, \ldots, \beta_{i,P}^0]'$. The LASSO–MD estimator can be implemented in three steps.

Step 1 (Demeaning)**.** We begin by eliminating the fixed effects by demeaning the data. This gives $\widetilde{y}_{i,t} := y_{i,t} - T^{-1}\sum_{s=1}^T y_{i,s}$ and $\widetilde{\mathbf{x}}_{i,t} := \mathbf{x}_{i,t} - T^{-1}\sum_{s=1}^T \mathbf{x}_{i,s}$, which can be stacked over time to obtain the following version of (2.1):

$$\widetilde{\mathbf{y}}_i = \widetilde{\mathbf{X}}_i\boldsymbol{\beta}_i^0 + \widetilde{\mathbf{u}}_i, \tag{2.4}$$

where $\widetilde{\mathbf{y}}_i := [\widetilde{y}_{i,1}, \ldots, \widetilde{y}_{i,T}]'$ and $\widetilde{\mathbf{u}}_i := [\widetilde{u}_{i,1}, \ldots, \widetilde{u}_{i,T}]'$ are $T \times 1$ vectors, while $\widetilde{\mathbf{X}}_i := [\widetilde{\mathbf{x}}_{i,1}, \ldots, \widetilde{\mathbf{x}}_{i,T}]'$ is a $T \times P$ matrix.

Step 2 (Estimating $\mathbb{G}^0$)**.** In the second step, we estimate $\mathbb{G}^0$ by minimizing the following objective function:

$$\mathcal{L}(\boldsymbol{\beta}) := \frac{1}{T}\sum_{i=1}^N \|\widetilde{\mathbf{y}}_i - \widetilde{\mathbf{X}}_i\boldsymbol{\beta}_i\|^2 + \lambda\sum_{i=1}^N\sum_{j=i+1}^N\sum_{p=1}^P \omega_{i,j,p}|\beta_{i,p} - \beta_{j,p}|, \tag{2.5}$$

where $\boldsymbol{\beta} := [\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_N']'$ is a $NP \times 1$ vector, $\|\mathbf{A}\| := \sqrt{\text{tr}(\mathbf{A}'\mathbf{A})}$ is the Frobenius norm of the generic matrix $\mathbf{A}$, $\lambda = \lambda(T) > 0$ is a tuning parameter and $\omega_{i,j,p}$ is a certain adaptive weight. Specifically, $\omega_{i,j,p} := |\dot{\beta}_{i,p} - \dot{\beta}_{j,p}|^{-\kappa}$, where $\kappa > 0$ is a user-specified constant, and $\dot{\beta}_{i,p}$ is the p-th element of the unit-specific ordinary least squares (OLS) estimator obtained by minimizing the first term on the right-hand side of (2.5) with respect to $\boldsymbol{\beta}_i$;

$$\dot{\boldsymbol{\beta}}_i := \begin{bmatrix} \dot{\beta}_{i,1} \\ \vdots \\ \dot{\beta}_{i,P} \end{bmatrix} = (\widetilde{\mathbf{X}}_i'\widetilde{\mathbf{X}}_i)^{-1}\widetilde{\mathbf{X}}_i'\widetilde{\mathbf{y}}_i. \tag{2.6}$$

By minimizing $\mathcal{L}(\boldsymbol{\beta})$, we obtain $\widehat{\boldsymbol{\beta}} := [\widehat{\boldsymbol{\beta}}_1', \ldots, \widehat{\boldsymbol{\beta}}_N']'$, where $\widehat{\boldsymbol{\beta}}_i := [\widehat{\beta}_{i,1}, \ldots, \widehat{\beta}_{i,P}]'$. With this estimate in hand, we check whether $|\widehat{\beta}_{i,p} - \widehat{\beta}_{j,p}| = 0$ or $|\widehat{\beta}_{i,p} - \widehat{\beta}_{j,p}| > 0$, and assign cross-sectional units i and j to the same group if the former condition holds and to different groups if the latter condition holds. This way we obtain estimators of both the group membership of each cross-sectional unit and the number of groups for each p. The resulting estimator of

$\mathbb{G}^0$ is henceforth denoted $\widehat{\mathbb{G}} := \{\widehat{\mathbb{G}}_1, \ldots, \widehat{\mathbb{G}}_P\}$, where $\widehat{\mathbb{G}}_p := \{\widehat{\mathbb{G}}_{1,p}, \ldots, \widehat{\mathbb{G}}_{\widehat{G}_p,p}\}$ with $\widehat{G}_p$ being the number of distinct coefficient estimates of slope p.

Step 3 (Estimating $\boldsymbol{\theta}^0$)**.** In the third and final step, we estimate $\boldsymbol{\theta}^0$ conditional on the Step-2 estimate of $\mathbb{G}^0$. The implementation of this step is inspired by the literature on structural breaks and thresholds (see, for example, Hansen, 1999). The basic idea is to first move the unknown group structure in $\mathbb{G}^0$ from $\boldsymbol{\beta}_i^0$ to the regressors using dummy variables, and then to replace $\mathbb{G}^0$ by $\widehat{\mathbb{G}}$.

We begin by nothing that $\beta_{i,p}^0$ and $\theta_{g,p}^0$ are related through $\beta_{i,p}^0 = \sum_{g=1}^{G_p^0} \mathbb{1}(i \in \mathbb{G}_{g,p}^0)\theta_{g,p}^0$, where $\mathbb{1}(A)$ is the indicator function for the event A, which equals 1 (0) if A is true (false). Let us therefore denote by $\widetilde{\mathbf{X}}_{i,p} := [\widetilde{x}_{i,1,p}, \ldots, \widetilde{x}_{i,T,p}]'$ the p-th column of $\widetilde{\mathbf{X}}_i$. Define the $T \times 1$ vector $\widetilde{\mathbf{X}}_{i,p}(\mathbb{G}_{g,p}) := \mathbb{1}(i \in \mathbb{G}_{g,p})\widetilde{\mathbf{X}}_{i,p}$, the $T \times G_p$ matrix $\widetilde{\mathbf{X}}_{i,p}(\mathbb{G}_p) := [\widetilde{\mathbf{X}}_{i,p}(\mathbb{G}_{1,p}), \ldots, \widetilde{\mathbf{X}}_{i,p}(\mathbb{G}_{G_p,p})]$, and the $T \times \sum_{p=1}^P G_p$ matrix $\widetilde{\mathbf{X}}_i(\mathbb{G}) := [\widetilde{\mathbf{X}}_{i,1}(\mathbb{G}_1), \ldots, \widetilde{\mathbf{X}}_{i,P}(\mathbb{G}_P)]$. In this notation, note how we have $\widetilde{\mathbf{X}}_i\boldsymbol{\beta}_i^0 = \sum_{p=1}^P \widetilde{\mathbf{X}}_{i,p}\beta_{i,p}^0 = \sum_{p=1}^P \sum_{g=1}^{G_p^0} \widetilde{\mathbf{X}}_{i,p}(\mathbb{G}_{g,p}^0)\theta_{g,p}^0 = \sum_{p=1}^P \widetilde{\mathbf{X}}_{i,p}(\mathbb{G}_p^0)\boldsymbol{\theta}_p^0 = \widetilde{\mathbf{X}}_i(\mathbb{G}^0)\boldsymbol{\theta}^0$. The model in (2.4) can therefore be rewritten very conveniently in the following way:

$$\widetilde{\mathbf{y}}_i = \widetilde{\mathbf{X}}_i(\mathbb{G}^0)\boldsymbol{\theta}^0 + \widetilde{\mathbf{u}}_i, \tag{2.7}$$

The proposed LASSO–MD estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^0$ is simply the pooled OLS estimator of (2.8) with $\widehat{\mathbb{G}}$ in place of $\mathbb{G}^0$;

$$\widehat{\boldsymbol{\theta}} := \begin{bmatrix} \widehat{\boldsymbol{\theta}}_1 \\ \vdots \\ \widehat{\boldsymbol{\theta}}_P \end{bmatrix} = \left(\sum_{i=1}^N \widetilde{\mathbf{X}}_i(\widehat{\mathbb{G}})'\widetilde{\mathbf{X}}_i(\widehat{\mathbb{G}}) \right)^{-1} \sum_{i=1}^N \widetilde{\mathbf{X}}_i(\widehat{\mathbb{G}})'\widetilde{\mathbf{y}}_i. \tag{2.8}$$

Some remarks are in order. In most situations of empirical relevance, the fixed effect α_i is a nuisance parameter that is not of any particular interest by itself. However, since α_i may be correlated with $\mathbf{x}_{i,t}$, it can also not be ignored or else our proposed LASSO–MD approach will be rendered omitted variables biased. It is therefore essential to be able to appropriately control for α_i. This is where the demeaning in Step 1 comes in. It ensures that LASSO–MD is exactly invariant with respect to α_i.

However, there are situations in which α_i is an object of interest. In our empirical application, α_i is the excess return on the zero-beta asset, often referred to as that asset's "alpha". If riskless borrowing and lending are allowed, the zero-beta asset earns the risk-free rate and its excess return is zero, that is, $\alpha_i = 0$. The Fama–French three-factor model predicts that only the MKT, SMB and HML factors are priced. A rejection of the null hypothesis of $\alpha_i = 0$ therefore means that these three factors are not enough to explain the average level of returns in excess of the risk free rate.[7] In this case, α_i measures the degree of "mispricing", which may well have a group structure, just like $\boldsymbol{\beta}_i$.

If the purpose is not just to eliminate α_i but to estimate it, then Step 1 of the estimation procedure must be suitably modified. This is done by replacing $\widetilde{y}_{i,t}$, $\widetilde{\mathbf{x}}_{i,t}$ and $\boldsymbol{\beta}_i^0$ by $y_{i,t}$, $(1, \mathbf{x}_{i,t}')'$ and $(\alpha_i, \boldsymbol{\beta}_i^{0\prime})'$, respectively. The resulting model for $\widetilde{\mathbf{y}}_i$ has exactly the same from as in (2.4) but with $\widetilde{\mathbf{X}}_i$ and $\boldsymbol{\beta}_i^0$ being of dimension $T \times (P+1)$ and $(P+1) \times 1$, respectively. Given this modification, Steps 2 and 3 are unaffected.

A word about Step 2: The second term appearing on the right-hand side of (2.5) is a penalty that to the best of our knowledge has not been used in the literature before. The paper that is closest to ours in this regard is Mehrabani (2023). However, his penalty is of the following group fused LASSO type: $\lambda \sum_{i=1}^{N} \sum_{j=i+1}^{N} \omega_{i,j} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|$, where $\omega_{i,j} := \|\dot{\boldsymbol{\beta}}_i - \dot{\boldsymbol{\beta}}_j\|^{-\kappa}$. The use of the Frobenius norm here means that Mehrabani's version of the LASSO penalizes small differences in entire vectors of slopes, which means that it shrinks those vectors towards one another and assigns the associated cross-sectional units to the same group. Conversely, sufficiently large differences causes the procedure to assign the associated units to different groups, and it does so regardless of whether those differences emanate from all vector elements or just a subset. This means that the estimated group structure will be one-dimensional even if the true structure is in fact multi-dimensional. In our Monte Carlo and empirical studies, we elaborate on this point.

The above concerns motivate the use of a penalty of the "ℓ_1-type" that penalizes individual slope differences using the ℓ_1 norm, as opposed to vector differences using the Frobenius

[7] The interactive effects model in (2.2) provides a means to allow for possible missing factors.

norm. As a result, the proposed penalty is suitable for estimating group structures that are slope-specific. To aid the group selection, we use adaptive weights where high initial difference estimates receive relatively low penalties, and vice versa.

3 Asymptotic results

Unlike existing studies, the results reported in this section do not require $N \to \infty$. In fact, as long as $N \geq 2$, such that $\beta_{i,p} - \beta_{j,p}$ can be computed at least once for every p, N does not even have to be large. This is an advantage, because in practice N is always finite. However, we do require $T \to \infty$. This means that the asymptotic approximation offered by our results is expected to work well in "long" panels where T is larger than N. The main drawback is that our theory is silent about the effect of increasing N. But then again in practice N is fixed. We now state the conditions that underlie our fixed-N theory.

Assumption 1.

(a) $\varepsilon_{i,s}$ is mean zero and independent of $\mathbf{x}_{i,t}$ for all i, t and s with $\sup_{1\leq i\leq N, 1\leq t\leq T} \mathbb{E}(\|\mathbf{x}_{i,t}\|^4) < \infty$, $\sup_{1\leq i\leq N, 1\leq t\leq T} \mathbb{E}(\varepsilon_{i,t}^4) < \infty$ and $\sup_{1\leq i\leq N} T^{-1} \sum_{t=1}^{T} \sum_{s=1}^{T} |\mathbb{E}(\widetilde{u}_{i,t} \widetilde{\mathbf{x}}_{i,s}' \widetilde{\mathbf{x}}_{i,t} \widetilde{u}_{i,s})| < \infty$.

(b) For $\mathbf{Q}_i := T^{-1} \widetilde{\mathbf{X}}_i' \widetilde{\mathbf{X}}_i$, we have $\inf_{1\leq i\leq N} \mu_{\min}(\mathbf{Q}_i) > 0$ with probability 1 (w.p.1) and $\mathbf{Q}_i \to_p \mathbf{Q}_{0,i} = \lim_{T\to\infty} \mathbb{E}(\mathbf{Q}_i)$ as $T \to \infty$, where $\inf_{1\leq i\leq N} \mu_{\min}(\mathbf{Q}_{0,i}) > 0$, $\mu_{\min}(\mathbf{A})$ signifies the smallest singular value of any matrix $\mathbf{A}$, and $\to_p$ signifies convergence in probability.

(c) $\mathbf{f}_t$ is mean zero and independent of $(\varepsilon_{i,s}, \mathbf{x}_{i,s})$ for all i, t and s with $\sup_{1\leq t\leq T} \mathbb{E}(\|\mathbf{f}_t\|^4) < \infty$. Also, $\boldsymbol{\lambda}_i$ is non-random with $\sup_{1\leq i\leq N} \|\boldsymbol{\lambda}_i\| < \infty$.

Assumption 2. For $\Delta_{\min} := \min_{1\leq k<k'\leq G_p, 1\leq p\leq p} |\theta_{k,p}^0 - \theta_{k',p}^0|$, we have $\sqrt{T}\lambda\Delta_{\min}^{-\kappa} \to c_1 < \infty$, $\sqrt{T}\Delta_{\min} \to c_2 \in (0, \infty]$ and $T^{(\kappa+1)/2}\lambda \to \infty$ as $T \to \infty$. Also, $\|\boldsymbol{\theta}^0\| < \infty$.

Assumption 3.

(a) For $\mathbf{Q}(\mathbb{G}^0) := T^{-1} \sum_{i=1}^{N} \widetilde{\mathbf{X}}_i(\mathbb{G}^0)' \widetilde{\mathbf{X}}_i(\mathbb{G}^0)$, we have $\mu_{\min}(\mathbf{Q}(\mathbb{G}^0)) > 0$ w.p.1 and $\mathbf{Q}(\mathbb{G}^0) \to_p \mathbf{Q}_0 := \lim_{T\to\infty} \mathbb{E}(\mathbf{Q}(\mathbb{G}^0))$ as $T \to \infty$, where $\mu_{\min}(\mathbf{Q}_0) > 0$ w.p.1.

(b) $\sqrt{T}\mathbf{U}(\mathbb{G}^0) \to_d \mathcal{N}(\mathbf{0}_{\sum_{p=1}^{P} G_p^0 \times 1}, \mathbf{\Sigma}_0)$ as $T \to \infty$, where $\mathbf{U}(\mathbb{G}) := T^{-1}\sum_{i=1}^{N} \widetilde{\mathbf{X}}_i(\mathbb{G})'\widetilde{\mathbf{u}}_i$, $\to_d$ signifies convergence in distribution, and $\mathcal{N}(\cdot,\cdot)$ denotes a normal distribution.

The estimation of the groups in Step 2 of the LASSO–MD estimation procedure is based on the unit-specific LASSO estimates $\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_N$ contained in $\widehat{\boldsymbol{\beta}}$. Assumption 1 ensures that these estimates are consistent and is not particularly restrictive. The condition in part (a) that $\varepsilon_{i,t}$ and $\mathbf{x}_{i,t}$ are independent of each other rules out the presence of lagged dependent variables in $\mathbf{x}_{i,t}$. This can be relaxed but then $\varepsilon_{i,t}$ must be serially uncorrelated.[8] The absolute summability of the autocovariances in the same part ensures that any serial correlation in $\widetilde{\mathbf{x}}_{i,t}\widetilde{u}_{i,t}$ is of the weak form. Part (b) is a standard non-collinearity condition that is easy to understand. It ensures that $T^{-1}\widetilde{\mathbf{X}}_i'\widetilde{\mathbf{X}}_i$ is invertible, which in turn rules out time invariant regressors.[9] As pointed out in Section 2, because of the demeaning, LASSO–MD is exactly invariant with respect to the fixed effect α_i; however, it is not invariant with respect to the interactive effects in $\boldsymbol{\gamma}_i'\mathbf{f}_t$. Assumption 1 (c), which is similar to Assumption 2 in Pesaran and Yamagata (2024), ensures that the latter effects do not interfere with the asymptotic validity of LASSO–MD.

Assumption 2 places restrictions on the tuning parameter λ and on $\Delta_{\min}$, which is our measure of minimum degree of slope heterogeneity required for it to be possible to separate one group from another. These are needed to ensure the consistency of the estimated group structure. If $\Delta_{\min} = O(T^{-1/2+\delta})$ for $\delta > 0$, the conditions placed on λ are satisfied if $\lambda \propto T^{-\rho}$ with $\rho \in [0, 1/2]$, which are mild enough to enable data-driven selection via an information criterion, as we explain in Section 4. The conditions placed on $\Delta_{\min}$ are also mild, yet more restrictive than those employed in most one-dimensional studies, as to be expected because of our more general group structure.[10]

[8] An advantage of the fixed-N asymptotic setup considered here is that lagged dependent variables can be permitted without for that matter requiring bias correction, which is known to be a difficult task in practice, to the point that it is sometimes better to simply ignore the issue altogether (see Judson and Owen, 1999, for a general discussion).

[9] Time invariant regressors in $\mathbf{x}_{i,t}$ has to be ruled out even if the data are not demeaned, as in the case discussed in Section 2 when the fixed effects are parameters of interest. In this case, a constant is appended to $\mathbf{x}_{i,t}$, which means that there can be no other time invariant regressors.

[10] For example, Mehrabani (2023) requires that $\sqrt{T}\min_{1\leq g<g'\leq G_p} \|\boldsymbol{\theta}_g^0 - \boldsymbol{\theta}_{g'}^0\|$ is positive in the limit. This

Assumption 3 is a "high-level" condition concerned with the LASSO–MD estimator of $\boldsymbol{\theta}^0$.[11] Part (a) does for this estimator what Assumption 1 (b) does for the unit-specific LASSO estimator. Part (b) is a central limit theorem that holds quite generally. Note in particular that since the assumption only calls for a large T, the cross-sectional properties of $\{\widetilde{\mathbf{X}}_i(\mathbb{G}^0)'\widetilde{\mathbf{u}}_i\}_{i=1}^N$ are essentially unrestricted.

An important point about Assumptions 1–3 worth reiterating is that they do not place any conditions on the group structure contained in $\mathbb{G}^0$. Hence, both the number of groups, G_p^0, and the number of units within each group, $\mathbb{G}_{g,p}$, are completely unrestricted. With these assumptions, we are ready to state our first result.

Theorem 1. *Suppose that Assumptions 1 and 2 hold. Then, uniformly in $i = 1, \ldots, N$,*

$$\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^0\| = O_p(T^{-1/2}).$$

Theorem 1 establishes that the unit-specific LASSO estimator is consistent and that the rate of convergence is given by $T^{-1/2}$. The theorem is important in itself but also because it is needed to establish our next theorem.

Theorem 2. *Suppose that Assumptions 1 and 2 hold. Then, uniformly in $p = 1, \ldots, P$, as $T \to \infty$,*

$$\mathbb{P}\big(|\widehat{\beta}_{i,p} - \widehat{\beta}_{j,p}| = 0 \text{ for all } i, j \in \mathbb{G}_{g,p}^0 \text{ and } g = 1, \ldots, G_p^0\big) \to 1.$$

Theorem 2 says that with probability approaching one $|\beta_{i,p}^0 - \beta_{j,p}^0|$ will be estimated to zero exactly whenever cross-sectional units i and j are members of the same group. But by Theorem 1 we know that $\widehat{\boldsymbol{\beta}}_i - \widehat{\boldsymbol{\beta}}_j$ is consistent for $\boldsymbol{\beta}_i^0 - \boldsymbol{\beta}_j^0$ for all i and j, regardless of whether they belong to the same group or not. This implies that the procedure can correctly identify the unknown group structure. The following corollary formalizes this.

condition is implied by our Assumption 2, as seen by noting that $\|\boldsymbol{\theta}_g^0 - \boldsymbol{\theta}_{g'}^0\|^2 = \sum_{p=1}^P |\theta_{g,p}^0 - \theta_{g',p}^0|^2 \geq P(\min_{1\leq p\leq P} |\theta_{g,p}^0 - \theta_{g',p}^0|)^2$.

[11] Assumption 3 holds under a variety of more primitive conditions. The high-level formulation is chosen in part because it is standard in the literature (see, for example, Mehrabani, 2023, Su and Ju, 2018, and Wang et al., 2018), in part because we are not particularly interested in the sets of primitive conditions under which Assumption 3 holds.

Corollary 1. *Suppose that the conditions of Theorem 2 hold. Then, uniformly in* $p = 1, \ldots, P$*, as* $T \to \infty$*,*

(a) $\mathbb{P}\big(\widehat{G}_p = G_p^0\big) \to 1$;

(b) $\mathbb{P}\big(\widehat{\mathbb{G}}_p = \mathbb{G}_p^0 | \widehat{G}_p = G_p^0\big) \to 1$.

Theorem 3 reports the asymptotic distribution of the LASSO–MD estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^0$, and it does so conditional on the event that $\widehat{\mathbb{G}} = \mathbb{G}^0$, which according to Corollary 1 holds with high probability.

Theorem 3. *Suppose that Assumptions 1–3 hold, and that* $\widehat{\mathbb{G}} = \mathbb{G}^0$*. Then, as* $T \to \infty$*,*

$$\sqrt{T}\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\big) \to_d \mathcal{N}\big(\mathbf{0}_{\sum_{p=1}^P G_p^0 \times 1}, \mathbf{Q}_0^{-1}\boldsymbol{\Sigma}_0\mathbf{Q}_0^{-1}\big).$$

Theorem 3 supports asymptotically valid standard normal and chi-squared inference. Of course, for such standard inference to be possible, we need a consistent estimator of $\mathbf{Q}_0^{-1}\boldsymbol{\Sigma}_0\mathbf{Q}_0^{-1}$. In view of Corollary 1, a naturally consistent estimator of $\mathbf{Q}_0$ is given by $\mathbf{Q}(\widehat{\mathbb{G}})$. For $\boldsymbol{\Sigma}_0$, we employ the usual heteroskedasticity and autocorrelation (HAC) robust estimator. Denote by $\mathbf{U}_t(\mathbb{G})'$ the t-th row of $\sum_{i=1}^N \widetilde{\mathbf{X}}_i(\mathbb{G})'\widetilde{\mathbf{u}}_i$ and by $\widehat{\mathbf{u}}_i := \widetilde{\mathbf{y}}_i - \widetilde{\mathbf{X}}_i(\widehat{\mathbb{G}})\widehat{\boldsymbol{\theta}}$ the residual obtained from the fit of (2.8). In this notation, $\boldsymbol{\Sigma}_0$ is identically the long-run covariance matrix of $T^{-1/2}\sum_{t=1}^T \mathbf{U}_t(\mathbb{G})$, which we estimate by applying the kernel method of Newey and West (1987) to $\sum_{i=1}^N \widetilde{\mathbf{X}}_i(\widehat{\mathbb{G}})'\widehat{\mathbf{u}}_i$. Similar arguments have been used before by, for example, Bai (2009).

The above results assume that the data have been demeaned prior to the application of LASSO–MD, as prescribed by Step 1 of the three-step estimation procedure. If the fixed effects are not demeaned away but included as additional regressors as discussed in Section 2, LASSO–MD can be used to test their significance. Within the context of our empirical application, testing the null hypothesis that $\alpha_1 = \ldots = \alpha_N = 0$ is tantamount to testing the suitability of the Fama–French three-factor model as an asset pricing model. It has therefore attracted considerable attention and there is by now a separate literature devoted to it (see Pesaran and Yamagata, 2024, for a recent overview). However, most tests are multivariate, and their performance can be unacceptably poor unless N (T) is very small (large).

Attempts have been made to combat this problem by imposing sparsity on the $N \times N$ covariance matrix of $u_{1,t}, \ldots, u_{N,t}$, and basing estimation on adaptive thresholding or other similar techniques. The present paper can be seen as another attempt in the same direction. It puts the sparsity assumption on the differences of the alphas, and uses LASSO–MD to estimate the resulting groups. Theorem 3 justifies standard significance tests of the grouped alphas.

4 Monte Carlo study

In this section, we report the results of a small-scale Monte Carlo study aimed in part at assessing the accuracy of our theoretical results, in part at evaluating the relative performance of LASSO–MD when compared to the penalized least squares (PLS) approach of Mehrabani (2023), which is arguably the closest competitor in the one-dimensional case. The DGP used for this purpose is given by a restricted version of the one in (2.1) based on setting $p = 3$ and $x_{i,t,p} = 0.2\alpha_i + e_{i,t,p}$, and drawing $e_{i,t,p}$ and α_i independently from $\mathcal{N}(0,1)$ (similarly to Wang et al., 2018). For the error term $u_{i,t}$, we consider the following DGP (taken from Bai and Ng, 2002):

$$u_{i,t} = \pi_1 u_{i,t-1} + \chi_{i,t} + \sum_{j=1}^{J} \pi_2 (\chi_{i-j,t} + \chi_{i+j,t}), \tag{4.1}$$

where $J = 10$, $\pi_1 = 0.6$, $\pi_2 = 0.1$, and $\chi_{i,t} \sim \mathcal{N}(0, \sigma_i^2)$ such that $\sigma_i \sim \mathcal{U}(0.5, 1)$ with $\mathcal{U}(\cdot,\cdot)$ being the uniform distribution. Hence, the errors are not only serially and cross-sectionally correlated but also heteroskedastic. Of course, in this paper the most important feature of the DGP is not the errors, but the slopes, $\boldsymbol{\beta}_i = [\beta_{i,1}, \beta_{i,2}, \beta_{i,3}]'$. Three DGPs are considered.

DGP1 (Multi-dimensional group structure)**.**

$$\beta_{i,1} = \begin{cases} 0 & \text{for} \quad i \in \mathbb{G}_{1,1}^0 = \{1, \ldots, 0.3N\} \\ 1 & \text{for} \quad i \in \mathbb{G}_{2,1}^0 = \{0.3N+1, \ldots, 0.6N\} \\ 2 & \text{for} \quad i \in \mathbb{G}_{3,1}^0 = \{0.6N+1, \ldots, N\} \end{cases},$$

$$\beta_{i,2} = \begin{cases} 0.5 & \text{for} \quad i \in \mathbb{G}_{1,2}^0 = \{1, \ldots, 0.3N\} \\ 1.5 & \text{for} \quad i \in \mathbb{G}_{2,2}^0 = \{0.3N+1, \ldots, N\} \end{cases}',$$

$$\beta_{i,3} = 3 \quad \text{for} \quad i \in \mathbb{G}_{1,3}^0 = \{1, \ldots, N\}.$$

DGP2 (One-dimensional group structure)**.**

$$\beta_{i,p} = \begin{cases} 1.6 & \text{for} \quad i \in \mathbb{G}_{1,p}^0 = \{1, \ldots, 0.4N\} \\ 0 & \text{for} \quad i \in \mathbb{G}_{2,p}^0 = \{0.4N+1, \ldots, N\} \end{cases}',$$

for $p \in \{1, 2, 3\}$.

DGP3 (Multi-dimensional group structure)**.**

$$\beta_{i,p} = \begin{cases} 1.6 & \text{for} \quad i \in \mathbb{G}_{1,p}^0 = \{1, \ldots, \tau_p N\} \\ 0 & \text{for} \quad i \in \mathbb{G}_{2,p}^0 = \{\tau_p N+1, \ldots, N\} \end{cases}',$$

for $p \in \{1, 2, 3\}$, where $\tau_1 = 0.25$, $\tau_2 = 0.5$ and $\tau_3 = 0.75$.

In DGP1, the size of the slopes, the number of groups and the group members all differ across p. This DGP is therefore multi-dimensional. DGP2 is of the usual one-dimensional type with $G_1^0 = G_2^0 = G_3^0 = 2$ groups. This DGP is included to assess the relative performance of LASSO–MD when the conditions of PLS are met. Like DGP1, DGP3 is multi-dimensional but here the difference when compared to DGP2 is more subtle. In fact, the only difference is that in DGP3 the groups are not the same for all p. The main reason for including this DGP is to illustrate how one-dimensional approaches like PLS may lead to the estimation of a large number of groups. In DGP3 there are just two multi-dimensional groups for each p. From a one-dimensional perspective, however, there are four groups within which slopes are constant. The first group has slope coefficient $\boldsymbol{\beta}_i = [1.6, 1.6, 1.6]'$ and covers the units $i \in \{1, ..., 0.25N\}$, while the second, third and fourth groups have $\boldsymbol{\beta}_i = [0, 1.6, 1.6]'$ for $i \in \{0.25N + 1, ..., 0.5N\}$, $\boldsymbol{\beta}_i = [0, 0, 1.6]'$ for $i \in \{0.5N + 1, ..., 0.75N\}$ and $\boldsymbol{\beta}_i = [0, 0, 0]'$ for $i \in \{0.75N + 1, ..., N\}$, respectively. Accounting for the two multi-dimensional groups in DGP3 therefore requires four one-dimensional groups.

We consider all combinations of $N \in \{20, 30, 40\}$ and $T \in \{60, 75, 100\}$, where T is intentionally larger than N to reflect our large-T and fixed-N asymptotic framework. The number of replications is set to 1000.[12]

[12] With this many replications the simulations were too time consuming for a personal computer. We therefore used the UPPMAX (Uppsala Multidisciplinary Center for Advanced Computational Science) cluster Rackham, which is accessible via the SNIC (Swedish National Infrastructure for Computing). Rackham consists of 486 nodes, each containing two 10-core Intel Xeon V4 central processing units. All coding was done in Python.

The three-step estimation procedure laid out in Section 2 is conditional on a particular choice of λ. While any choice satisfying Assumption 2 will in principle do, in practice it may be preferable to set λ in a data-driven fashion. The previous literature has employed information criteria to handle this issue and so do we. Specifically, the following information criterion will be used:

$$\mathrm{IC}(\lambda) := \frac{1}{T-1}\sum_{i=1}^{N}\|\widehat{\mathbf{u}}_i(\lambda)\|^2 + \varphi \cdot \sum_{p=1}^{P}\widehat{G}_p(\lambda), \tag{4.2}$$

where $\varphi = 0.5\log(T)T^{-1/2}$ and $\widehat{\mathbf{u}}_i(\lambda)$ is again the residual obtained from (2.8).[13] We write $\widehat{G}_p(\lambda)$ and $\widehat{\mathbf{u}}_i(\lambda)$ as explicit functions of λ to stress that they have been computed for a particular value of this parameter.[14] We follow the convention in the literature of setting the power parameter in the LASSO weight to $\kappa = 2$ (see, for example, Mehrabani, 2023).[15]

We evaluate the accuracy of both the estimated group structure ($\widehat{\mathbb{G}}$) and the resulting LASSO–MD coefficient estimates ($\widehat{\boldsymbol{\theta}}$). In particular, while in Table 1 we report the average and the incorrect selection frequency of the estimated number of groups ($\widehat{G}_1$, $\widehat{G}_2$ and $\widehat{G}_3$), in Table 2 we report the root mean squared error (RMSE) and the bias of the estimated coefficients. Because PLS is one-dimensional, it can only estimate one number of groups. We have already explained how the appropriate target to consider for this approach in DGP3 is given by 4, as this is the minimum number of one-dimensional groups needed to account for the heterogeneity of $\boldsymbol{\beta}_i$. The corresponding PLS targets in DGP1 and DGP2 are given by 3 and 2, respectively. We also computed the frequency with which the procedure incorrectly allocates cross-sectional units to groups conditional on estimating the correct number of groups. However, this frequency was always zero across all replications and so we do not report the results.

INSERT TABLES 1 AND 2 ABOUT HERE

[13] As usual, φ is not unique, but it is enough if it satisfies $\varphi \to 0$ and $T\varphi \to \infty$. The proposed specification not only satisfies these conditions but has been found work well across a broad range of simulation designs.

[14] The minimization of $\mathrm{IC}(\lambda)$ is carried out by first calibrating λ so that all cross-sectional units are assigned to one group for all slopes. Let $\lambda_{\max}$ denote this value. We then minimize $\mathrm{IC}(\lambda)$ by searching over 50 evenly distributed log grids in the interval $[0, \lambda_{\max}]$.

[15] The choice of κ is not particularly important as the selection of λ adapts.

We begin by considering the results reported in Table 1 for LASSO–MD. The first thing to note is that the accuracy of the estimated number of groups is increasing in T, which is a reflection of Corollary 1 and the consistency of $\widehat{G}_p$. A large T is not necessary, though, but performance is actually quite good already when $T = 60$, which is noteworthy given the complexity of the estimation problem. Accuracy increases also as the number of groups decreases, which is reasonable since more groups makes for a more challenging estimation problem.

As already explained, PLS is estimating three groups in DGP1, two in DGP2 and four in DGP3. Interestingly enough, the accuracy of LASSO–MD is almost uniformly better than that of PLS even though the estimation problem solved by LASSO–MD is relatively more complex. This is true not only in DGP1 and DGP3 but also in DGP2. Hence, LASSO–MD does not lose out to the competition even in the one-dimensional case, which makes it very attractive in empirical work where the dimensionality of any grouping is rarely known.

Moving on to Table 2, we see that the LASSO–MD coefficient estimator generally works well, which is just as expected given the accuracy of the estimated number of groups. We also see that performance is better the larger is T, which corroborates Theorem 3 and the $\sqrt{T}$-consistency of $\widehat{\boldsymbol{\theta}}$. Interestingly, the LASSO–MD RMSE is decreasing also in N. Of course, being based on the unit-specific LASSO estimator, the accuracy of the estimated groups is not expected to get any better as N increases, and this is confirmed by the results in Table 1. Our fixed-N theory therefore provides a good characterization of the types of samples in which the proposed procedure is likely to perform well as a whole.

The PLS coefficient estimator also works well but not quite as well as LASSO–MD. As already pointed out, the reason for why PLS works even in DGP1 and DGP3 when the group structure is not one-dimensional is that it is always possible to rewrite multi-dimensional structures in terms of a number of one-dimensional groups. The PLS coefficient estimator is therefore robust in this sense. It is important to note, however, that unless the true group structure is one-dimensional, the PLS group estimator does not take into account the information contained in that structure, which is wasteful. It is also likely to lead to results that

are difficult to interpret. In particular, not accounting for the multi-dimensionality of the problem is likely to lead to the estimation of many groups, as we illustrate in Section 5.

5 Empirical application

Although asset pricing models are supposed to work for individual firms as well as portfolios, they are often estimated and tested using portfolios only. In fact, the use of portfolios is so widespread that it is hardly ever questioned (see, for example, Berk, 2000). This is certainly the true for the Fama–French three-factor model, which is almost always fitted to the 25 size and value portfolios initially prescribed by Fama and French (1993). However, there is recent evidence to suggest that this practice can be problematic.

Ang et al. (2010) point out that grouping firms into a few large portfolios destroys the cross-sectional variation of the slope coefficients, which is wasteful. They instead recommend using either a larger number of small portfolios or individual stocks. Berk (2000) argues that portfolio grouping is, to a large extent, arbitrary and that the results can be very sensitive to this choice. Data snooping is therefore a concern, as researchers can alter their results by changing how the portfolio grouping is carried out. However, even in the absence of such data snooping, Berk shows that estimation results can be biased if the grouping is based on either a variable that is correlated with returns or a variable measured within the sample. Size and value are not only known to be highly correlated with returns, but they are also measured within the sample. Lewellen et al. (2010) are also skeptical of using size- and value-based portfolios. In order to avoid situations in which the data mechanically validates the model, they recommend using portfolios that do not correlate as strongly with the SMB and HML factors, such as industry portfolios.

In the present paper, we take the aforementioned concerns and recommendations seriously. The idea is to provide an agnostic assessment of the Fama–French three-factor model using LASSO–MD. We begin by estimating the model based on industry-level data. If the size and value effects are at work, we would expect the estimated coefficients, or betas, of the MKT, SMB, and HML factors to be related to the size and value of the included industries.

Take SMB as an example. Logic based on the existing empirical evidence using size and value portfolios suggests that industries with similarly sized firms should group together and that industries with relatively small firms should have relatively large betas (see Fama and French, 1993). Moreover, the betas of SMB should be unrelated to value. For HML, we expect high-value industries to have relatively larger betas.

The monthly data set we use is taken from the Kenneth French's website.[16] We include the largest sample period possible with no missing values; this means that the sample starts in July 1969 and that it ends in July 2024, for a total of $T = 661$ monthly observations. The included regressors are MKT (MKT_t), SMB (SMB_t) and HML (HML_t), which are all cross-sectionally invariant and therefore vary only by month. Hence, in terms of the model in (2.1), in this application, $\mathbf{x}_{i,t} = [MKT_t,\ SMB_t,\ HML_t]'$. The dependent variable, $y_{i,t}$, is the (average value-weighted) excess stock returns for $N = 49$ industries (similarly to Fama and French, 1997, and Sun et al., 2024). The panel data set that we consider in this section is therefore of the long type, which means that our theoretical results should provide a good approximation to the small-sample properties of LASSO–MD. We therefore expect it to perform well.

INSERT TABLE 3 ABOUT HERE

The industry return file contains size and value per industry. We average these over the sample period and rank industries descendingly so that the industries with the largest (smallest) size and value are assigned a rank of 1 (49). Table 3 reports the average and standard deviation of these size and value ranks for each of the estimated groups alongside the estimated beta and the number of members of each group when applying LASSO–MD to the demeaned data.

The first thing to note is that there is not just one group but several, and that the number of groups vary from one regressor to another, which we take as evidence in support of a multi-dimensional group structure. Another observation is that the estimated groups are not equal in size but that there are a few groups that are large and many that are very

[16] The link follows: https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

small.[17] This means that it is important to use an approach like LASSO–MD that do not require groups to be large. It also means that it may be problematic to arrange data in equal-sized groups based on size and value percentiles, and to assume that slopes are equal within groups, as is customary in the previous empirical literature. For SMB and HML, the general pattern is that for betas that are at the end of the estimated range the associated groups are small and that the large groups are located closer to the middle. For MKT, on the other hand, the largest group is at the top end of the estimated beta range.

The estimated betas of SMB and HML seem related to size and value. In particular, the betas are generally increasing (decreasing) in the average size (value) rank, which means that they are decreasing (increasing) in size (value) itself. The fact that the SMB (HML) betas are decreasing (increasing) in size (value) is, as already explained, in accordance with our a priory expectations. However, we also see that the SMB (HML) betas are not flat in value (size) but that they are decreasing (increasing), too, although the effect is not as pronounced as for size (value). Hence, the SMB (HML) betas are related not only to size (value) but also to value (size), which goes against the idea that SMB and HML should capture distinct sources of risk.

There are essentially three large groups, one for each regressor. For MKT, in terms of the labelling of Table 3, it is Group 4 (27 industries) that is largest, while for HML it is Group 5 (40 industries). For SMB, Groups 4 (9 industries) and 5 (19 industries) are estimated to be different but the difference in beta, 0.235 versus 0.237, is economically unimportant. We therefore treat these groups as one. For the large groups, the average size and value ranks are about the same, around 25, and the standard deviations are about half of those averages, which implies that a 95% confidence interval for the mean will cover almost the entire $[1, 49]$ range of ranks. Most industries therefore have the same betas regardless of their size and value. In other words, while the SMB and HML betas seem related to size and value, this effect is driven by only a small number of industries.

The large-group betas of SMB and HML are about the same and fall in the $[0.227, 0.237]$

[17] In order to avoid cluttering, Table 3 does not report the members of each group. These results are available upon request.

range. The MKT large-group beta is given by 1.095. Fama and French (1997) estimate industry-specific regressions using a sample covering the months June 1963–December 1994. They report an average beta of 1.04 for MKT, 0.39 for SMB and 0.02 for HML. The corresponding averages for our more recent sample are given by 1.01, 0.05 and 0.01, respectively. Hence, while close to the large-group beta for MKT, the average industry betas for SMB and HML are markedly smaller than our estimated large-group betas. The reason is that there are many small groups whose betas are smaller than the large-group ones, and this pulls the average down. In fact, taken at face value, these average estimates suggest that there are no size and value effects. This illustrates the risk involved when pooling across units that are not really poolable.[18]

As already pointed out, the results reported in Table 3 are based on demeaning the data with respect to industry-specific fixed effects. If we instead include these effects as additional regressors as discussed in Section 2, the estimated groups for the betas are unaffected. For the alphas we estimate only one group and the common alpha estimate, -0.039, is very close to zero. To test whether this estimate is significantly different from zero, we computed a t-test along the lines explained in Section 3. The associated p-value is 0.125, which does not provide any evidence against the Fama–French three-factor model.

INSERT TABLE 4 ABOUT HERE

According to our Monte Carlo results, LASSO–MD works relatively well not only if the group structure is multi-dimensional, which it seems to be in the present application, but also if it is one-dimensional. Even so, for completeness, Table 4 reports the results obtained by using Mehrabani's (2023) PLS approach. As expected given the discussion in Section 4 (of the results for DGP3), PLS estimates many groups, 25 in total. Moreover, because there are multiple regressors whose estimated slope coefficients all vary to some extent between groups, it seems very difficult, if not impossible, to determine if the difference from one group to another is due to one or more coefficients changing. We also see that the procedure

[18] Averaging across unit-specific slope estimates is a very common form of pooling, often referred to as "between" pooling.

estimates one group, Group 16, that is substantially larger than the rest and whose betas are very close to the large-group LASSO–MD ones. However, the size of this group (21 industries) is much smaller than the largest groups estimated by LASSO–MD, which is again just as expected given the discussion in Section 4. These results illustrate quite clearly the loss of information incurred by not accounting for the multi-dimensionality of the grouping problem.

6 Conclusion

This study proposes a new approach to the estimation of panel data regression models with interactive effects and possibly heterogenous slope coefficients. The heterogeneity is modelled through a multi-dimensional group structure in which each individual slope coefficient may have its own set of groups, which means that it is very flexible. The estimation problem is viewed as model selection issue that is addressed using a novel version of the LASSO dubbed "LASSO–MD". The new approach is computationally fast, it does not require large groups, it is asymptotically valid provided only that T is large, and it has excellent small-sample properties. It should therefore be a valuable addition to the already existing menu of estimators of panel data regression models with grouped heterogeneity.

7 References

Ang, A., J. Liu and K. Schwarz (2020). Using stocks or portfolios in tests of factor models. *Journal of Financial and Quantitative Analysis* **55**, 709–750.

Athey, S. and G. Imbens (2015). Machine learning methods in economics and econometrics: A measure of robustness to misspecification. *American Economic Review: Papers & Proceedings* **105**, 476–480.

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

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

Berk, J. B. (2000). Sorting out sorts. *Journal of Finance* **55**, 407–427.

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

Cheng, X., F. Schorfheide and P. Shao (2023). Clustering for multi-dimensional heterogeneity with an application to production function estimation. PIER Working Paper 23–016.

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

Chordia, T., A. Goyal and J. A. Shanken (2017). Cross-sectional asset pricing with individual stocks: Betas versus characteristics. Available at SSRN: `https://ssrn.com/abstract=2549578 or http://dx.doi.org/10.2139/ssrn.2549578`.

Cytrynbaum, M. (2021). Blocked clusterwise regression. ArXiv: 2001.11130.

Eberhardt, M., and F. Teal (2013). No Mangoes in the Tundra: Spatial Heterogeneity in Agricultural Productivity Analysis. *Oxford Bulletin of Economics and Statistics* **75**, 914–939.

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

Fama, E. F., and K. R. French (1997). Industry costs of equity. *Journal of Financial Economics* **43**, 153–193.

Hansen, B. E. (1999). Threshold effects in non-dynamic panels: Estimation, testing, and inference. *Journal of Econometrics* **93**, 345–368.

Judson, R. A., and A. L. Owen (1999). Estimating dynamic panel data models: a guide for macroeconomists. *Economics Letters* **65**, 9–15.

Leng, X., H. Chen and W. Wang (2023). Multi-dimensional latent group structures with heterogeneous distributions. *Journal of Econometrics* **233**, 1–21.

Lewellen J., S. Nagel and J. Shanken (2010). A skeptical appraisal of asset pricing tests. *Journal of Financial Economics* **96**, 175–194.

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

Lu, X., and H. White (2014). Robustness checks and robustness tests in applied economics. *Journal of Econometrics* **178**, 194–206.

Mehrabani, A. (2023). Estimation and identification of latent group structures in panel data. *Journal of Econometrics* **235**, 1464–1482.

Miao, K., L. Su and W. Wang (2020). Panel threshold regressions with latent group structures. *Journal of Econometrics* **214**, 451–481.

Newet, W. K., and K. D. West (1987). A simple positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55**, 703–708.

Pesaran, H. M., and T. Yamagata (2024). Testing for alpha in linear factor pricing models with a large number of securities. *Journal of Financial Econometrics* **22**, 407–460.

Sarafidis, V., and N. Weber (2015). A partially heterogeneous framework for analyzing panel data. *Oxford Bulletin of Economics and Statistics* **77**, 274–296.

Su, L., and G. Ju (2018). Identifying latent grouped patterns in panel data models with interactive fixed effects. *Journal of Econometrics* **206**, 554–573.

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

Sun, Y., C. Wan, W. Zhang and W. Zhong (2024). A multi-kink quantile regression model with common structure for panel data analysis. *Journal of Econometrics* **239**, 105304.

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

Table 1: Monte Carlo results for the estimated number of groups.

		DGP1		DGP2		DGP3	
N	T	LMD	PLS	LMD	PLS	LMD	PLS
				Incorrect estimation frequency			
20	60	[0.286, 0.188, 0.008]	0.376	[0.024, 0.029, 0.025]	0.034	[0.040, 0.024, 0.045]	0.214
20	75	[0.215, 0.153, 0.002]	0.298	[0.014, 0.010, 0.015]	0.016	[0.019, 0.009, 0.021]	0.121
20	100	[0.102, 0.101, 0.002]	0.155	[0.003, 0.006, 0.006]	0.005	[0.006, 0.006, 0.009]	0.048
30	60	[0.391, 0.285, 0.017]	0.516	[0.055, 0.052, 0.070]	0.054	[0.074, 0.035, 0.076]	0.321
30	75	[0.279, 0.234, 0.007]	0.362	[0.033, 0.025, 0.027]	0.029	[0.047, 0.017, 0.031]	0.210
30	100	[0.151, 0.132, 0.001]	0.199	[0.009, 0.009, 0.010]	0.009	[0.016, 0.008, 0.012]	0.083
40	60	[0.696, 0.647, 0.012]	0.770	[0.051, 0.046, 0.065]	0.474	[0.048, 0.033, 0.052]	0.673
40	75	[0.626, 0.560, 0.001]	0.681	[0.020, 0.021, 0.011]	0.021	[0.030, 0.022, 0.020]	0.132
40	100	[0.119, 0.098, 0.001]	0.152	[0.007, 0.006, 0.003]	0.006	[0.015, 0.004, 0.009]	0.054
				Average			
20	60	[3.337, 2.189, 1.009]	3.409	[2.024, 2.031, 2.027]	2.038	[2.042, 2.025, 2.047]	4.266
20	75	[3.235, 2.154, 1.002]	3.317	[2.014, 2.010, 2.015]	2.018	[2.020, 2.009, 2.022]	4.137
20	100	[3.106, 2.101, 1.002]	3.162	[2.003, 2.006, 2.006]	2.005	[2.006, 2.006, 2.009]	4.049
30	60	[3.493, 2.313, 1.017]	3.751	[2.056, 2.055, 2.075]	2.057	[2.078, 2.036, 2.082]	4.448
30	75	[3.330, 2.250, 1.007]	3.474	[2.034, 2.026, 2.028]	2.029	[2.048, 2.017, 2.032]	4.252
30	100	[3.162, 2.136, 1.001]	3.235	[2.009, 2.010, 2.010]	2.010	[2.017, 2.010, 2.012]	4.095
40	60	[3.783, 2.685, 1.012]	4.032	[2.054, 2.047, 2.071]	2.481	[2.052, 2.035, 2.055]	5.639
40	75	[3.668, 2.575, 1.002]	3.781	[2.020, 2.021, 2.011]	2.023	[2.031, 2.022, 2.020]	4.177
40	100	[3.138, 2.104, 1.002]	3.185	[2.007, 2.006, 2.003]	2.006	[2.015, 2.004, 2.009]	4.061

Notes: "LMD" and "PLS" refer to LASSO–MD and the penalized least squares (PLS) approach of Mehrabani (2023), respectively. In DGP1 and DGP3, the group structure is multi-dimensional, while in DGP2 it is one-dimensional. The top panel reports the frequency of incorrectly estimating the number of groups across replications. For LASSO–MD, the results have the following form "[x, y, z]", where x, y and z are the frequency of incorrectly estimating G_1^0, G_2^0 and G_3^0, respectively. For PLS, we report the frequency of not estimating $\max\{G_1^0, G_2^0, G_3^0\}$. The bottom panel reports the average of the estimated number of groups across replications. For LASSO–MD, the results have the same form as in the top panel but now x, y and z are the average $\widehat{G}_1$, $\widehat{G}_2$ and $\widehat{G}_3$, respectively.

Table 2: Monte Carlo results for the estimated coefficients.

		DGP1		DGP2		DGP3	
N	T	LMD	PLS	LMD	PLS	LMD	PLS
				RMSE			
20	60	0.1249	0.1641	0.0492	0.0500	0.0943	0.0451
20	75	0.0691	0.1128	0.0317	0.0336	0.0751	0.0378
20	100	0.0613	0.0846	0.0329	0.0348	0.0583	0.0311
30	60	0.0418	0.0590	0.0225	0.0279	0.0219	0.0540
30	75	0.0637	0.0707	0.0397	0.0323	0.0671	0.0358
30	100	0.0325	0.0477	0.0193	0.0199	0.0247	0.0118
40	60	0.0497	0.0956	0.0202	0.0224	0.0916	0.0293
40	75	0.0300	0.0703	0.0200	0.0133	0.0316	0.0179
40	100	0.0310	0.0315	0.0200	0.0201	0.0335	0.0217
				Bias			
20	60	0.0050	0.0168	-0.0072	-0.0081	-0.0135	-0.0077
20	75	0.0012	0.0185	0.0052	0.0049	0.0099	0.0022
20	100	0.0026	-0.0037	-0.0039	-0.0044	-0.0067	-0.0002
30	60	0.0053	0.0055	-0.0008	-0.0012	0.0002	0.0030
30	75	0.0036	0.0034	0.0008	0.0013	0.0023	0.0046
30	100	-0.0008	-0.0050	-0.0006	-0.0010	0.0012	0.0000
40	60	0.0004	0.0022	-0.0021	-0.0022	-0.0085	-0.0017
40	75	0.0004	0.0030	0.0013	0.0012	0.0043	0.0010
40	100	0.0038	0.0031	0.0019	0.0019	0.0029	0.0021

Notes: "LMD" and "PLS" refer to LASSO–MD and the penalized least squares (PLS) approach of Mehrabani (2023), respectively. In DGP1 and DGP3, the group structure is multi-dimensional, while in DGP2 it is one-dimensional. The numbers reported in the table are the average bias and root mean square error (RMSE) across replications and coefficients times 100. All results are conditional on the groups being correctly estimated.

Table 3: Empirical results based on LASSO–MD.

		Size		Value		
Group	Beta	Mean	STD	Mean	STD	GS
			MKT			
1	0.667	21.5	14.8	18.5	21.9	2
2	0.855	8.5	10.6	36.0	7.1	2
3	0.896	14.8	12.2	34.9	16.5	8
4	1.095	28.3	13.6	22.6	13.0	37
			SMB			
1	-0.037	11.0	14.1	45.0	5.7	2
2	-0.011	13.6	8.6	23.2	16.3	12
3	0.195	20.5	17.7	42.0	7.1	2
4	0.235	19.1	11.5	23.1	15.1	9
5	0.237	33.1	10.3	24.8	11.2	19
6	0.315	42.0	–	22.0	–	1
7	0.423	46.0	–	2.0	–	1
8	0.447	49.0	–	14.0	–	1
9	0.654	31.0	24.0	28.0	24.0	2
			HML			
1	-0.511	14.0	–	45.0	–	1
2	-0.272	9.5	3.5	39.5	5.0	2
3	-0.218	23.0	–	40.0	–	1
4	-0.089	27.0	8.5	48.0	1.41	2
5	0.227	24.8	14.4	23.7	12.92	40
6	0.453	38.0	14.1	8.5	3.54	2
7	0.488	46.0	–	2.0	–	1

Notes: This table reports the estimation results based on the proposed LASSO–MD approach. The dependent variable is excess stock returns. The regressors are given by MKT, SMB and HML. "Beta" refers to the estimated slope coefficient, "Size" and "Value" refer to the size and value rank for each group, respectively, "Mean" and "STD" refer to the sample mean and standard deviation, respectively, and "GS" refers to the size of each estimated group. For groups that contain only one industry, the STD cannot be computed and is hence missing.

Table 4: Empirical results based on PLS.

	Beta			
Group	MKT	SMB	HML	GS
1	0.665	0.217	0.135	1
2	0.726	-0.069	0.257	1
3	0.820	-0.150	-0.099	1
4	0.831	-0.036	0.209	1
5	0.843	-0.026	0.200	5
6	0.933	0.092	0.010	1
7	0.935	0.108	0.278	1
8	0.968	0.066	0.291	1
9	0.983	0.072	0.275	1
10	0.988	0.093	0.212	1
11	1.004	0.104	0.252	1
12	1.011	0.223	0.205	1
13	1.029	0.123	0.249	1
14	1.038	0.170	0.170	1
15	1.093	0.309	0.221	1
16	1.095	0.229	0.213	21
17	1.097	0.038	0.427	1
18	1.115	0.241	0.219	1
19	1.121	0.219	-0.212	1
20	1.129	0.244	0.194	1
21	1.134	0.530	0.576	1
22	1.162	0.341	-0.183	1
23	1.165	0.633	0.548	1
24	1.227	0.330	-0.230	1
25	1.298	0.718	-0.509	1

Notes: This table reports the estimation results based on the penalized least squares (PLS) approach of Mehrabani (2023). The dependent variable is excess stock returns. The regressors are given by MKT, SMB and HML. "Beta" and "GS" refer to the estimated slope coefficient and the size of each estimated group, respectively.