Multidimensional Interactive Fixed-Effects *

Hugo Freeman

September 22, 2022

Latest version here

Abstract

This paper studies a linear and additively separable model for multidimensional panel data of three or more dimensions with unobserved interactive fixed effects. Two approaches are considered to account for these unobserved interactive fixed-effects when estimating coefficients on the observed covariates. First, the model is embedded within the standard two-dimensional panel framework and restrictions are derived under which the factor structure methods in Bai (2009) lead to consistent estimation of model parameters. The second approach considers group fixed-effects and kernel methods that are more robust to the multidimensional nature of the problem. Theoretical results and simulations show the benefit of standard two-dimensional panel methods when the structure of the interactive fixed-effect term is known, but also highlight how the group fixed-effects and kernel methods perform well without knowledge of this structure. The methods are implemented to estimate the demand elasticity for beer under a handful of models for demand.

1 Introduction

Models of multidimensional data – panel data with more than two dimensions – are fast becoming popular in econometric analysis as large data sets with a multidimensional structure become available. For example, in gravity models of trade that are repeated over time one may be interested in studying trade patterns between an importer, i, an exporter, j that is repeated every quarter or year, t. One may also be interested in studying demand elasticities through consumption data that may vary by product, i, store, j with repeated observation over week

*First and foremost I would like to greatly thank my advisor, Martin Weidner, who has been exceptionally generous with his time and patience with me during my PhD and especially on this paper. I would also like to thank my secondary advisor Lars Nesheim for his guidance through my various PhD milestones and projects. For this paper specifically, additional to the above I would like to thank Andrei Zeleneev for his exceedingly helpful suggestions along with the audience at The International Panel Data Conference 2022, The IAAE Conference 2022 and the Bristol Econometrics Study Group 2022. This research was supported by the European Research Council grant ERC-2018-CoG-819086-PANEDA.

or month, t.[1] In these examples it is clear that there may exist unobserved characteristics in each dimension that can determine variation across both the dependent and independent variables that needs to be controlled for to avoid issues with endogeneity. For example, this could be shifts in taste preferences, that are unobserved by the econometrician, that may effect sales of particular products in certain stores differently over time. Thus far most analysis has addressed unobserved heterogeneity in the higher-dimensional setting by using a combination of additive scalar fixed-effects. These additive scalar fixed-effects approaches, however, can only accommodate variation in unobserved heterogeneity over a subset of dimensions with any one of the scalar fixed-effects terms. For example, in the three-dimensional model this type of fixed-effect approach can only control for variation over ij, it and jt, but not over all ijt. In the face of more complicated relationships that admit multiplicative variation across dimensions, these additive effects are unsatisfactory to control for unobserved heterogeneity. This paper develops tools to control for unobserved heterogeneity in the form of interactive fixed-effects in models of multidimensional panel data. The main body of the paper focuses on the linear and additively separable model. More generic applications of these tools are discussed in the introduction but are not formally studied.

To fix ideas consider linear parameter estimation in the following interactive fixed-effects model with three dimensions,

$$Y_{ijt} = X_{ijt}'\beta + \sum_{\ell=1}^{L} \varphi_{i\ell}^{(1)} \varphi_{j\ell}^{(2)} \varphi_{t\ell}^{(3)} + \varepsilon_{ijt}, \tag{1}$$

where all terms in $\sum_{\ell=1}^{L} \varphi_{i\ell}^{(1)} \varphi_{j\ell}^{(2)} \varphi_{t\ell}^{(3)}$ are unobserved and L, which is defined later, can be interpreted as a rank parameter. Reducing the problem to three dimensions is without loss of generality for the methods considered herein. Additive fixed-effects are omitted for brevity but can be removed with a simple within transformation. Let X_{ijt} be arbitrarily correlated with the unobserved interactive fixed-effects term, $\sum_{\ell=1}^{L} \varphi_{i\ell}^{(1)} \varphi_{j\ell}^{(2)} \varphi_{t\ell}^{(3)}$, but uncorrelated with the noise term, ε_{ijt}. The challenge to estimating β is isolating variation in X_{ijt} that is not correlated with the interactive fixed-effects term. This paper develops the multidimensional group fixed-effects and kernel weighted transformations to project out this unobserved heterogeneity and also shows settings were standard factor methods work well. The group fixed-effects method used below is similar to Bonhomme, Lamadon and Manresa (2021) and the within-cluster transformation in Freeman and Weidner (2022).

The model for interactive fixed-effects has precedent in the standard two-dimensional panel

[1] A non-exhaustive list of related examples can be found in the introduction of Matyas (2017) in trade, housing and prices, migration, country productivity and consumer price setting.

data setting. For instance take the model considered in Bai (2009) and similar to Pesaran (2006),

$$Y_{it} = X_{it}'\beta + \sum_{\ell=1}^{L} \lambda_{i\ell} f_{t\ell} + e_{it}. \tag{2}$$

In that setting, Bai (2009) show that the interactive term $\sum_{\ell=1}^{L} \lambda_{i\ell} f_{t\ell}$ also sufficiently captures variation in additive individual and time effects without the need to specify these separately, so these are again naturally omitted. For multidimensional applications it may be preferable to simply transform the problem in (1) to a two dimensional problem and estimate (2) directly using the transformed data. However, and as will be explained in further detail in Section 3.1, problems persist when L is large and only a subset of the unobserved heterogeneity parameters are low-dimensional. For consistent estimation of β, transforming the multidimensional array to a matrix then estimating (2) requires either: (a) all fixed-effects are low-dimensional, or; (b) that a subset of the fixed-effects are low-dimensional and the analyst knows which ones are. The requirement that the analyst has this knowledge can be highly restrictive. Alternatively, the within-cluster and kernel weighted transformations analysed in this paper requires only that a subset of the fixed-effect parameters are low-dimensional, though the analyst does not need to know which of the fixed-effect parameters make up this subset.

Under sufficient regularity conditions, the methods considered in this paper may also control for variation from arbitrary functions of the fixed-effects. Similar to that considered in Zeleneev (2020) and Freeman and Weidner (2022), the functional representation of model (1) could be,

$$Y_{ijt} = X_{ijt}'\beta + h(\varphi_i^{(1)}, \varphi_j^{(2)}, \varphi_t^{(3)}) + \varepsilon_{ijt},$$

for vector-valued $\varphi_i^{(1)}$, $\varphi_j^{(2)}$ and $\varphi_t^{(3)}$. The set of fixed-effects to be transformed by the function $h(\cdot,\cdot,\cdot)$ could also extend to fixed-effects over multiple indices, e.g. α_{ij} from above. It should be noted that the setting considered in Zeleneev (2020) requires that the transformation is non-smooth, and it is not trivial to see that a "within-type" transformation will sufficiently project this type of heterogeneity. With sufficient smoothness conditions on the function transforming the fixed-effects, existing literature could be generalised to show consistency using the proposed within-cluster transformation in the multidimensional case.

Models with discrete explanatory variables (Chernozhukov, Fernández-Val, Hahn and Newey, 2013; Hoderlein and White, 2012; Evdokimov, 2010; Fernández-Val, Freeman and Weidner, 2021), provide another interesting application of these group fixed-effects estimators. Take the following regression line for discrete valued X_{ijt},

$$Y_{ijt} = h\left(X_{ijt}, \varphi_i^{(1)}, \varphi_j^{(2)}, \varphi_t^{(3)}, \varepsilon_{ijt}\right).$$

Then, under sufficient smoothness conditions on the function h, the unobserved heterogeneity may also be projected out with a group fixed-effect estimator. Tensor completion techniques also

have useful generalisations in this setting, for example, Tomioka, Hayashi and Kashima (2010); Li, Wang, Lu and Tang (2019); Xu (2020), for some examples of methods that consider sparse multidimensional arrays. The sparse multidimensional array problem has similar complexities to the low-rank tensor approximation problem in that they do not extend from the matrix problems in a straightforward way, hence require non-trivial extensions.

It is also important to consider unobserved heterogeneity in applications that admit discrete dependent variables. For example, for binary response variable with known $F(\cdot)$,

$$Y_{ijt} = F\big(X_{ijt}'\beta + \sum_{\ell=1}^{L} \varphi_{i\ell}^{(1)}\varphi_{j\ell}^{(2)}\varphi_{t\ell}^{(3)}\big). \tag{3}$$

Estimation of the unobserved heterogeneity term may then be performed with a similar iterative scheme as that proposed in Chen, Fernández-Val and Weidner (2021) or the sufficient statistic approach in Chapter 6 of Matyas (2017). The incidental parameter problem in this setting can be alleviated using methods in this paper by allowing cluster sizes to grow with data size coupled with taking grouped fixed-effects along fewer dimensions, for example in Bonhomme, Lamadon and Manresa (2021) and also Appendix C.

Menzel (2021) consider a special case of multidimensional data for bootstrapping methods where the data is D-adic. That is, each dimension of the data refers to the same set of observations, like a network graph where each index refers to an individual in the network. An example of the multidimensional version of this could be a binary indicator of a three step path, $Y_{ijk} = G_{ij}G_{jk}$, detailing if there exists a path from i to k. In any case, the type of multidimensional data considered in that work is a distinct special case of the type of data structures considered in this paper.

The technical component of this paper is highly related to the numerical analysis literature on low-rank approximations of multidimensional arrays. As pointed out in De Silva and Lim (2008), the optimisation problem of finding low-rank approximations in the tensor setting is not well-posed, hence most results in this literature rely on numerical evidence. See Kolda and Bader (2009) for a summary of the multidimensional array decomposition problem and Vannieuwenhoven, Vandebril and Meerbergen (2012); Rabanser, Shchur and Günnemann (2017) for examples of numerical results. As such, it is necessary to innovate on this tensor low-rank problem to find appropriate analytical results. To this end, this paper utilises well-posed components of the numerical analysis literature for use in nuisance parameter applications. These applications have the advantage that they do not require the multidimensional array of fixed-effects to be reconstructed, hence do not attempt to directly solve the low-rank tensor problem. It is worth a note that Elden and Savas (2011), along with related papers, suggest a reformulation of the low multilinear rank problem that may have promising applications in econometrics, but this is left for future research.

The paper is organised as follows. Section 2 introduces the model and motivates the within-cluster transformation, Section 3 discusses the existing literature on matrix methods and how these extend to the multidimensional setting, Section 4 discusses estimation with group fixed-effects and kernel methods, which is the main methodological contribution of the paper, Section 5 discusses the asymptotic performance of the estimators, Section 6 displays the simulation results and Section 8 concludes.

2 Motivation

The model in full dimensional generality is, [2]

$$\boldsymbol{Y} = \sum_{k=1}^{K} \boldsymbol{X}_k \beta_k + \boldsymbol{\mathcal{A}} + \boldsymbol{\mathcal{B}} + \boldsymbol{\varepsilon}, \tag{4}$$

where $\boldsymbol{Y}, \boldsymbol{X}_k, \boldsymbol{\varepsilon} \in \mathbb{R}^{N_1 \times N_2 \times \cdots \times N_d}$. $\boldsymbol{\mathcal{A}} = \sum_{\ell=1}^{L} \varphi_\ell^{(1)} \circ \cdots \circ \varphi_\ell^{(d)}$ where $\varphi_\ell^{(n)} \in \mathbb{R}^{N_n}$ for each $n = 1, \ldots, d$ and "$\circ$" is the outer product. L is naturally restricted to have upper bound $\min_n\{\prod_{n' \neq n} N_{n'}\}$, see Kruskal (1989). $\boldsymbol{\mathcal{B}}$ is the sum of additive fixed-effects terms that can each vary over $d-1$ indices. For example, this could be, $\mathcal{B}_{i_1,i_2,\ldots,i_d} = \gamma^{(1)}_{i_2,\ldots,i_d} + \gamma^{(2)}_{i_1,i_3,\ldots,i_d} + \gamma^{(3)}_{i_1,i_2,i_4,\ldots,i_d} + \cdots + \gamma^{(d)}_{i_1,\ldots,i_{d-1}}$. However, each of these scalar terms need not vary over $d-1$ dimensions and may vary over $d-2$ or $d-3$ dimensions, and so forth. $\boldsymbol{\varepsilon}$ is a noise term uncorrelated with all $\boldsymbol{X}_k$ and all unobserved fixed-effects terms. Take $i_n \in \{1, \ldots, N_n\}$ for all $n \in \{1, \ldots, d\}$ as the dimension specific index. The regressors $\boldsymbol{X}_k$ may be arbitrarily correlated with $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{B}}$. $\boldsymbol{\mathcal{B}}$ cannot be ignored without loss of generality for arbitrary estimators hence is included for completeness. The main estimation procedure considered below is robust to $\boldsymbol{\mathcal{B}}$ being included or excluded, so it may be ignored in that context.

Model (4) can be seen as a natural extension of the Bai (2009) model to three (or more) dimensions with the $\boldsymbol{\mathcal{A}}$ term interpreted as a "higher-dimensional" factor stucture. Similar to this strain of the literature, all terms in $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{B}}$ are considered fixed nuisance parameters. There are potentially many extensions to the factor model setting in Bai (2009) to the higher dimension case. This paper starts with what seems the most natural extension.

[2] For example, in index notation this model can be written as,

$$Y_{i_1,i_2,\ldots,i_d} = \sum_{k=1}^{K} X_{i_1,i_2,\ldots,i_d;k} \beta_k + \mathcal{A}_{i_1,i_2,\ldots,i_d} + \mathcal{B}_{i_1,i_2,\ldots,i_d} + \varepsilon_{i_1,i_2,\ldots,i_d}$$

with $\mathcal{A}_{i_1,i_2,\ldots,i_d} = \sum_{\ell=1}^{L} \varphi^{(1)}_{i_1\ell} \ldots \varphi^{(d)}_{i_d\ell}$ and the additive fixed-effects term could be for example, $\mathcal{B}_{i_1,i_2,\ldots,i_d} = \gamma^{(1)}_{i_2,\ldots,i_d} + \gamma^{(2)}_{i_1,i_3,\ldots,i_d} + \gamma^{(3)}_{i_1,i_2,i_4,\ldots,i_d} + \cdots + \gamma^{(d)}_{i_1,\ldots,i_{d-1}}$.

2.1 Within-cluster transformation

Ignore the additive fixed-effects and consider the model,

$$\mathbf{Y} = \sum_{k=1}^{K} \beta_k \mathbf{X}_k + \boldsymbol{\mathcal{A}} + \boldsymbol{\varepsilon}.$$

Allowing for arbitrary correlation between $\mathbf{X}_k$ and $\boldsymbol{\mathcal{A}}$ for some or all k introduces an endogeneity problem unless estimation methods can account for this. For example, $\mathbf{X}_k$ may be generated as,

$$\mathbf{X}_k = h_k(\boldsymbol{\mathcal{A}}) + \boldsymbol{\nu}_k.$$

For the below motivation it is useful to reconsider (4) with just three-dimensions again. A popular method to project fixed-effects is the within transformation on both $\mathbf{Y}$ and $\mathbf{X}$ as follows,

$$\widetilde{Y}_{ijt} = Y_{ijt} - \bar{Y}_{\cdot jt} - \bar{Y}_{i\cdot t} - \bar{Y}_{ij\cdot} + \bar{Y}_{\cdot\cdot t} + \bar{Y}_{\cdot j\cdot} + \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdots} \tag{5}$$

where the variables with bars simply denote the average taken over the "dotted" index for the entire sample. That is, $\bar{Y}_{\cdot jt} := \frac{1}{N_1}\sum_{i=1}^{N_1} Y_{ijt}$, $\bar{Y}_{\cdot\cdot t} := \frac{1}{N_1 N_2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} Y_{ijt}$, etc. Note, this is not a unique representation of the within transformation, and can be achieved with many different transformations. In the presence of interactive fixed-effects this transformation leaves the following fixed-effects residual,

$$\widetilde{\mathcal{A}}_{ijt} = \sum_{\ell=1}^{L} (\varphi_{i\ell}^{(1)} - \bar{\varphi}_{\ell}^{(1)})(\varphi_{j\ell}^{(2)} - \bar{\varphi}_{\ell}^{(2)})(\varphi_{t\ell}^{(3)} - \bar{\varphi}_{\ell}^{(3)}),$$

where $\bar{\varphi}_{\ell}^{(1)} := \frac{1}{N_1}\sum_{i=1}^{N_1} \varphi_{i\ell}^{(1)}$, $\bar{\varphi}_{\ell}^{(2)} := \frac{1}{N_2}\sum_{j=1}^{N_2} \varphi_{j\ell}^{(2)}$, and $\bar{\varphi}_{\ell}^{(3)} := \frac{1}{N_3}\sum_{t=1}^{N_3} \varphi_{t\ell}^{(3)}$. Sufficient projection of the unobserved heterogeneity term, $\widetilde{\mathcal{A}}_{ijt}$, then relies on either $\varphi_{i\ell}^{(1)} \to \bar{\varphi}_{\ell}^{(1)}$, $\varphi_{j\ell}^{(2)} \to \bar{\varphi}_{\ell}^{(2)}$ or $\varphi_{t\ell}^{(3)} \to \bar{\varphi}_{\ell}^{(3)}$ for each ℓ, which is highly restrictive and actually implies the simple additive fixed-effect model is the true model.

Consider as an alternate the within-cluster transformation. Take for now as given some grouping of units within each dimension. That is, there exist some allocations of all i's into groups, all j's into groups and all t's into groups. How these groups are constructed is formally defined later. With a slight abuse of notation in how each $\bar{\mathcal{A}}$ is defined, the within-cluster transformation is,

$$\widetilde{\mathcal{A}}_{ijt} = \mathcal{A}_{ijt} - \bar{\mathcal{A}}_{i^*jt} - \bar{\mathcal{A}}_{ij^*t} - \bar{\mathcal{A}}_{ijt^*} + \bar{\mathcal{A}}_{i^*j^*t} + \bar{\mathcal{A}}_{i^*jt^*} + \bar{\mathcal{A}}_{ij^*t^*} - \bar{\mathcal{A}}_{i^*j^*t^*} \tag{6}$$

where the bar variables combined with the star indices denote means taken within that index's group. For example, $\bar{\mathcal{A}}_{i^*jt}$ is the mean value of all i^*'s assigned to i's group, $\bar{\mathcal{A}}_{i^*jt^*}$ is the mean

across both i^* in i's group and t^* in t's group, and so on. Taking group assignments as a given, this projection is equivalent to solving the following least squares problem

$$Q(\beta, \mathbf{g}) = \min_{\alpha,\gamma,\delta} \sum_{i,j,t} \left(Y_{ijt} - X'_{ijt}\beta - \alpha_{g_1(i)jt} - \gamma_{ig_2(j)t} - \delta_{ijg_3(t)}\right)^2 \tag{7}$$

where $\alpha \in \mathbb{R}^{\text{ncol}(g_1)\times N_2 \times N_3}$, $\gamma \in \mathbb{R}^{N_1 \times \text{ncol}(g_2) \times N_3}$ and $\delta \in \mathbb{R}^{N_1 \times N_2 \times \text{ncol}(g_3)}$ where $\text{ncol}(\cdot)$ returns the number of columns of a matrix. $\mathbf{g}$ is a list of group assignment for each dimension. For example, $g_1(i)$ maps to the group identity of individual i. This is why α is restricted to vary across only $\text{ncol}(g_1)$ different values in the first dimension, which is less than N_1. Notice α, γ and δ can be defined as combinations of the within-cluster projection as follows,

$$\begin{aligned}
\alpha_{g_1(i)jt} &:= \bar{\mathcal{A}}_{i^*jt} - \bar{\mathcal{A}}_{i^*j^*t} + \bar{\mathcal{A}}_{i^*j^*t^*} \\
\gamma_{ig_2(j)t} &:= \bar{\mathcal{A}}_{ij^*t} - \bar{\mathcal{A}}_{i^*j^*t} \\
\delta_{ijg_3(t)} &:= \bar{\mathcal{A}}_{ijt^*} - \bar{\mathcal{A}}_{i^*jt^*},
\end{aligned}$$

though this representation is not unique.

This projection also differences out any additive fixed effects and leaves the following interactive fixed-effects residual,

$$\widetilde{\mathcal{A}}_{ijt} = \sum_{\ell=1}^{L} (\varphi_{i\ell}^{(1)} - \bar{\varphi}_{i^*\ell}^{(1)})(\varphi_{j\ell}^{(2)} - \bar{\varphi}_{j^*\ell}^{(2)})(\varphi_{t\ell}^{(3)} - \bar{\varphi}_{t^*\ell}^{(3)}). \tag{8}$$

where $\bar{\varphi}_{i^*\ell}^{(1)}$ is the group mean of $\varphi_{i^*\ell}^{(1)}$ for the i^*'s in i's group, and so on for the other terms. Hence, sufficient projection of the interactive fixed-effects terms relies on the weaker condition that parameters converge to their group means, namely, $\varphi_{i\ell}^{(1)} \to \bar{\varphi}_{i^*\ell}^{(1)}$, $\varphi_{j\ell}^{(2)} \to \bar{\varphi}_{j^*\ell}^{(2)}$ or $\varphi_{t\ell}^{(3)} \to \bar{\varphi}_{t^*\ell}^{(3)}$ for each ℓ. This condition is clearly weaker than requiring the parameters to converge to the population means. In fact, the group mean differencing could be seen as a weighted mean difference across the population, with equal weight given to observations within the cluster and zero weight to observations outside of each cluster. This fact will be utilised to also develop a more generic kernel type weighted difference estimator, synonymous to a Nadaraya-Watson type estimator.

So far it has been shown that with the relatively innocuous shift from the within transformation to the within-cluster transformation, projection of the additive heterogeneity term has been preserved and less restrictive conditions are required to also eliminate the interactive term. Choice of clusters for this transformation is key to suffice this less restrictive condition. Given a set of proxies to cluster on, clustering or matching methods can be used to find these groups, for example Bonhomme, Lamadon and Manresa (2021). Developing a set of proxies and choice of cluster approach is discussed in Section 4. A sufficient condition for consistency can be

concisely summed up by Assumption 1, which considers the model in full generality as in (4). In Assumption 1, the set $\mathcal{M}$ is defined to be referenced at a later stage in the statement of the consistency result but can be ignored at this stage. Importantly this set must be non-empty. Reference to the cluster mechanism $\mathcal{C}$ is also made in the assumption as these may change and impact how reasonable each bound in Assumption 1 is. Generically, these cluster mechanisms are simply any way to group units together. Again, at this stage reference to $\mathcal{C}$ can be ignored, it is just useful for later reference.

Assumption 1 (Clustering)**.**

Let $j_n(i_n)$ be any unit in the same cluster as i_n from using cluster mechanism $\mathcal{C}$. Then,

(i). For all n as $N_n \to \infty$,

$$\frac{1}{N_n}\sum_{i_n=1}^{N_n}\left\|\varphi_{i_n}^{(n)}-\varphi_{j_n(i_n)}^{(n)}\right\|^2 \lesssim O_p(1)$$

and,

(ii). For a non-empty subset $\mathcal{M} \subset \{1,\ldots,d\}$ take for any $n^* \in \mathcal{M}$ a sequence $\xi_{N_{n^*}} \to 0$ as $N_{n^*} \to \infty$. Then,

$$\frac{1}{N_{n^*}}\sum_{i_n=1}^{N_{n^*}}\left\|\varphi_{i_{n^*}}^{(n^*)}-\varphi_{j_{n^*}(i_{n^*})}^{(n^*)}\right\|^2 = O_p(\xi_{N_{n^*}})$$

Assumption 1.(i) restricts the chosen clustering procedure to allocate stable clusters. This implies that as $\{N_1,\ldots,N_d\} \to \infty$ the cluster allocations do not become increasingly disparate in the underlying parameter space. Assumption 1.(ii) states that for at least one dimension the clustering procedure finds matches with asymptotically negligible difference in the underlying parameter space. These assumptions restrict the cluster mechanism to uncover closeness in the true parameter space, which implies a restriction on how the cluster proxies are sampled and on how these proxies are used to cluster. Since proxies are not always estimated, and actually sometimes group assignments may be given extraneously, it is useful to state Assumption 1 in generic terms that ignore these clustering mechanics.

With some additional regularity conditions on the set of regressors, this high level assumption is sufficient to guarantee consistency using the within-cluster transformation in conjunction with pooled OLS. Details of the estimation procedure are given in Section 4 and of the regularity conditions and theoretical bounds in Section 5. Some data generating processes and clustering mechanisms that suffice these high level conditions are discussed in Section 5. These are not exhaustive, however, and may suggest avenues for future research.

3 Existing Literature

This section details some existing literature on factor models and multidimensional low-rank problems. This section is also utilised to make some useful definitions that are vital to the remainder of the paper. The section starts with a brief heuristic explanation of how the problem may be transformed into a simple two-dimensional panel data model and estimated using methods in that literature. The multidimensional array framework is then discussed along with the literature related to low-rank approximation in those settings.

3.1 Matrix low-rank approximation estimator

This section provides a heuristic description of some matrix methods that can be applied directly to the multidimensional model. Specific regularity conditions are left to the results in Section 5 for brevity. As such, this subsection serves to highlight settings where matrix methods are sufficient and where they are not, and to justify the use of the more elaborate multidimensional methods. Note that Kapetanios, Serlenga and Shin (2021) employ a similar approach for three-dimensional arrays in conjunction with the Pesaran (2006) common correlated effects estimator.

Consider, first, recasting the multidimensional array problem into a two dimensional panel problem by reorganising the indices, e.g. with the first dimension varying over the row space and the remaining indices jointly varying over the column space. This procedure is known as flattening, and the formal definition is reserved until Section 4 for reference. Again, assume a three dimensional model for simplicity. This amounts to the regression line,

$$Y_{(1)} = X'_{(1)}\beta + \varphi^{(1)}\Gamma' + \varepsilon_{(1)}$$

where $Y_{(1)}, X_{(1)}, \varepsilon_{(1)} \in \mathbb{R}^{N_1 \times N_2 N_3}$, $\varphi^{(1)}$ is an $N_1 \times L_1$ matrix and Γ is an $N_2 N_3 \times L_1$ matrix that accounts for variation in $\varphi^{(2)}$ and $\varphi^{(3)}$. The term L_1 is indexed by the dimension 1 because it may vary non-trivially according to the flattened dimension. It should then be apparent that this is exactly the model described in (2), that is, the standard linear model with factor structure unobserved heterogeneity as studied in Bai (2009), where the second dimension is simply a combination of j and t.

Hence, a standard factor model that estimates at least L_1 factors should result in consistent estimation of the slope coefficients, see Moon and Weidner (2017). However, this relies on an important structural feature of the unobserved heterogeneity term. When flattened in the chosen dimension – the first dimension in the above example – the rank of the matrix after flattening must be low relative to data size. This implies that to successfully project out the variation in the fixed-effect term either the matrix of fixed-effects from any flattening is low-rank, or, at least one flattening leads to a low-rank matrix of fixed-effects and the analyst knows which

flattening this is. To use the above example again, this means the analyst knows that $\varphi^{(1)}\Gamma'$ is low-rank, hence flattening in the first dimension is the correct way to recast the model to a panel data model, and so forth for the other flattenings. Whilst requiring low-rankness in at least one dimension may be an acceptable restriction, having knowledge of which dimension this low-rankness resides in is potentially more restrictive.

To understand the problem, consider the following two examples, one where the flattening is not low-rank and one where it is. First, assume $\varphi^{(1)}$ varies in a high-dimensional parameter space, e.g. with $N_1 < N_2N_3$, $\varphi^{(1)} \in \mathbb{R}^{N_1 \times N_1}$ and $\Gamma \in \mathbb{R}^{N_2N_3 \times N_1}$ with each column mutually orthogonal for both these matrices. Then the product of these matrices is full-rank and any factor projection approach will not fully control for this term. On the contrary, consider $\varphi^{(1)} \in \mathbb{R}^{N_1 \times N_1}$ where all columns are linearly dependent. Then the matrix $\varphi^{(1)}\Gamma'$ is rank-1 regardless of L and of how $\varphi^{(2)}$ and $\varphi^{(3)}$ vary, thus can be projected with a factor model estimated with 1 factor. Hence it is important which dimension the analyst chooses to flatten over.

Well established diagnostics in Bai and Ng (2002), Ahn and Horenstein (2013) and Hallin and Liška (2007) can be used to determine the number of factors. These diagnostics can be repeated across different flattenings, which may be informative of the dimension to use for flattening. Note these procedures require an initial guess of β and relies on this guess not eradicating the factor structure in the residual; see the beginning of Section 4.2 for a concrete example of this. It should also be noted that these diagnostics are not without restrictions and can lead to spurious conclusions on the optimal number of factors. For example, the eigenvalue ratio test in Ahn and Horenstein (2013) can undershoot the number of factors when singular values decay quickly for the leading few factors. This does not interfere with the asymptotic result in that paper but can have implications in small sample estimation. Indeed, however, these diagnostics can be helpful in both the matrix recasting of the problem and the group fixed-effects estimation in the sequel.

3.2 Multidimensional arrays

Discussed here are some important features of multidimensional arrays and their use in this paper. This includes a discussion on how to uncover cluster proxies from multidimensional array data. It also includes a discussion on why standard matrix methods do not extend well to the multidimensional setting. This section is also utilised to define important objects referenced in later sections.

First, consider how to cluster in the within-cluster transformation. In most clustering algorithms, for example K-means or K-nearest neighbour, there is some notion of a distance metric between units considered for each cluster. To arrive at a distance there must be some space to measure that distance over. For example, using some vector u_i and the Euclidean norm of

differences, $\|u_i - u_j\|$, to measure the distance between units. Algorithms to arrive at these groupings are well established when the distance metric and variable to take distance over are given. However, in this setting there is no clear variable through which to take distance over. Motivated here are methods to extract proxies that serve to measure distance across units in a way that isolates variation in each dimension of the unobserved heterogeneity term.

It is important to find proxies that isolate variation in each dimension since clustering is to be performed one index at a time. Discussed here are decompositions of multidimensional arrays that can perform this, Kolda and Bader (2009) contains a nice summary of some candidate decompositions. The method discussed here uses the higher order singular value decomposition (HOSVD), and focuses on components of this decomposition that have well formulated theoretical properties. The HOSVD is traditionally used in pursuit of a low-rank tensor decomposition by either direct truncation of left singular vectors or by some iteration approach similar to this, see for example the higher order orthogonal iteration scheme (HOOI). The problem of direct truncation, however, is not well-posed because the solution to the low-rank tensor problem may not be unique and reformulating the original tensor after the aforementioned truncation is not guaranteed to be lower tensor rank. See De Silva and Lim (2008) for an extensive explanation of the ill-posedness issues. Hence this method cannot be used in the pursuit of analytic consistency results. Problems also arise in this setting where the reformulated tensors can be arbitrarily well approximated by a tensor of lower tensor rank, which is a result of the border rank issue of the tensor rank decomposition. Whilst the HOOI scheme is implemented in simulations in Section 6 and may be well motivated from a numerical standpoint, this type of tensor decomposition is not discussed further in this paper.

Reconsider the three dimensional model with heterogeneity of the following form

$$\mathcal{A}_{ijt} = \sum_{\ell=1}^{L} \varphi_{i\ell}^{(1)} \varphi_{j\ell}^{(2)} \varphi_{t\ell}^{(3)}. \tag{9}$$

It would be convenient to use some Eckart–Young–Mirsky like theorem to arrive at a low-rank approximation for the multidimensional array $\boldsymbol{\mathcal{A}}$. This would provide an appropriate method to project out low-rank variation in the unobserved heterogeneity in the original regression model. Unfortunately such a theorem does not exist in the multidimensional setting with more than two dimensions. In fact, if multidimensional array rank is defined as the least number of outer products required to exactly decompose a multidimensional array, the low-rank problem is not well-posed, see again De Silva and Lim (2008). This motivates the use of the group fixed-effects as an alternative solution.

Take the standard singular value decomposition for some $N_1 \times N_2$ matrix A,

$$A = U\Sigma V'.$$

This decomposition is well-known in the literature and can be used, for example, to estimate fixed-effects parameters in the interactive fixed-effects model. The elements of any fixed column of U explains variation across a row space in the original matrix A, likewise the elements of any fixed column of V explain variation over a column space of A. As is convention, the columns of U and V are arranged in descending order according to their joint explanatory value of the original matrix. Thus, if A is a low-rank object then the leading few columns of U, respectively V, are useful proxies to measure closeness in the row space, respectively the column space of A. Extending this concept to the multidimensional setting turns out to be remarkably simple.

First to define some objects. Take the factor-n flattening, abbreviated to just n-flattening, of a d-order multidimensional array $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{N_1 \times \cdots \times N_d}$ to be the matrix $\mathcal{A}_{(n)} \in \mathbb{R}^{N_n \times N_{n+1} \ldots N_d N_1 \ldots N_{n-1}}$ that is simply a rearrangement of the original array. This matrix arranges the data so the n^{th} dimension varies along the row space and data from the remaining $d-1$ dimensions vary along the column space. How these remaining $d-1$ dimensions are arranged is not important - what is important is isolating variation in each dimension one at a time along the row space. With this definition of a flattening the multilinear rank can be defined as the vector $\mathbf{r} = \{r_1, \ldots, r_d\}$ where each r_n is

$$r_n := rank\big(\mathcal{A}_{(n)}\big).$$

Reconsider the singular value decomposition for matrices, applied to each of the n-flattenings of $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{N_1 \times, \ldots, \times N_d}$ as

$$\mathcal{A}_{(n)} = U^{(n)} \Sigma_n V^{(n)\prime}. \tag{10}$$

The rank of $\mathcal{A}_{(n)}$, defined above as r_n, is simply the number of non-zero elements of the diagonal matrix Σ_n. By the same logic as in the matrix case, variation over the rows of each $U^{(n)}$ explains variation over the n^{th} dimension of the multidimensional array $\boldsymbol{\mathcal{A}}$. Thus, if $\mathcal{A}_{(n)}$ is low rank, the leading few columns of $U^{(n)}$ provide good proxies for closeness in the row space of $\mathcal{A}_{(n)}$, which is the vector space of the n^{th} dimension. This shows that a simple rearrangement of the data provides readily available techniques to measure closeness in each dimension separately.

Consider for any of the dimensions n the corresponding matrix of left singular vectors from above, $U^{(n)}$, estimated with noise ε_{ijt}. That is, each U_n are calculated from the object $\boldsymbol{\mathcal{V}} = \boldsymbol{\mathcal{A}} + \boldsymbol{\varepsilon}$. Under reasonable regularity conditions on the noise term $\boldsymbol{\varepsilon}$, the left singular vectors from this decomposition comprise of a signal of the underlying parameter and noise from $\boldsymbol{\varepsilon}$. For example, in the three dimensional case, define the L_1-vector $\widehat{U}_i^{(1)}$ as the i-th row of the left singular matrix of $\boldsymbol{\mathcal{A}} + \boldsymbol{\varepsilon}$ flattened in the first dimension. Then the vector $\widehat{U}_i^{(1)}$ may comprise of,

$$\widehat{U}_i^{(1)} = \varphi_i^{(1)} + O_p\left(\frac{1}{\sqrt{\min\{N_1, N_2 N_3\}}}\right).$$

Likewise, for any dimension n define $\widehat{U}^{(n)}$ as the matrix of singular vectors from $\boldsymbol{\mathcal{A}}+\boldsymbol{\varepsilon}$ flattened in the n-th dimension, where $\boldsymbol{\mathcal{A}}$ is the unobserved fixed-effects component of interest and $\boldsymbol{\varepsilon}$ is the usual idiosyncratic noise term. Then, Bai and Ng (2002) detail conditions required for the following "up-to-rotation" consistency result, which has been amended to this paper's setting;

Lemma 1 (Theorem 1 from Bai and Ng (2002))**.** *For any fixed integer $k \geq 1$, there exists an $(r_n \times k)$ matrix H_n^k with $\operatorname{rank}(H_n^k) = \min\{k, r_n\}$ and $C_n = \min\left\{\sqrt{N_n}, \prod_{n' \neq n} \sqrt{N_{n'}}\right\}$ such that for each n under some regularity conditions*

$$C_n^2 \left\| \widehat{U}_{i_n}^{(n)} - H_n^{k\prime} \varphi_{i_n}^{(n)} \right\|^2 = O_p(1).$$

This establishes a consistency result for estimating cluster proxies and suggests these left singular vectors are viable options to cluster in each dimension. It also makes concrete the limitation implied by the value of C_n for each index - that short indices have poorly estimated proxies. However, given that the error term displayed in (8) is multiplicative across dimension, the error from this poor approximation should become negligible as long as enough other dimension proxies are well estimated. Also, the presence of the rotation matrices, H_n^k, in Lemma 1 can be ignored since these do not change relative distances of each unit under standard distance metrics used to cluster.

4 Group fixed-effects estimator

Discussed below is the main estimation procedure for the group fixed-effects estimator. In the main description one type of clustering mechanism is used, however, this can be extended to other methods to group units. These other methods, including a kernel function approach, are discussed later in the section.

4.1 Main estimator

Let $\Theta_{\mathcal{C}}$ be the space of group fixed-effects parameters associated to cluster mechanism $\mathcal{C}$, which is stated generically here. That is, each $\boldsymbol{\theta} \in \Theta_{\mathcal{C}}$ is a d-list of $\bigtimes_{n=1}^{d} N_n$ tensors. For each n in $\{1, \ldots, d\}$, the tensor θ_n varies freely over dimensions $\{1, \ldots, n-1\}$ and $\{n+1, \ldots, d\}$ but is fixed within each cluster along dimension n. The objective function for the group fixed-effect estimation of β under cluster mechanism $\mathcal{C}$ is

$$Q(\beta, \mathcal{C}) = \min_{\theta \in \Theta_{\mathcal{C}}} \left\| \mathbf{Y} - \sum_{k=1}^{K} \mathbf{X}_k \beta_k - \sum_{n=1}^{d} \theta_n \right\|_F^2 \tag{11}$$

and $\widehat{\beta}_{GFE,\mathcal{C}} := \operatorname{argmin}_{\beta \in \mathbb{R}^K} Q(\beta, \mathcal{C})$. It should be clear that the parameter space $\Theta_{\mathcal{C}}$ is indexed by clustering mechanism $\mathcal{C}$ because this mechanism defines how the parameters may vary. That

is, this is the estimated parameter space under a specific group fixed-effects estimator, which may only be an approximation of the true parameter space.

Consider taking cluster proxies from the estimated error term $\mathbf{W} = \mathbf{Y} - \mathbf{X}'\widetilde{\beta}$. Define $\widetilde{\beta}$ as the interim estimator used to obtain $\mathbf{W}$, and notice that this forms the basis of an iterative procedure, between forming clusters and estimating slope coefficients. This is illustrated in the following two-step procedure. For the below let $\widehat{L}_n$ be a hyperparameter that defines the number of singular vectors to use in the clustering stage.

1. For given $\widetilde{\beta}$, take the left singular matrices from each n-flattening of $\mathbf{W} = \mathbf{Y} - \mathbf{X}'\widetilde{\beta}$ to obtain $\{\widehat{U}_1, \widehat{U}_2, \widehat{U}_3\}$.

2. Cluster on the leading $\widehat{L}_n$ columns of $\widehat{U}_n$ to generate cluster assignments in the n^{th} dimension. Use these cluster assignments in the within-cluster transformation on $\mathbf{Y}$ and $\mathbf{X}$ then perform pooled OLS to obtain $\widehat{\beta}$.

3. Iterate steps 1 and 2 until convergence in the slope coefficients

This procedure may also be used as a debias estimator for a given initial estimate of $\widetilde{\beta}$ by ignoring step 3. Iteration here may not be stable given that step 1 and 2 do not optimise the same objective function, hence for theoretical purposes it may be convenient to only consider this as a debias procedure. In practice, iterating between step 1 and 2 after some initial grid search to initialise β may be optimal.

Of course, other clustering or transformations may be used in place of the residual clustering and within-cluster transformation. In the below, two alternatives are provided. The first maintains the within-cluster transformation but considers a different set of proxies. The second approach considers a kernel weighted transformation procedure that uses a generic set of proxies. At this stage and in the below estimator refinements the analyst may be concerned with the number of parameters required to conduct these transformation. Appendix C discusses a number of ways to reduce the size of the parameter space, including only projecting fixed-effects over a subset of dimensions and letting group sizes increase to reduce the number of groups.

4.2 Clusters on covariates

Whether used as an iterative scheme or an update, the above method has some identification issues. As an illustration take the data generating process for model (1) with just one covariate,

$$X_{ijt} = -\mathcal{A}_{ijt} + \mu_{ijt},$$

where μ_{ijt} is a white noise term. Consider an initial guess of $\widetilde{\beta} = 0$ when the true value is $\beta^0 = 1$. This leaves the residual term from Step 1 to base cluster assignment on as, $\mathbf{W} = \mathbf{Y} - \mathbf{X}\widetilde{\beta} = \mathbf{Y}$,

which reduces to $\mathbf{W} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$. Thus, clustering is based solely on noise and can be reasonably described as random. The associated within-cluster transformation will not project variation in the $\mathcal{A}$ terms that appear in both $\mathbf{Y}$ and $\mathbf{X}$ such that the OLS step in stage 2 produces

$$\widehat{\beta} \approx \frac{Var(\mu_{ijt})}{Var(\mu_{ijt}) + Var(\mathcal{A}_{ijt})} + o_p(1).$$

For $\frac{Var(\mu_{ijt})}{Var(\mathcal{A}_{ijt})} \to 0$, $\widehat{\beta} \to 0$ and the algorithm does not update the initial guess of $\widetilde{\beta} = 0$. This problem also arises in the matrix methods in Section 3.1 and is a more fundamental issue with this algorithmic approach.

This example clearly displays some identification issues with the above method. Worth noting is that this may be alleviated with a grid search approach, though this can be computationally infeasible even for a moderate number of covariates since the grid grows exponentially in the number of covariates. To avoid this, proposed below is a method to extract cluster allocations from only variation in the set of covariates. As discussed below, this clustering may also be conducted on control variables extraneous to the regression line. The two-step procedure works as follows.

1. Take the left singular matrices from each n-flattening of X to obtain $\{\widehat{U}_1, \widehat{U}_2, \widehat{U}_3\}$.

2. Cluster on the leading $\widehat{L}_n$ columns of $\widehat{U}_n$ to generate cluster assignments in the n^{th} dimension. Use these cluster assignments in the within-cluster transformation on $\mathbf{Y}$ and $\mathbf{X}$ then perform pooled OLS to obtain $\widehat{\beta}$.

An advantage of using covariate clustering is that it can make use of control variables that are a good signal of cluster but are not included in the regression line. For example, a control variable Z_i that is constant across j and t may be a good candidate to cluster along the i dimensions but will be projected out with the within-cluster transformation, so cannot be used directly in the pooled OLS estimation of β stage. This refinement also makes optimisation over β a convex problem, and no iteration is required because clustering is not a function of β estimates like in the approach in Section 4.1.

4.3 Kernel Methods

The distance metrics used for the group fixed-effects methods may also be used in kernel functions to difference fixed-effects with a Nadaraya-Watson style estimator. The estimator is then simply the within-cluster transformation with weighted means across the whole population rather than the mean within each cluster. To see this take the usual distance metric as $d(u_{i_n}, u_{i_n^*}) := \left\| u_{i_n} - u_{i_n^*} \right\|$ and apply a kernel function to take weighted difference over the fixed-effects parameter space. Take the proxy parameters used in the distance metric calculations

generically as $\widehat{\varphi}_{i_n}^{(n)}$. Then for each $n \in \{1, \ldots, d\}$ and $i_n \in \{1, \ldots, N_n\}$, define the kernel weight

$$w(i_n, j_n) := \frac{k\left(\left\|\widehat{\varphi}_{i_n}^{(n)} - \widehat{\varphi}_{j_n}^{(n)}\right\|\right)}{\sum_{i'_n} k\left(\left\|\widehat{\varphi}_{i_n}^{(n)} - \widehat{\varphi}_{i'_n}^{(n)}\right\|\right)}.$$

These weights can then be used in the standard within-transformation (5). Take, for example, the term $Y_{i_1^* i_2^* i_3}$ from (6), where the more generic index notation is used here, as,

$$Y_{i_1^* i_2^* i_3} = \sum_{j_1=1}^{N_1} \sum_{j_2=1}^{N_2} w(i_1, j_1) w(i_2, j_2) Y_{j_1 j_2 i_3}.$$

This transformation then creates an analogous estimation procedure to the cluster methods mentioned above. Proxies used in the kernel functions to create the weights can be the same singular vectors used for forming clusters and a similar analysis for the estimator performance follows. Of course, as with the group fixed-effects case, the proxies used in these kernel functions need not be the singular vectors and may be other observed control variables or other unobserved proxies. After this weighted difference, pooled OLS is used on the transformed variables to arrive at the estimator for β. Specifically, let $\check{\mathbf{A}}$ denote a tensor after the kernel weighted differencing transformation described above is applied. Then $\widehat{\beta}_{KER}$ is the pooled OLS estimator of $\check{\mathbf{Y}}$ regressed on $\check{\mathbf{X}}$.

5 Asymptotic results

This section discusses the main asymptotic results for the low-rank approximation matrix method and group fixed-effects method. The result for the group fixed-effects method is stated generically and applies to different approaches that comply with Assumption 1. Results for the kernel methods are then stated. In the following, let β^0 denote the true value of the slope coefficient.

5.1 Matrix methods

The following restrictions are required for consistency when using the Bai (2009) estimator after first flattening over a chosen dimension. The tensor rank parameter, L, may without loss be restricted to the upper bounded by $L \leq \min_n\{\prod_{n' \neq n} N_{n'}\}$. This is a result of elementary bounds on the tensor rank of an arbitrary tensor. If the multilinear rank terms depend on data dimensions then the tensor rank may be indexed as L_N, where N generically refers to the collection of dimensions N_n for $n \in \{1, \ldots, d\}$. The multilinear rank of $\boldsymbol{\mathcal{A}}$ is restricted such that it is low-rank along at least one of the flattenings. Let β^0 denote the true parameter value for the slope coefficients.

In the following assumptions let $\widehat{L}_n$ be the estimated number of factors for the (n)-flattening of the regression line when applying the least square methods in Bai (2009). Also, let $\mathcal{L} \subset \{1, \ldots, d\}$, and this this subset be arbitrary for now.

Assumption 2 (Bounded norms of covariates and exogenous error)**.**

(i). $\left\|X_k\right\|_F = O_p\left(\prod_{n=1}^{d} \sqrt{N_n}\right)$ *for each* k

(ii). $\left\|\varepsilon_{(n)}\right\| = O_p\left(\max\{\sqrt{N_{n^*}}, \prod_{m \neq n^*} \sqrt{N_m}\}\right)$ *for each* $n^* \in \mathcal{L}$

Assumption 3 (Weak exogeneity)**.** $\quad \text{vec}(X_k)'\text{vec}(\varepsilon) = O_p\left(\prod_{n=1}^{d} \sqrt{N_n}\right)$ *for each* k

Assumption 4 (Low multilinear rank)**.** *Let* $(r_1, \ldots, r_d)$ *be the multilinear rank of* $\boldsymbol{\mathcal{A}}$*. Then for some positive integer,* c*,* $r_{n^*} < c$ *for all* $n^* \in \mathcal{L}$*.*

Assumption 5 (Non-singularity)**.** *Let* $\sigma_s(A)$ *be the* s^{th} *singular value for a matrix* A*. For each dimension* $n^* \in \mathcal{L}$ *that satisfies Assumption 4, there exists a* $K \times 1$ *unit vector* δ_{n^*} *such that*

$$\sum_{s=2\widehat{L}^{(n^*)}+1}^{\min\{N_{n^*}, \prod_{m \neq n^*} N_m\}} \sigma_s\left(\frac{(\delta_{n^*} \cdot X_{(n^*)})(\delta_{n^*} \cdot X_{(n^*)})'}{\prod_n \sqrt{N_n}}\right) > b > 0 \qquad wpa1.$$

Assumptions 2, 3 and 5 are standard regularity assumptions already well established in the literature, e.g. see Moon and Weidner (2015). Assumption 2.(i) ensures that the covariates have bounded norms, for example having bounded second moments. Assumption 2.(ii) allows for some weak correlation across dimensions, see Moon and Weidner (2015), or is otherwise implied if the noise terms are independently distributed with bounded fourth moments, see Latała (2005). Assumption 3 is implied if $X_{i_1,i_2,\ldots,i_d;k}\varepsilon_{i_1,i_2,\ldots,i_d}$ are zero mean, bounded second moment and only admits weak correlation across dimensions for each $k = 1, \ldots, K$. Assumption 5 simply states that, after factor projection, the set of covariates still collectively admit full-rank variation.

Assumption 4 is new and asserts that there exists at least one flattening of the interactive term, $\boldsymbol{\mathcal{A}}$, that is low-dimensional or simply low-rank. Given that the true value for L is left mostly unrestricted at this stage, this requires that at least one of the unobserved terms $\varphi^{(n)}$ is low dimensional. Note that not all dimensions must satisfy Assumption 4 for the below result. If the correct dimension is chosen then variation from the interactive term can be sufficiently projected out using the factor model approach. This makes up the statement of the following Proposition.

Proposition 1. *Let* $\widehat{\beta}_{(n)}$ *be the estimator from Bai (2009) after first flattening along dimension* n*. If Assumptions 2-5 hold, the subset* $\mathcal{L}$ *is non-empty, and the estimated number of factors*

$\widehat{L}_{n^*} \geq r_{n^*}$, *then, for each* $n^* \in \mathcal{L}$ *satisfying Assumption 4,*

$$\left\|\widehat{\beta}_{(n^*)} - \beta^0\right\| = O_p\left(\frac{1}{\sqrt{\min\{N_{n^*}, \prod_{n\neq n^*} N_n\}}}\right). \tag{12}$$

Proposition 1 follows directly from Moon and Weidner (2015) since the flattening procedure reduces the problem to the standard linear factor model. Notice that this result only applies to estimates in the dimension(s) that satisfy the low-rank assumption in Assumption 4. That is, implicit in Proposition 1 is that the analyst has chosen the correct dimension to flatten over when reformulating the problem as a two-dimensional panel. Assumption 4 can be relaxed to $r_{n^*} = o\left(\min\{N_{n^*}, \prod_{n\neq n^*} N_n\}\right)$ as long as the estimated number of factors is allowed to increase with data size at a faster rate than this. The constraint $\widehat{L}_{n^*} \geq r_{n^*}$ can also be changed to $\widehat{L}_{n^*} \geq c$, however, this is more conservative than required for the statement of the result.

The estimation procedure from Proposition 1 can also be augmented to flatten over multiple indices. For instance, the analyst may flatten such that both the rows and columns in the matrix contain multiple indices from the original array. Of course, this augmentation makes Assumption 4 harder to satisfy as it requires multiple parameters to vary in low-dimensional space. To see this take the tensor $\boldsymbol{\mathcal{A}}$ flattened over the first two indices as $\mathcal{A}_{(1,2)} \in \mathbb{R}^{N_1 N_2 \times \prod_{n\notin\{1,2\}} N_n}$. If the parameters $\varphi^{(n)}$ for $n = 3, \ldots, d$ are high-dimensional, Assumption 4 is only satisfied when both $\varphi^{(1)}$ and $\varphi^{(2)}$ and their product space is low-dimensional. Clearly this is more restrictive than requiring only one of the parameter spaces to be low-dimensional. However, flattening along multiple dimensions can improve the convergence rate in Proposition 1 to $O_p\left(\frac{1}{\sqrt{\min\{N_1 N_2, \prod_{n\notin\{1,2\}} N_n\}}}\right)$, so there are benefits if this more restrictive assumption can be made.

As stated already, Proposition 1 takes for granted the correct choice in dimension to flatten across. Under Assumption 2.(ii) the singular values of the flattened normalised noise term dissipates as follows;

$$\frac{1}{\sqrt{\prod_n N_n}} \left\|\varepsilon_{(n)}\right\| = O_p\left(\frac{1}{\sqrt{\min\{N_n, \prod_{m\neq n} N_m\}}}\right).$$

Since $\mathcal{A}$ is a collection of fixed-effects, the normalised singular values of its flattenings are $O_p(1)$, that is, the singular values are not asymptotically negligible like those of the noise term.[3] This

[3]To see this consider the standard two dimension model and take the Frobenius norm any arbitrary component of the interactive fixed-effects term, $\lambda_r f_r'$, normalised by $1/\sqrt{NT}$

$$\frac{1}{\sqrt{NT}} \left\|\lambda_r f_r'\right\|_F = \sqrt{\frac{1}{NT}\sum_i \sum_t (\lambda_{ir} f_{tr})^2} = \sqrt{\frac{1}{N}\sum_i \lambda_{ir}^2}\sqrt{\frac{1}{T}\sum_t f_{tr}^2} = O(1).$$

The last equality comes from λ_{ir} and f_{tr} having bounded second moments.

ensures that, after flattening $\mathcal{A}$, each of the singular values eventually dominate those of the noise term. These conditions make up similar restrictions imposed in Ahn and Horenstein (2013) that allow for the use of the eigenvalue ratio test (ER) to diagnose the number of factors. Hence, in large samples, the analyst maybe able to use this test or similar to not only decide how many factors to use but also decide which dimension is likely to be low-dimensional.

Consider the factor model applied to a flattening that may not be low-rank. For a concrete example of when this can occur see the data generating process in the simulations in Section 6, where $\varphi^{(n)}$ are designed to be low-dimensional for some n, and high-dimensional otherwise. Along the dimensions of $\mathcal{A}$ that do not conform to the low-rank assumption in Assumption 4, the tail singular values may become difficult to discern from the singular values of the noise term in small samples. This means variation from those tail factors are less likely to be projected out from the factor model unless many factors are used in this projection. If r_n for $n \notin \mathcal{L}$ is allowed to increase adversely, for example at exactly the upper bound, then factor projection may never sufficiently project all relevant factors. Also, as the number of estimated factors increases, Assumption 5 becomes harder to satisfy since variation in the set of covariates is also projected out. This demonstrates the importance of choosing the correct dimension to flatten over, which is supported by the simulation results in Section 6.

5.2 Group fixed-effects methods

Below some regularity conditions additional to Assumption 1 are stated that are required for consistency of the group fixed-effects estimator. A brief discussion of some alternative restrictions precedes the statement of the result.

Below are refinements to the regularity conditions contained within the Assumptions listed in Section 5.1 that account for the within-cluster transformation.

Assumption 6 (Regularity conditions)**.** *Let $\widetilde{T}_{i_1,\ldots,i_d}$ be the entries of tensor $\mathbf{T}$ with group fixed-effects projected out according to the within-cluster transformation with cluster mechanism $\mathcal{C}$. Then,*

(i). $\left(\frac{1}{\prod_n N_n}\sum_{i_1}\cdots\sum_{i_d}\widetilde{X}_{i_1,\ldots,i_d}\widetilde{X}'_{i_1,\ldots,i_d}\right) = \Omega_{X,N}$ *exists and is positive definite along the sequence $N_1,\ldots,N_d \to \infty$. Specifically, let $\Omega_{X,N} = O_p(1)$.*

(ii). $\frac{1}{\prod_n N_n}\sum_{i_1}\cdots\sum_{i_d}\widetilde{X}_{i_1,\ldots,i_d}\varepsilon_{i_1,\ldots,i_d} = \Omega_{\varepsilon,N}$ *exists as $N_1,\ldots,N_d \to \infty$.*
Specifically, $\Omega_{\varepsilon,N} = O_p\left(\frac{1}{\sqrt{\prod_n N_n}}\right)$.

Assumption 6.(i) is very similar to Assumption 5 except that here full rank is required after the within-cluster projection rather than the factor projection. Assumption 6.(ii) is an

exogeneity condition that requires weak exogeneity in the covariates after the within-cluster transformation, which can be viewed as similar to Assumption 3. This is stricter than Assumption 3 because the noise term ε can foreseeably impact cluster allocation via its impact on the estimation of cluster proxy terms. This limitation is alleviated with, for instance, the use of cluster proxies extraneous to the regression line or perhaps through some sample splitting methods such as that proposed in Freeman and Weidner (2022). The method proposed in Section 4.2, clustering on covariates only, should also alleviate this issue if the non-transformed covariates are weakly exogenous.

Proposition 2 (Upper bound on group fixed-effects estimator)**.** *Let Assumptions 1 and 6 hold for cluster mechanism $\mathcal{C}$. Let $\mathcal{M}$ be the set defined in Assumption 1.(ii). Then, for tensor rank L_N that may depend on data size,*

$$\left\|\widehat{\beta}_{GFE,\mathcal{C}} - \beta^0\right\| = \sqrt{L_N} O_p\left(\prod_{n^* \in \mathcal{M}} \sqrt{\xi_{N_{n^*}}}\right) + O_p\left(\prod_{n=1}^{d} \frac{1}{\sqrt{N_n}}\right).$$

Before discussing this result an alternative restriction on the clustering mechanism is proposed. If clustering is performed on proxies then Assumption 1 can be stated in terms of the proxies, which forms the statement of Remark 1. This requires that the proxies are estimated consistently and form an injective mapping to the true parameters. An example of this are the conditions imposed in Freeman and Weidner (2022), stated in similar terms here:

Remark 1 (Clustering)**.** *The statement of Assumption 1 can be reformulated in terms of the cluster proxies as follows. Let $\widehat{\varphi}_{i_n}^{(n)} := \widehat{\varphi}^{(n)}\big(\varphi_{i_n}^{(n)}\big) \in \mathbb{R}^{\widehat{L}_n}$ be the proxy for individual i_n used to cluster along dimension n. Then,*

(i). For all n as $N_n \to \infty$,

$$\frac{1}{N_n} \sum_{i_n}^{N_n} \left\| \widehat{\varphi}_{i_n}^{(n)} - \widehat{\varphi}_{j_n(i_n)}^{(n)} \right\|^2 \lesssim O_p(1)$$

(ii). For a non-empty subset $\mathcal{M} \subset \{1, \ldots, d\}$ take for any $n^* \in \mathcal{M}$ a sequence $\xi_{N_{n^*}} \to 0$ as $N_{n^*} \to \infty$. Then,

$$\frac{1}{N_{n^*}} \sum_{i_{n^*}}^{N_{n^*}} \left\| \widehat{\varphi}_{i_{n^*}}^{(n^*)} - \widehat{\varphi}_{j_{n^*}(i_{n^*})}^{(n^*)} \right\|^2 = O_p(\xi_{N_{n^*}})$$

(iii). Let $\varphi_{i_n}^{(n)} \in \Phi_n$ be the L_n-column vector of fixed effects, where Φ_n are convex sets for each n. For each $a, b \in \Phi_n$ there exists a scalar $c_n > 0$ such that $\|a - b\| \leq c_n \cdot \left\|\widehat{\varphi}^{(n)}(a) - \widehat{\varphi}^{(n)}(b)\right\|$

If these alternate restrictions hold along with Assumption 6, then the bound in Proposition 2 holds for the GFE estimator.

Restrictions (i) and (ii) in Remark 1 are exactly Assumption 1.(i) and (ii) but with cluster proxies in place of the true parameter values. These are high level restrictions on the clustering mechanism that requires the mechanism to find closeness in the proxy space. Restriction (iii) in Remark 1 is an injectivity assumption on the proxy functions that demands closeness in the underlying parameter space given closeness in the proxy space. This requires that the proxies do actually provide a mapping to the true parameter space, that is, that they are reasonable proxies. An example of proxies that do this are the singular vectors from Section 3.2 that fit the requirements of Lemma 1. To see this expand the term $\|a-b\|$ and use the triangle inequality to see, $\|a-b\| \leq \|a-\widehat{\varphi}^{(n)}(a)\| + \|\widehat{\varphi}^{(n)}(b)-b\| + \|\widehat{\varphi}^{(n)}(a)-\widehat{\varphi}^{(n)}(b)\|$, where the first two terms are bound at the rate $O_p(C_n^{-1})$. Note the rotation matrices are ignored for brevity and C_n is the convergence rate from Lemma 1. Hence, asymptotically, Remark 1.(iii) can be achieved with $c_n = 1$.

This display also makes clear the bottle-neck when clustering in high-dimensional objects. The distance of the proxies, $\|\widehat{\varphi}^{(n)}(a)-\widehat{\varphi}^{(n)}(b)\|$, is difficult to bound using clustering methods when the dimension of the proxies are larger than two, see Graf and Luschgy (2002) and further discussion in Bonhomme, Lamadon and Manresa (2021). This implies that a low-dimensional set of proxies must bound the true parameter values for clustering methods to work well in this setting. Hence, whilst the relationship in restriction (iii) of Remark 1 may be satisfied for an arbitrarily high-dimensional set of proxies, for a reasonable family of cluster mechanisms to bound these proxies as per restrictions (i) and (ii) of this remark, restriction (iii) must also hold for a low-dimensional set of proxies. This can be highly restrictive with, for example, fixed-effects $\varphi_{i_n}^{(n)}$ that are high-dimensional.

Of note in Assumption 1 and Remark 1 is the subset $\mathcal{M}$ of dimensions that admit the well-behaved cluster property. As per the above, an implicit requirement on the latent parameters from this subset of dimensions is some form of low-dimensionality in the vectors $\varphi_{i_{n^*}}^{(n^*)}$ for $n^* \in \mathcal{M}$. A sufficient condition for parameters in these dimension to be low-dimensional is low multilinear rank for each $n^* \in \mathcal{M}$ for a given sample. Hence, it is expected that the set $\mathcal{M}$ should roughly coincide with the set $\mathcal{L}$, or at least be a subset of $\mathcal{L}$, from Assumption 4.

How the sequences $\xi_{N_{n^*}}$ converges to zero and how L_N is bounded are important for the convergence result in Proposition 2. First note that for fixed L_N the first term in the result simplifies to $O_p\left(\prod_{n^*\in\mathcal{M}} \sqrt{\xi_{N_{n^*}}}\right)$. Also note, if the conditions for Lemma 1 hold and clustering is based on singular vector estimates that adhere to Remark 1, then it is possible to achieve $\xi_{N_{n^*}} = O_p\left(\frac{1}{\min\{N_{n^*}, \prod_{n\neq n^*} N_n\}}\right)$. If each N_n grow at the same rate then the consistency result is,

$$\left\|\widehat{\beta}_{GFE,\mathcal{C}} - \beta^0\right\| = \sqrt{L_N} O_p\left(N_n^{-|\mathcal{M}|/2}\right).$$

In the worst case scenario $\sqrt{L_N}$ is upper bound by $\sqrt{L_N} \lesssim N_n^{(d-1)/2}$, which is taken from $L_N \leq \min_n \prod_{n' \neq n} N_{n'}$. The convergence result is then $O_p\left(N_n^{(d-1-|\mathcal{M}|)/2}\right)$, which is of course conservative but shows that if $|\mathcal{M}| = d$, then consistency is guaranteed albeit at the slow rate of $N_n^{1/2}$. This means that all dimensions must have good cluster assignments, which is obviously not an ideal worst case but shows the limitations of this method when L_N is unrestricted.

For the special case of $d = 3$ it can be shown that $L_N \leq \min_n \prod_{n' \neq n} r_{n'}$. From the discussion above, it is expected that $n \in \mathcal{M}$ is sufficient for $n \in \mathcal{L}$, that is, r_n is small for the set of dimensions $n \in \mathcal{M}$. This tightens the bound in Proposition 2 to $O_p\left(N_n^{\max\{-|\mathcal{M}|/2, 1-|\mathcal{M}|\}}\right)$, such that only $|\mathcal{M}| \geq 1$ is required for consistency. The analogous tensor rank bound is so far not known for the case with $d \geq 4$.

5.3 Kernel Methods

Described here are the restrictions required to show consistency using the kernel methods in Section 4.3.

Assumption 7 (Kernels)**.** *Take the class of kernel functions used as $\mathcal{K}$. For $k \in \mathcal{K}$, k is bounded. Then for $h > 0$ there exists an $\alpha > 0$ such that $k(a/h)a \lesssim O(h^\alpha)$.*

Assumption 7 refers to a bandwidth parameter, h, and restricts the kernels to penalise distance at a rate equal to or faster than $O(h^\alpha/a)$. For consistency using the kernel methods, the sequence $h \to 0$ is considered, such that an upper bound on α is the critical object of interest in Assumption 7 for the kernel function under consideration. As an example of a class of kernel functions that satisfies Assumption 7, the exponential class of the form considered in Remark 2 may be utilised.

Remark 2. *For $c_1, c_2 > 0$, let $k'(a) \propto c_1 exp(-c_2 a^2)$ for all $a \geq 0$ and $k' \in \mathcal{K}'$. Then* $\operatorname{argmax}_a k'(a/h)a = h/\sqrt{2c_2}$*, and,*

$$\max_a k'(a/h)a \propto \frac{c_1}{\sqrt{2c_2}} e^{-1/2} h = O(h)$$

Thus, Assumption 7 is satisfied for the exponential class of kernel functions $\mathcal{K}'$ with $\alpha = 1$. Further, for $h \to 0$, it suffices that $\alpha \in (0, 1]$.

Assumption 7 is stated more generically than Remark 2 as there is a larger class of bounded kernel functions that satisfy the sufficient restriction for result below.

Assumption 8 (Sampling)**.** *Let $\widehat{\varphi}_{i_n}^{(n)} \in \widehat{\Phi}_n$ be the proxy space for the fixed-effects and let $\mathcal{K}$ be the family of kernel functions used. Let $K_{i_n}(h_n) := \max_{j_n} k\left(\frac{1}{h_n}\left\|\widehat{\varphi}_{i_n}^{(n)} - \widehat{\varphi}_{j_n}^{(n)}\right\|\right)$. For $k \in \mathcal{K}$ and*

$0 < \varepsilon_{i_n} < K_{i_n}(h_n)$ *define*

$$M_n\left(\widehat{\varphi}_{i_n}^{(n)}, \varepsilon_{i_n}\right) := \sum_{j=1}^{N_n} \mathbb{1}\left(k\left(\frac{1}{h_n}\left\|\widehat{\varphi}_{i_n}^{(n)} - \widehat{\varphi}_{j_n}^{(n)}\right\|\right) > \varepsilon_{i_n}\right).$$

Then for any $\varepsilon_{i_n} \in (0, K_{i_n}(h_n))$,

$$\operatorname*{plim}_{N_n \to \infty} \frac{M_n\left(\widehat{\varphi}_{i_n}^{(n)}, \varepsilon_{i_n}\right)}{N_n} \geq c_{i_n} \in (0, 1]. \tag{13}$$

for all $i_n \in 1, \ldots, N_n$.

The upper bound on ε_{i_n}, K_{i_n}, is expected to be $k(0)$ for most classes of kernels. That is, the kernel function evaluated at $\widehat{\varphi}_{i_n}^{(n)} = \widehat{\varphi}_{j_n}^{(n)}$ should maximise the value of the kernel function. For example, $k(0) = c_1$ in the exponential class of kernel functions in Remark 2. Assumption 8 may, for example, be satisfied for a reasonable class of kernel functions when the parameter space is generated such that the neighbourhood around each realisation grows proportionally with the sample size. Remark 3 formalises this.

Remark 3 (Sampling)**.** *Let* $\widehat{\varphi}_{i_n}^{(n)} \in \widehat{\Phi}_n$ *and redefine* $M_n^{\varepsilon}\left(\widehat{\varphi}_{i_n}^{(n)}\right)$ *as*

$$M_n^{\varepsilon}\left(\widehat{\varphi}_{i_n}^{(n)}\right) := \sum_{j_n=1}^{N_n} \mathbb{1}\left(\widehat{\varphi}_{j_n}^{(n)} \in B_{\varepsilon}\left(\widehat{\varphi}_{i_n}^{(n)}\right)\right),$$

where $B_{\varepsilon}(x)$ *is the* ε*-neighbourhood around* x*. Then Assumption 8 is satisfied for the dimension* n *for any* $\varepsilon > 0$ *if,*

$$\operatorname*{plim}_{N_n \to \infty} \frac{M_n^{\varepsilon}\left(\widehat{\varphi}_{i_n}^{(n^*)}\right)}{N_n} \geq c_{i_n} \in (0, 1].$$

for all $\widehat{\varphi}_{i_n}^{(n)} \in \widehat{\Phi}_n$.

Assumption 8 and Remark 3 are restrictions on the data generating process of the fixed-effect proxy parameter space. These are similar to Assumption 5.5 in Altonji and Matzkin (2005) except related to the fixed-effect parameter space. Note that for these to be satisfied, the probability over the support of the fixed-effect space must be strictly positive. Defined in Remark 3 is essentially the building blocks of the probability space for the fixed-effects proxy parameter space. Whilst this paper focuses on fixed-effects, that is, effects that are taken as given and not modelled as random variables, it is still useful to understand that these parameters are sampled from some space. This space is what the restrictions in Assumption 8 and Remark 3 pertain to.

Proposition 3 (Upper bound on kernel estimator)**.** *Let the class of kernel functions used to perform weighted means difference and the proxy space used in these kernel functions satisfy Assumption 7 and 8. Extend Assumption 6 to the set of covariates after kernel weighted means are differenced. Let* $\left\|\varphi_{i_{n^*}}^{(n^*)} - \widehat{\varphi}_{i_{n^*}}^{(n^*)}\right\|^2 = O_p(C_{n^*}^{-2})$ *for* $n^* \in \mathcal{M}$ *and* $\left\|\varphi_{i_{n^*}}^{(n^*)} - \widehat{\varphi}_{i_{n^*}}^{(n^*)}\right\|^2 = O_p(1)$ *for* $n' \notin \mathcal{M}$*, where* $\mathcal{M}$ *is a non-empty subset of dimensions. Let* h_n *be the bandwidth parameter from Assumption 8. Then,*

$$\left\|\widehat{\beta}_{KER} - \beta^0\right\| = \sqrt{L_N} O_p\left(\prod_{n^* \in \mathcal{M}} \sqrt{O_p\left(C_{n^*}^{-2}\right) + O_p\left(C_{n^*}^{-1} h_{n^*}^{\alpha}\right) + O_p\left(h_{n^*}^{2\alpha}\right)}\right) + O_p\left(\prod_{n=1}^{d} \frac{1}{\sqrt{N_n}}\right).$$

For $h_n^{\alpha} \lesssim O(C_n^{-1})$ *this reduces to*

$$\left\|\widehat{\beta}_{KER} - \beta^0\right\| = \sqrt{L_N} O_p\left(\prod_{n^* \in \mathcal{M}} O_p\left(C_{n^*}^{-1}\right)\right) + O_p\left(\prod_{n=1}^{d} \frac{1}{\sqrt{N_n}}\right).$$

Proposition 3 shows that the convergence rate for the kernel estimator is bounded by the convergence rate of the proxy estimates. That is, as long as the bandwidth parameter approaches zero sufficiently fast, the kernel estimator converges at a rate no worse than the convergence of the proxies when proxies are estimated at or slower than $\sqrt{N}$-convergence. This is expected and also a good result that the kernel method does not hinder the convergence rate from these proxies. These kernel methods do, however, suffer the same curse of dimensionality as the group fixed-effects estimators since Assumption 8 becomes increasingly difficult to satisfy as the dimension of the fixed-effects increases.

6 Simulation

Table 1 shows simulation results for the following DGP,

$$\begin{aligned} Y_{ijt} &= X_{ijt}\beta + \mathcal{A}_{ijt} + \mathcal{B}_{ijt} + \varepsilon_{ijt} \\ X_{ijt} &= \mathcal{A}_{ijt} + \mathcal{B}_{ijt} + \nu_{ijt} \end{aligned}$$

with $\mathcal{A}_{ijt} = \sum_{\ell=1}^{N_1} \varphi_{i\ell}^{(1)} \varphi_{j\ell}^{(2)} \varphi_{t\ell}^{(3)}$, $\mathcal{B}_{ijt} = \alpha_{ij} + \gamma_{it} + \delta_{jt}$. Also,

$$\begin{gathered} \varepsilon_{ijt}, \nu_{ijt}, \alpha_{ij}, \gamma_{it} \text{ and } \delta_{jt} \overset{i.i.d.}{\sim} N(0,1) \text{ and for each } \ell,\ \varphi_{i\ell}^{(1)}, \varphi_{t\ell}^{(3)} \overset{i.i.d.}{\sim} N(0,1). \\ \varphi_{j1}^{(2)} \overset{i.i.d.}{\sim} N(0,1) \text{ with } \varphi_{j1}^{(2)} = \varphi_{j2}^{(2)} = \cdots = \varphi_{jN_1}^{(2)} \end{gathered}$$

$\mathcal{A}_{ijt}$ and $\mathcal{B}_{ijt}$ are normalised to have unit variance. $\boldsymbol{\mathcal{A}}$ is specified such that it is rank 1 when flattened in the second dimension and rank N_1 when flattened in either dimension one or three. That is, the multilinear rank is $\mathbf{r} = (N_1, 1, N_1)$. This comes directly from the data generating

process for each $\varphi^{(n)}$, where the matrix $\varphi^{(2)}$ is designed to be rank-1 and the matrices $\varphi^{(1)}$ and $\varphi^{(3)}$ are designed to be rank-N_1.

In Table 1, the estimators OLS and Fixed-effects are simply the pooled OLS estimator and the pooled OLS estimator after additive fixed-effects are projected out, respectively. As expected both of these two have poor bias. The four GFE estimators perform well with reasonably low bias and standard deviation. GFE (K-means) is GFE estimator with clustering based on the K-means algorithm, with proxies taken from the residual. GFE (K-means on X) is the same estimator with proxies taken from the scalar covariate of interest. GFE (1-NN) and GFE (1-NN on X) are likewise the same estimators but using the one nearest neighbours clustering. The factor model is used after first flattening along each dimension as Factor(dim $= n$), where n is the dimension used for flattening. In each case, 2 factors are projected. The results show the theoretical result succinctly, where the bias is close to zero when the correct dimension is flattened over (the second dimension in this case) and very poor bias when the incorrect dimension is used (the first and third dimensions). Lastly, the kernel differencing estimator is estimated with kernel function bandwidths 0.5, 1 and 1.5; which all have comparable bias but substantially better standard deviation for bandwidth equal 1 and 1.5.

This analysis is repeated for the four dimensional case in Table 2, where the second and third dimensions admit low-dimensional unobserved interactive fixed-effects parameters. For computational reasons, the GFE nearest neighbour estimators are omitted. The simulations suggest similar results as the three dimensional case, where the factor models perform well when flattened in the low-dimensional dimensions (second and third) and poorly in the high-dimensional dimensions (first and fourth).

3-D	Mean bias	St. dev.	MSE
OLS	0.6668	0.0033	4.45e-01
Fixed-effects	0.4997	0.0114	2.50e-01
GFE (K-means)	0.0118	0.0096	2.32e-04
GFE (K-means on X)	0.0129	0.0112	2.91e-04
GFE (1-NN)	0.0112	0.0153	3.61e-04
GFE (1-NN on X)	0.0111	0.0154	3.61e-04
Kernel (h = 0.5)	0.0030	0.0090	8.94e-05
Kernel (h = 1.0)	0.0031	0.0068	5.58e-05
Kernel (h = 1.5)	0.0037	0.0062	5.24e-05
Factor (dim = 1)	0.4319	0.0135	1.87e-01
Factor (dim = 2)	0.0030	0.0050	3.40e-05
Factor (dim = 3)	0.4319	0.0135	1.87e-01

Table 1: 3D model ($N_1 = N_2 = N_3 = 36$), with 10,000 Monte Carlo rounds. All results are in relation to β estimation.

4-D	Mean bias	St. dev.	MSE
OLS	0.6670	0.0018	4.45e-01
Fixed-effects	0.4981	0.0282	2.49e-01
GFE (K-means)	0.0012	0.0049	2.58e-05
GFE (K-means on X)	0.0013	0.0051	2.82e-05
Kernel (h = 0.5)	5.23e-05	0.0114	1.00e-04
Kernel (h = 1.0)	4.42e-05	0.0057	3.26e-05
Kernel (h = 1.5)	-1.32e-05	0.0045	2.03e-05
Factor (dim = 1)	0.3733	0.0311	1.40e-01
Factor (dim = 2)	0.0030	0.0030	1.82e-05
Factor (dim = 3)	0.0030	0.0030	1.82e-05
Factor (dim = 4)	0.3734	0.0311	1.40e-01

Table 2: 4D model ($N_1 = N_2 = N_3 = N_4 = 20$), with 10,000 Monte Carlo rounds.

A two-dimensional simulation exercise is also performed to compare the grouped fixed-effects approach to the factor model approach in a setting where theoretical results for the factor model are well known. Table 3 shows the results of this two-way setting where the data generating process is a factor model with two factors. The GFE estimators have less bias than the factor

model even when the factor model overestimates the number of factors. To see this, compare the factor estimates with 2, 4 and 6 factors projected out with the GFE estimator. For increase in variance of order ≈ 4, the GFE estimator reduces bias by an order ≈ 10. This is a surprising improvement in estimates for a setting that is purpose designed for the factor model. Where this comparison falls down is for models with a larger number of factors because generally clustering does not perform well when the latent parameter space has dimension greater than 2.

	Mean bias	St. dev.	MSE
OLS	0.6672	0.0033	4.45e-01
Fixed-effects	0.5000	0.0043	2.50-e01
GFE	0.0002	0.0090	8.05e-05
Kernel (h = 0.5)	0.0003	0.0055	3.06e-05
Kernel (h = 1.0)	0.0003	0.0054	2.87e-05
Kernel (h = 1.5)	0.0004	0.0053	2.81e-05
Factor (R = 2)	0.0024	0.0047	2.76e-05
Factor (R = 4)	0.0031	0.0048	3.25e-05
Factor (R = 6)	0.0024	0.0049	3.03e-05

Table 3: 2D model ($N_1 = N_2 = 216$), with 10,000 Monte Carlo rounds.

7 Empirical application - demand estimation for beer

The methods proposed in this paper are applied to estimated the demand elasticity for beer. Price and quantity for beer sales is taken from the Dominick's supermarket dataset for the years 1991-1995 and is related to supermarkets across the Chicago area. Price and quantity vary over three dimensions in this example – product (i), store (j) and month (t). Fixed-effects that interact across all three dimensions can control for taste shocks to beer consumption that differ over both product and store. Take for instance a large sporting event (temporary t shock) that changes preferences differently across locations (j) and across certain subsets of sponsored beer (i). For example, in the stadiums for the many NBA finals playoffs the Chicago Bulls played in the early 1990's, Miller Lite beer advertisements could be seen alongside advertisements for the substitute product Canadian Club whisky. This suggests these events attracted large marketing campaign spends for these and other beer substitute brands that most likely also included price offers at local supermarkets. Whilst the impact of these advertisements and price offers on the demand for or price of beer is not clear and, further, that it is reasonably safe to assume the econometrician does not observe the plethora of marketing campaigns around these events, the analyst would most likely still want to control for aggregate shocks like these. For this reason it is

important to use methods that robustly control for unobserved fixed-effects, such as unobserved marketing campaigns, that may impact both quantity demanded and prices in unforeseen ways.

Models for demand estimation ideally account for endogenous variation in prices and quantity. The classic instrumental variable approach is to find a variable that varies exogenously to the production process but can reasonably describe price fluctuations. A popular instrument in the estimation of beer demand is the commodity price for barley, one of the product's main ingredients, see e.g. Saleh (2014); Tremblay and Tremblay (1995); Richards and Rickard (2021). Since the price of barley is arguably not driven by the demand for it by any one supplier of beer, it can be a useful variable to instrument for price shifts. In the following, it is taken as given that the price of barley is exogenous with respect to the noise term, ε.

For validity the instrument is also required to be strong, in the sense that it is strongly correlated with price. In this dataset correlation between the price of barley, which varies over only t, and price of beer depends on how beer price is aggregated. If beer price is first integrated over i and j, such that it only varies over t, then it is highly correlated with the price of barley, at 0.61. However, if beer price is not aggregated at all it is only correlated at 0.05. This suggests there are important product and store level price drivers for beer that are not accounted for by fluctuations in the price of barley. This implies that price fluctuations in barley alone may not be viable to fully capture beer prices when considering variation over all three dimensions. For exogeneity, the price of barley must be independent of common unobserved shocks to both price and demand, which translates to being independent of $\varphi_{t,\ell}^{(3)}$ and any scalar fixed-effects that vary over t in the interactive fixed-effects model. More details are deferred to Appendix B.

An alternative method is to follow the external variable approach of Altonji and Matzkin (2005), that requires exogeneity of beer price conditional on an external variable. If the price of barley is treated as the external variable, the beer demand example requires conditional exogeneity of beer prices across time periods of similar barley prices. For example, that the price of beer is exogenous when compared to other beer prices during time periods of high barley price, and so on. Note, the external variable approach provides much more flexibility than the traditional IV approach in the multidimensional setting because it may viably be used in conjunction with the group and kernel fixed-effects approaches even when the external variable varies over a subset of dimensions. See Appendix B for an explanation of how external variables are used in this setting and how the assumption works in conjunction with the fixed-effects estimators.

Table 4 refers to estimates from the standard logit demand model,

$$\log(quantity_{ijt}) - log(quantity_{1jt}) = price_{ijt}\beta + \mathcal{A}_{ijt} + \varepsilon_{ijt}$$

where $\mathcal{A}_{ijt}$ is the usual interactive fixed-effects and no control variables are included since

the set of available controls are rank-deficient and automatically projected out with standard scalar fixed-effects and from differencing out the outside option. The outside option is encoded as product number 1 and is the aggregate consumption of products with small quantities consumed. This serves the purpose of creating an outside option to do the necessary logit demand transformation as well as to avoid issues related to an unbalanced panel for the many niche products with sparse consumption amounts. Own price elasticity is calculated as $\eta_{ijt} = price_{ijt}\beta(1 - quantity_{ijt}/\sum_{ijt} quantity_{ijt})$ and the mean elasticity is taken as the mean of this measure for each estimator. The pooled instrumental variable and external variable estimates estimate relatively large elasticities. However, all of the fixed-effects approaches estimate statistically similar slope coefficients and elasticities at the mean. This implies that whilst some fixed-effects may exist in the true model for demand, they are unlikely complex enough to require the high-dimensional projections from the GFE or kernel methods. This of course also takes for granted that the IIA logit model is the true model for demand.

Estimator	Coefficient (bootstrap st. dev.)	Elasticity at mean
Pooled OLS	-0.60 (0.04)	-3.26
Pooled IV	-0.72 (0.04)	-3.91
Pooled EV	-0.71 (0.05)	-3.86
Additive FE	-0.32 (0.05)	-1.74
Factor (dim = 1)	-0.29 (0.04)	-1.58
Factor (dim = 2)	-0.32 (0.05)	-1.74
Factor (dim = 3)	-0.37 (0.05)	-2.01
GFE	-0.32 (0.05)	-1.74
GFE (EV)	-0.31 (0.05)	-1.68
Kernel (Gaussian)	-0.30 (0.05)	-1.63
Kernel (EV, Gaussian)	-0.34 (0.08)	-1.85

Table 4: Logit demand slope estimates and elasticities at the mean (73 products, 41 stores, 57 months).

Standard deviations were bootstrapped by resampling along each dimension separately. In the first dimension, product 1 is fixed across bootstrap samples as the outside option and the remaining products are resampled with replacement.

The second column from Table 5 refers to the estimates for demand elasticities for the following regression model,

$$\log(quantity_{ijt}) = \log(price_{ijt})\beta + \mathcal{A}_{ijt} + \varepsilon_{ijt} \tag{14}$$

where $\mathcal{A}_{ijt}$ is the usual interactive fixed-effects term from the prequel. This amounts to estimating the standard log-log model for demand with fixed-effects. That is,

$$quantity_{ijt} = price_{ijt}^{\beta} \exp(\mathcal{A}_{ijt} + \varepsilon_{ijt}).$$

Again, no controls are included here since they are low-dimensional and subsumed by the fixed-effects term. This model specification estimates reasonably similar elasticities as the logit case across each of the different fixed-effects estimators but relatively large differences in estimates for pooled OLS, IV and external variable estimators. The similar elasticities for the different fixed-effects estimators within Table 5 again suggests that whilst some form of fixed-effects should be included, they may not need be as complex as implied by the GFE and kernel methods.

The third column from Table 5 reports estimates of the same log-log model controlling for the average log price of other products,

$$\log(quantity_{ijt}) = \log(price_{ijt})\beta + \delta \sum_{i' \neq i} \log(price_{i'jt}) + \mathcal{A}_{ijt} + \varepsilon_{ijt}. \tag{15}$$

This model assumes homogeneous cross-elasticity over all other beer products. That is, it refers to the demand model,

$$quantity_{ijt} = price_{ijt}^{\beta} \prod_{i' \neq i} price_{i'jt}^{\delta_{ii'}} \exp(\mathcal{A}_{ijt} + \varepsilon_{ijt}),$$

where $\delta_{ii'} = \delta$ for all i and i'. Whilst this may oversimplify the system of cross-elasticities in the market for beer, it does significantly change the estimates for β in the log-log model. This suggests that cross-elasticities should probably be controlled for since β estimates do seem sensitive to their inclusion. Note that most estimators returned a negative value for δ, which opposes the theory that other brands of beer, on aggregate, are substitutes. However, since prices are aggregated in such a crude way, the cross-elasticity estimates should not be taken too seriously. If interested in the cross-elasticities, then some care should be taken to segment or group products in such a way that actual substitution is being identified here, not just aggregate market forces. For this model, all fixed-effects estimates are within statistical noise of each other, this time with the external variable approach being closely aligned. IV is estimated with very high variation in both log-log models, which may be due to barley being a weak instrument.

8 Conclusion

This paper shows methods to generalise the interactive fixed-effect to the multidimensional case with more than two dimensions. Theoretical results show that standard matrix methods can be applied to this setting but require additional knowledge of the data generating process. The

Estimator	$\widehat{\beta}$ (St. dev.) from (14)	$\widehat{\beta}$ (St. dev.) from (15)
Pooled OLS	0.06 (0.31)	0.09 (0.30)
Pooled IV	-4.76 (2.76)	-3.47 (2.30)
Pooled EV	-3.05 (0.48)	-2.74 (0.27)
Additive FE	-1.78 (0.34)	-2.80 (0.28)
Factor (dim = 1)	-1.61 (0.30)	-2.65 (0.25)
Factor (dim = 2)	-1.78 (0.33)	-2.78 (0.27)
Factor (dim = 3)	-2.09 (0.32)	-2.89 (0.27)
GFE	-1.85 (0.33)	-2.86 (0.30)
GFE (EV)	-1.84 (0.36)	-2.70 (0.30)
Kernel (Gaussian)	-1.72 (0.33)	-2.58 (0.29)
Kernel (EV, Gaussian)	-1.97 (0.47)	-2.81 (0.43)

Table 5: Log-log demand elasticities (73 products, 41 stores, 57 months).

Standard deviations were bootstrapped by resampling along each dimension separately. In the first dimension, product 1 is fixed across bootstrap samples.

multiplicative interactive error from the group fixed-effects and kernel methods show a potential improvement on the asymptotic rate of convergence and suggest a more robust approach to projecting fixed-effects. Simulations corroborate these theoretical results and show the relative advantage of using a standard factor model when the structure of the interactive term is known. They also show the robustness of the group fixed-effects estimator to not having this same knowledge. Inference in these models is still an open question for further research.

The model is applied to a simple demand model for beer consumption. The application demonstrated how the GFE and kernel methods integrate well with the external variable approach in Altonji and Matzkin (2005) as apposed to instrumental variable approaches that do not allow for fixed-effect estimation when the instrument is rank-deficient. The application shows that whilst some fixed-effects should likely be included in the model for beer demand, they are unlikely to be overly complicated to justify the GFE or kernel methods. This is a useful analysis, as it provides a robustness check for the specification of fixed-effects in model specifications.

References

Ahn, S. C. and A. R. Horenstein (2013). Eigenvalue ratio test for the number of factors. *Econometrica 81*(3), 1203–1227.

Altonji, J. G. and R. L. Matzkin (2005). Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica 73*(4), 1053–1102.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica 77*(4), 1229–1279.

Bai, J. and S. Ng (2002, January). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

Bonhomme, S., T. Lamadon, and E. Manresa (2021). Discretizing unobserved heterogeneity. *Econometrica (Forthcoming)*.

Chen, M., I. Fernández-Val, and M. Weidner (2021). Nonlinear factor models for network and panel data. *Journal of Econometrics 220*(2), 296–324.

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2013). Average and quantile effects in nonseparable panel models. *Econometrica 81*(2), 535–580.

De Silva, V. and L.-H. Lim (2008). Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis and Applications 30*(3), 1084–1127.

Elden, L. and B. Savas (2011). Perturbation theory and optimality conditions for the best multilinear rank approximation of a tensor. *SIAM journal on matrix analysis and applications 32*(4), 1422–1450.

Evdokimov, K. (2010). Identification and estimation of a nonparametric panel data model with unobserved heterogeneity. *Department of Economics, Princeton University*.

Fernández-Val, I., H. Freeman, and M. Weidner (2021). Low-rank approximations of nonseparable panel models. *The Econometrics Journal 24*(2), C40–C77.

Freeman, H. and M. Weidner (2022). Linear panel regressions with two-way unobserved heterogeneity. *arXiv preprint arXiv:2109.11911*.

Graf, S. and H. Luschgy (2002). Rates of convergence for the empirical quantization error. *The Annals of Probability 30*(2), 874–897.

Hallin, M. and R. Liška (2007). Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association 102*(478), 603–617.

Hoderlein, S. and H. White (2012). Nonparametric identification in nonseparable panel data models with generalized fixed effects. *Journal of Econometrics 168*(2), 300–314.

Kapetanios, G., L. Serlenga, and Y. Shin (2021). Estimation and inference for multi-dimensional heterogeneous panel datasets with hierarchical multi-factor error structure. *Journal of Econometrics 220*(2), 504–531.

Kolda, T. G. and B. W. Bader (2009). Tensor decompositions and applications. *SIAM review 51*(3), 455–500.

Kruskal, J. B. (1989). Rank, decomposition, and uniqueness for 3-way and n-way arrays. In *Multiway data analysis*, pp. 7–18.

Latała, R. (2005). Some estimates of norms of random matrices. *Proceedings of the American Mathematical Society 133*(5), 1273–1282.

Li, X., A. Wang, J. Lu, and Z. Tang (2019). Statistical performance of convex low-rank and sparse tensor recovery. *Pattern Recognition 93*, 193–203.

Matyas, L. (2017). The econometrics of multi-dimensional panels. *Advanced studies in theoretical and applied econometrics. Berlin: Springer*.

Menzel, K. (2021). Bootstrap with cluster-dependence in two or more dimensions. *Econometrica 89*(5), 2143–2188.

Moon, H. R. and M. Weidner (2015). Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica 83*(4), 1543–1579.

Moon, H. R. and M. Weidner (2017). Dynamic linear panel regression models with interactive fixed effects. *Econometric Theory 33*(1), 158–195.

Pesaran, M. H. (2006, 07). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica 74*(4), 967–1012.

Rabanser, S., O. Shchur, and S. Günnemann (2017). Introduction to tensor decompositions and their applications in machine learning. *arXiv preprint arXiv:1711.10781*.

Richards, T. J. and B. J. Rickard (2021). Dynamic model of beer pricing and buyouts. *Agribusiness 37*(4), 685–712.

Saleh, J. C. (2014). Simple estimators for cross price elasticity parameters with product differentiation: panel data methods and testing. *Revista de la Competencia y la Propiedad Intelectual 10*(18), 15–40.

Tomioka, R., K. Hayashi, and H. Kashima (2010). Estimation of low-rank tensors via convex optimization. *arXiv preprint arXiv:1010.0789*.

Tremblay, C. H. and V. J. Tremblay (1995). Advertising, price, and welfare: evidence from the us brewing industry. *Southern Economic Journal*, 367–381.

Vannieuwenhoven, N., R. Vandebril, and K. Meerbergen (2012). A new truncation strategy for the higher-order singular value decomposition. *SIAM Journal on Scientific Computing 34*(2), A1027–A1052.

Xu, A.-B. (2020). Tensor completion via a low-rank approximation pursuit. *arXiv preprint arXiv:2004.08872*.

Zeleneev, A. (2020). Identification and estimation of network models with nonparametric unobserved heterogeneity. *Department of Economics, Princeton University*.

A Proofs

Proof of Proposition 2. In the following, let $\text{vec}(\widetilde{\mathbf{X}})$ be the $\prod_n N_n \times K$ matrix of vectorised covariates after the within-cluster transformations where each column is a vectorised transformed covariate. The vec operator on other variables is the standard vectorisation operator. Also let $N = \prod_n N_n$ and the subscript $i = 1, \ldots, N$ be the index for the vectorised data when i has no subscript. Then,

$$\begin{aligned}
\beta_{GFE,\mathcal{C}} &= \left(\text{vec}(\widetilde{\mathbf{X}})'\text{vec}(\widetilde{\mathbf{X}})\right)^{(-1)} \text{vec}(\widetilde{\mathbf{X}})'\text{vec}(\widetilde{\mathbf{Y}}) \\
&= \left(\text{vec}(\widetilde{\mathbf{X}})'\text{vec}(\widetilde{\mathbf{X}})\right)^{(-1)} \text{vec}(\widetilde{\mathbf{X}})' \left(\text{vec}(\widetilde{\mathbf{X}})\beta^0 + \text{vec}(\widetilde{\boldsymbol{\mathcal{A}}}) + \text{vec}(\widetilde{\boldsymbol{\varepsilon}})\right) \\
&= \beta^0 + \left(\text{vec}(\widetilde{\mathbf{X}})'\text{vec}(\widetilde{\mathbf{X}})\right)^{(-1)} \text{vec}(\widetilde{\mathbf{X}})' \left(\text{vec}(\widetilde{\boldsymbol{\mathcal{A}}}) + \text{vec}(\widetilde{\boldsymbol{\varepsilon}})\right),
\end{aligned}$$

such that,

$$\begin{aligned}
\left\|\beta_{GFE,\mathcal{C}} - \beta^0\right\| &= \left\| \left(\text{vec}(\widetilde{\mathbf{X}})'\text{vec}(\widetilde{\mathbf{X}})\right)^{(-1)} \text{vec}(\widetilde{\mathbf{X}})' \left(\text{vec}(\widetilde{\boldsymbol{\mathcal{A}}}) + \text{vec}(\widetilde{\boldsymbol{\varepsilon}})\right) \right\| \\
&\leq \|\kappa_N\| + \|\omega_N\|
\end{aligned}$$

where

$$\begin{aligned}
\|\kappa_N\| &:= \left\| \left(\text{vec}(\widetilde{\mathbf{X}})'\text{vec}(\widetilde{\mathbf{X}})\right)^{(-1)} \text{vec}(\widetilde{\mathbf{X}})'\text{vec}(\widetilde{\boldsymbol{\mathcal{A}}}) \right\|; \\
\|\omega_N\| &:= \left\| \left(\text{vec}(\widetilde{\mathbf{X}})'\text{vec}(\widetilde{\mathbf{X}})\right)^{(-1)} \text{vec}(\widetilde{\mathbf{X}})'\text{vec}(\widetilde{\boldsymbol{\varepsilon}}) \right\|.
\end{aligned}$$

The terms κ_N and ω_N are dealt with separately.

First to bound κ_N. Notice,

$$\|\kappa_N\| \leq \left\| \left(\text{vec}(\widetilde{\mathbf{X}})'\text{vec}(\widetilde{\mathbf{X}})\right)^{(-1)} \right\|_F \left\|\text{vec}(\widetilde{\mathbf{X}})'\text{vec}(\widetilde{\boldsymbol{\mathcal{A}}})\right\|$$

Focus on the right hand part, and let $\langle ., . \rangle_F$ be the Frobenius inner product,

$$\left\|\text{vec}(\widetilde{\mathbf{X}})'\text{vec}(\widetilde{\boldsymbol{\mathcal{A}}})\right\| = \left\| \begin{bmatrix} \langle \widetilde{X}_1, \widetilde{\boldsymbol{\mathcal{A}}} \rangle_F \\ \vdots \\ \langle \widetilde{X}_K, \widetilde{\boldsymbol{\mathcal{A}}} \rangle_F \end{bmatrix} \right\| \leq \left\| \begin{bmatrix} \sum_{i=1}^N \left| \widetilde{X}_{i,1} \widetilde{\mathcal{A}}_i \right| \\ \vdots \\ \sum_{i=1}^N \left| \widetilde{X}_{i,K} \widetilde{\mathcal{A}}_i \right| \end{bmatrix} \right\| \tag{A.1}$$

where the triangle inequality is used entry-wise. By Hölder's inequality

$$\sum_{i=1}^N \left| \widetilde{X}_{i,k} \widetilde{\mathcal{A}}_i \right| \leq \left\|\text{vec}(\widetilde{\mathbf{X}}_k)\right\| \left\|\text{vec}(\widetilde{\boldsymbol{\mathcal{A}}})\right\| \quad \text{for each } k = 1, \ldots, K$$

This bounds the norm in (A.1) as,

$$\left\|\text{vec}(\widetilde{\mathbf{X}})'\text{vec}(\widetilde{\boldsymbol{\mathcal{A}}})\right\| \leq \sqrt{\sum_{k=1}^K \left\|\text{vec}(\widetilde{\mathbf{X}}_k)\right\|^2} \left\|\text{vec}(\widetilde{\boldsymbol{\mathcal{A}}})\right\|.$$

From Assumption 2.(i) there is $\sqrt{\sum_{k=1}^{K}\left\|\text{vec}(\widetilde{\mathbf{X}}_k)\right\|^2} = O_p\left(\sqrt{\prod_n N_n}\right)$. This leaves $\left\|\text{vec}(\widetilde{\boldsymbol{\mathcal{A}}})\right\|$. Take $g_n(i_n)$ as the indices in i_n's cluster such that $|g_n(i_n)|$ is the cluster size. Also, let $\bar{\varphi}_{i_n^*}^{(n)}$ be the cluster average for i_n's cluster. Then,

$$\begin{aligned}
\left\|\text{vec}(\widetilde{\boldsymbol{\mathcal{A}}})\right\|^2 = \left\|\widetilde{\boldsymbol{\mathcal{A}}}\right\|_F^2 &= \sum_{i_1,\ldots,i_d}\left(\sum_{\ell=1}^{L}\prod_{n=1}^{d}\left(\varphi_{i_n,\ell}^{(n)} - \bar{\varphi}_{i_n^*,\ell}^{(n)}\right)\right)^2 \\
\text{(Jensen's inequality)} \qquad &\leq L^2\sum_{i_1,\ldots,i_d}\sum_{\ell=1}^{L}\frac{1}{L}\prod_{n=1}^{d}\left(\varphi_{i_n,\ell}^{(n)} - \bar{\varphi}_{i_n^*,\ell}^{(n)}\right)^2 \\
&\leq L\left(\prod_{n=1}^{d}N_n\right)\prod_{n=1}^{d}\frac{1}{N_n}\sum_{i_n}\left\|\varphi_{i_n}^{(n)} - \bar{\varphi}_{i_n^*}^{(n)}\right\|^2.
\end{aligned}$$

Expand the term, $\left\|\varphi_{i_n}^{(n)} - \bar{\varphi}_{i_n^*}^{(n)}\right\|^2$,

$$\begin{aligned}
\left\|\varphi_{i_n}^{(n)} - \bar{\varphi}_{i_n^*}^{(n)}\right\|^2 &= \left\|\varphi_{i_n}^{(n)} - \frac{1}{|g_n(i_n)|}\sum_{j_n\in g_n(i_n)}\varphi_{j_n}^{(n)}\right\|^2 \\
&\leq \frac{1}{|g_n(i_n)|^2}\left(\sum_{j_n\in g_n(i_n)}\left\|\varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)}\right\|\right)^2 \\
&\leq \max_{j_n\in g_n(i_n)}\left\|\varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)}\right\|^2 \qquad \text{(A.2)}
\end{aligned}$$

Then by Assumption 1,

$$\left\|\text{vec}(\widetilde{\mathbf{X}})'\text{vec}(\widetilde{\boldsymbol{\mathcal{A}}})\right\| \leq \sqrt{L}\left(\prod_{n=1}^{d}N_n\right)O_p\left(\prod_{n\in\mathcal{M}}\sqrt{\xi_{N_n}}\right)$$

Lastly, Assumption 6.(i) implies the left hand term of $\|\kappa_N\|$ is $O_p(1/\prod_n N_n)$. This leaves

$$\|\kappa_N\| = \sqrt{L}O_p\left(\prod_{n^*\in\mathcal{M}}\sqrt{\xi_{N_{n^*}}}\right)$$

Finally, to bound $\|\omega_N\|$. Note that

$$\|\omega_N\| \leq \left\|\left(\text{vec}(\widetilde{\mathbf{X}})'\text{vec}(\widetilde{\mathbf{X}})\right)^{(-1)}\right\|_F\left\|\text{vec}(\widetilde{\mathbf{X}})'\text{vec}(\widetilde{\boldsymbol{\varepsilon}})\right\|.$$

Use Assumption 3 to bound the right hand term, $\left\|\text{vec}(\widetilde{\mathbf{X}})'\text{vec}(\widetilde{\boldsymbol{\varepsilon}})\right\| = O_p\left(\sqrt{\prod_n N_n}\right)$. Then, as above, the left hand term is $O_p(1/\prod_n N_n)$ such that

$$\|\omega_N\| = O_p\left(\frac{1}{\sqrt{\prod_n N_n}}\right).$$

■

Proof of Remark 1. Begin from A.2 in the proof of Proposition 2. The right hand terms, $\left\|\varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)}\right\|^2$, are bound as,

$$\left\|\varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)}\right\|^2 \leq c_n^2 \left\|\widehat{\varphi}_{i_n}^{(n)} - \widehat{\varphi}_{j_n}^{(n)}\right\|^2,$$

by Remark 1.(iii). The result then follows immediately by applying the conditions Remark 1.(i) and (ii) after this inequality. ■

Proof of Proposition 3. The interactive fixed-effect approximation error can be summarised as,

$$\left\|\mathrm{vec}(\widetilde{\boldsymbol{\mathcal{A}}})\right\|^2 = \sum_{i_1,\ldots,i_d} \left(\sum_{\ell=1}^{L}\prod_{n=1}^{d}\left(\varphi_{i_n,\ell}^{(n)} - \bar{\varphi}_{i_n^*,\ell}^{(n)}\right)\right)^2.$$

By similar steps as the proof of Proposition 2 this can be bound by

$$\left\|\mathrm{vec}(\widetilde{\boldsymbol{\mathcal{A}}})\right\|^2 \leq L\left(\prod_{n=1}^{d} N_n\right)\prod_{n=1}^{d}\frac{1}{N_n}\sum_{i_n}\left\|\varphi_{i_n}^{(n)} - \bar{\varphi}_{i_n^*}^{(n)}\right\|^2.$$

Concentrate on the last term, $\frac{1}{N_n}\sum_{i_n}\left\|\varphi_{i_n}^{(n)} - \bar{\varphi}_{i_n^*}^{(n)}\right\|^2$,

$$\frac{1}{N_n}\sum_{i_n}\left\|\varphi_{i_n}^{(n)} - \bar{\varphi}_{i_n^*}^{(n)}\right\|^2 \leq \frac{1}{N_n}\sum_{i_n}\frac{\left(\sum_{j_n} k\left(\frac{1}{h_n}\left\|\widehat{\varphi}_{i_n}^{(n)} - \widehat{\varphi}_{j_n}^{(n)}\right\|\right)\left\|\varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)}\right\|\right)^2}{\left(\sum_{j_n} k\left(\frac{1}{h_n}\left\|\widehat{\varphi}_{i_n}^{(n)} - \widehat{\varphi}_{j_n}^{(n)}\right\|\right)\right)^2}, \tag{A.3}$$

where elementary norm bounds are used to bound the left hand side.

Use as shorthand $\widehat{a}_{ij}^{(n)} := \left\|\widehat{\varphi}_{i_n}^{(n)} - \widehat{\varphi}_{j_n}^{(n)}\right\|$. Expand $\left\|\varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)}\right\|$ around the proxies for each fixed-effect term and bound using the triangle inequality as,

$$\begin{aligned}\left\|\varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)}\right\| &\leq \left\|\varphi_{i_n}^{(n)} - \widehat{\varphi}_{i_n}^{(n)}\right\| + \left\|\varphi_{j_n}^{(n)} - \widehat{\varphi}_{j_n}^{(n)}\right\| + \left\|\widehat{\varphi}_{i_n}^{(n)} - \widehat{\varphi}_{j_n}^{(n)}\right\| \\ &= O_p(C_n^{-1}) + \widehat{a}_{ij}^{(n)}.\end{aligned}$$

Then by the Cauchy-Schwarz inequality the term (A.3) can be bound as,

$$\begin{aligned}\frac{1}{N_n}\sum_{i_n}\left\|\varphi_{i_n}^{(n)} - \bar{\varphi}_{i_n^*}^{(n)}\right\|^2 &\leq \frac{1}{N_n}\sum_{i_n}\frac{\left(\sum_{j_n} k\left(\widehat{a}_{ij}^{(n)}/h_n\right)\left(O_p(C_n^{-1}) + \left(\widehat{a}_{ij}^{(n)}\right)\right)\right)^2}{\left(\sum_{j_n} k\left(\widehat{a}_{ij}^{(n)}/h_n\right)\right)^2} \\ \text{(Assumption 8)} \quad &= \frac{1}{N_n}\sum_{i_n}\left(\sum_{j_n} k\left(\widehat{a}_{ij}^{(n)}/h_n\right)\left(O_p(C_n^{-1}) + \widehat{a}_{ij}^{(n)}\right)\right)^2 O_p(N_n^{-2}) \\ &= \frac{1}{N_n}\sum_{i_n}\left(O_p(C_n^{-1})\sum_{j_n} k\left(\widehat{a}_{ij}^{(n)}/h_n\right) + \sum_{j_n} k\left(\widehat{a}_{ij}^{(n)}/h_n\right)\widehat{a}_{ij}^{(n)}\right)^2 O_p(N_n^{-2})\end{aligned} \tag{A.4}$$

Where Assumption 8 implies that $\left(\sum_{j_n} k\left(\widehat{a}_{ij}^{(n)}/h_n\right)\right)^2$ converges in probability at the rate N_n^2 to a strictly positive constant such that the inverse is also convergent by the continuous mapping theorem. Note that the discontinuity of the inverse function is at $\left(\sum_{j_n} k\left(\widehat{a}_{ij}^{(n)}/h_n\right)\right)^2/N_n^2 = 0$, which has zero probability by Assumption 8. This shows, $\left(\sum_{j_n} k\left(\widehat{a}_{ij}^{(n)}/h_n\right)\right)^{-2} = O_p(N_n^{-2})$.

Since the class of kernels are bounded, $\sum_{j_n} k\left(\widehat{a}_{ij}^{(n)}/h_n\right) = O(N_n)$ and from Assumption 7 there is $\sum_{j_n} k\left(\widehat{a}_{ij}^{(n)}/h_n\right)\widehat{a}_{ij}^{(n)} = O(h_n^{\alpha})$. Thus, (A.4) continues,

$$\begin{aligned}\frac{1}{N_n}\sum_{i_n}\left\|\varphi_{i_n}^{(n)} - \bar{\varphi}_{i_n^*}^{(n)}\right\|^2 &\leq \frac{1}{N_n}\sum_{i_n}\left(O_p(C_n^{-1}N_n) + O(h_n^{\alpha}N_n)\right)^2 O_p(N_n^{-2}) \\ &= O_p(C_n^{-2}) + O_p(C_n^{-1}h_n^{\alpha}) + O_p(h_n^{2\alpha}) \end{aligned} \tag{A.5}$$

Summarising leaves

$$\frac{1}{N_n}\sum_{i_n}\left\|\varphi_{i_n}^{(n)} - \bar{\varphi}_{i_n^*}^{(n)}\right\|^2 \leq O_p\left(C_n^{-2}\right) + O_p\left(C_n^{-1}h_n^{\alpha}\right) + O_p\left(h_n^{2\alpha}\right).$$

By similar arguments in the proof of Proposition 2 this makes up the first component of Proposition 3. The second component comes directly from extending Assumption 6 to the kernel estimator, as per the proof of Proposition 2. ■

B Demand application: fixed-effects with external variables

This section provides an explanation of how the external variable approach can work in conjunction with the GFE and kernel fixed-effects approach and briefly how the instrumental variable approach is limited in the presence of fixed-effects.

Take the interactive fixed-effects model and consider the usual IV approach. For simplicity, take the two equation model for IV estimation as,

$$\begin{aligned} Y &= X\beta + \mathcal{A} + \mathcal{B} + \varepsilon \\ X &= Z\delta + \mathcal{A} + \mathcal{B} + \nu \end{aligned} \tag{A.6}$$

where, as before, $\mathcal{A} = \sum_{\ell=1}^{L} \varphi_{\ell}^{(1)} \circ \varphi_{\ell}^{(2)} \circ \varphi_{\ell}^{(3)}$ and $\mathcal{B}$ is a collection of scalar fixed-effects. If Z varies over only one dimensions, like the price of barley, then it is not possible to project these fixed-effects out. This is because once X is projected onto Z in the first stage, it then too only varies over t and reduces the second stage problem to one dimension. Hence, Z must be mean independent of $\mathcal{A}+\mathcal{B}$ to maintain the exogeneity condition. In the beer demand application this insists that Z_t must be independent of any time related shocks that impact both demand and supply for beer; that is, the usual IV exogeneity condition with respect to the time fixed-effect in

the interactive term. For mean independence with $\mathcal{B}$ it is required that Z_t is independent of any scalar fixed-effect that varies over t. For independence with $\mathcal{A}$, it suffices that Z_t is independent of $\varphi_{t\ell}^{(3)}$ for all ℓ.

The external variable model is implemented using the external variable as a kernel proxy to difference out a weighted mean in the t dimension. The advantage here is that this method does not project the covariates onto the external variable space like with instrumental variables, hence does not reduce the covariates dimensionality. This means it can be used in conjunction with standard fixed-effects methods or GFE and kernel methods. The exogeneity condition here, though, is different to the standard instrumental variable condition.

Required in the external variable setting is some form of continuous mapping from the fixed-effect parameter in the t dimension and the external variable. For example, there could be a Lipschitz continuous mapping as,

$$\|\varphi^{(3)}(z) - \varphi^{(3)}(z_0)\| \leq C\|z - z_0\|.$$

Then, as long as Z_t is sampled from a distribution with strictly positive probability over its support, closeness in the external variable can be found, which in turns bounds variation in the fixed-effect parameter. Hence, with some form of contraction from Z_t to $\varphi_t^{(3)}$, $\mathcal{A}$ can be projected out. Because other fixed-effects can be included, $\mathcal{B}$ can be projected out with simple additive fixed-effects and the other terms in $\mathcal{A}$ can also be differenced out using the GFE or kernel methods. This leaves the remainder error, from GFE or kernel,

$$\sum_{\ell=1}^{L} \left(\varphi_{i,\ell}^{(1)} - \bar{\varphi}_{i^*,\ell}^{(1)}\right) \left(\varphi_{j,\ell}^{(2)} - \bar{\varphi}_{j^*,\ell}^{(2)}\right) \left(\varphi_{\ell}^{(3)}(Z_t) - \bar{\varphi}_{\ell}^{(3)}(Z_{t^*})\right).$$

Thus, the error reduction from the first two dimensions, over i and j, is maintained and any endogeneity related to the third dimension can be dealt with by finding closeness in the external variable, much like is done with the usual proxies considered throughout.

C Reducing the number of estimated parameters

Analysts may be concerned with the number of parameters implied by the least squares problem (7). In practice, this equation implies a total of $N_1N_2g(N)+N_1g(N)N_3+g(N)N_2N_3$ parameters, where $g(N)$ is the number of groups in each dimension that may depend on total data size $N = \prod_n N_n$. This implies the number of fixed-effects parameters with respect to total data size is

$$\frac{g(N)\sum_{n=1}^{d}\prod_{n'\neq n} N_{n'}}{\prod_n N_n} = g(N) \cdot O\left(\frac{1}{\min_{n\in\{1,\ldots,d\}} N_n}\right) \tag{A.7}$$

Hence, in the linear setting, the loss of degrees of freedom is negligible as long as the group size $g(N)$ does not grow too fast with respect to total data size. However, this makes estimation in non-linear settings like (3) problematic because of the incidental parameter bias, see in Chen, Fernández-Val and Weidner (2021). For this reason it is useful to consider versions of the within-cluster transformation that do not require so many parameters. The following is a non-exhaustive list of methods to reduce the number of estimated parameters.

The first approach to consider is to simply ensure the group sizes are small with respect to data size. To do this consider $g_n := g(N_n)$ as the number of groups in dimension n. Take a similar calculation to (A.7) to obtain the total number of parameters $\sum_n g_n \prod_{n' \neq n} N_{n'}$. It should then be clear that as long as $g_n = o(N_n)$, the number of estimated parameters is small with respect to total data size and the incidental parameter problem is asymptotically negligible. However, the condition that $g_n = o(N_n)$ may be highly restrictive. For example, if a sample of the unobserved parameter space is very disparate then this condition restricts the analyst to make poor approximations of the fixed-effects terms as each $\varphi^{(n^*)}_{i_{n^*},\ell} - \varphi^{(n^*)}_{j_{n^*}(i_{n^*}),\ell}$ will be very large. This is why it is important to consider the alternatives provided below. As can be seen in (8), the approximation error is multiplicative across dimensions, which means the analyst needs only to approximate a subset of these well. This fact is utilised in the below displays.

Consider clusters along just one dimension. The within-cluster transformation associated with this is simply,

$$\widetilde{A}_{ijt} = A_{ijt} - A_{i^*jt} = \sum_{\ell=1}^{L} (\varphi^{(1)}_{i\ell} - \bar{\varphi}^{(1)}_{i^*\ell}) \varphi^{(2)}_{j\ell} \varphi^{(3)}_{t\ell}.$$

Under some high-level assumptions on the unobserved fixed-effects, $\bar{\varphi}^{(1)}_{i^*\ell} = \varphi^{(1)}_{i\ell} + O\left(\frac{1}{N_1}\right)$. Also, the term $\bar{\varphi}^{(1)}_{i^*\ell}$ may have to be estimated - call the estimate $\widehat{\varphi}^{(1)}_{i^*\ell}$. Again, under some high-level assumptions, this could be estimated as $\widehat{\varphi}^{(1)}_{i^*\ell} = \bar{\varphi}^{(1)}_{i^*\ell} + O_p\left(\frac{1}{\sqrt{N_2 N_3}}\right)$. Combining this leaves the estimated $\widetilde{A}_{ijt} = O_p\left(\frac{1}{\min\{N_1, \sqrt{N_2 N_3}\}}\right)$. So selection of which dimension, d^*, to cluster and difference over solves the optimisation $d^* = \text{argmax}_{d \in \{1,2,3\}} \min\{N_d, \prod_{n,m \neq d; n \neq m} \sqrt{N_n N_m}\}$. This procedure requires $N_{n \neq d^*} N_{m \notin \{d^*, n\}} \times g(d^*)$ parameters to estimate, where $g(d^*)$ is the number of groups for dimension d^*. Of course, choice of d^* may also incorporate the number of parameters required for estimation. Note that this method does not automatically project the additive terms from $\mathcal{B}$, so this should be performed after an initial within projection.

This logic can be extended to a difference across two-dimensions as,

$$\widetilde{A}_{ijt} = A_{ijt} - A_{i^*j^*t} = \sum_{\ell=1}^{L} \left(\varphi^{(1)}_{i\ell} \varphi^{(2)}_{j\ell} - \bar{\varphi}^{(1)}_{i^*\ell} \bar{\varphi}^{(2)}_{j^*\ell}\right) \varphi^{(3)}_{t\ell}.$$

By the same reasoning as above this leads to

$$\widehat{A}_{ijt} = O_p\left(\left(\min_{d\in\{1,2\}} \min\left\{N_d, \sqrt{N_{\{1,2\}\setminus d}N_3}\right\}\right)^{-1}\right)$$

The optimal dimensions to cluster and difference on is

$$\{d_1^*, d_2^*\} = \underset{d_1,d_2}{\operatorname{argmax}} \min_{d\in\{d_1,d_2\}} \min\left\{N_d, \sqrt{N_{\{d_1,d_2\}\setminus d}N_{n\notin\{d_1,d_2\}}}\right\}.$$

This requires $g(d_1^*)g(d_2^*) \times N_{n\notin\{d_1^*,d_2^*\}}$ parameters.

Take a further difference to obtain

$$\widetilde{A}_{ijt} = (A_{ijt} - A_{i^*j^*t}) - (A_{ijt^*} - A_{i^*j^*t^*}) = \sum_{\ell=1}^{L}(\varphi_{i\ell}^{(1)}\varphi_{j\ell}^{(2)} - \bar{\varphi}_{i^*\ell}^{(1)}\bar{\varphi}_{j^*\ell}^{(2)})(\varphi_{t\ell}^{(3)} - \bar{\varphi}_{t^*\ell}^{(3)}).$$

This reduces to

$$\widetilde{A}_{ijt} = O_p\left(\left(\min_{d\in\{1,2\}} \min\left\{N_d, \sqrt{N_{\{1,2\}\setminus d}N_3}\right\} \min\left\{N_3, \sqrt{N1N2}\right\}\right)^{-1}\right),$$

which is smaller than the two cluster difference. d^* can be found similarly. This requires $g(d_1^*)g(d_2^*) \times N_{n\notin\{d_1^*,d_2^*\}} + N_{n\notin\{d_1^*,d_2^*\}} \min_{m\in\{d_1^*,d_2^*\}} N_m g(d_{n\neq m}^*)$.

The above parameter reduction exercises and specifically the choice of which dimension(s) to cluster on are also subject to the proxies used for clustering. For example, along some dimensions there may exist observable characteristics that provide a good signal of individual unobserved fixed-effect cluster. Diagnostics discussed in Section 3.1 also uncover which dimension exhibits low-rank variation, making it a good candidate for single dimension clustering. The d^*'s above are given as guides in applications where there is no obvious dimension to concentrate on when parameter reduction is required. It should also be clear that more estimated parameters can lead to tighter asymptotic rates of decay in the unobserved remainder term, which becomes obvious in the asymptotic results discussed later. One last consideration when choosing from these reduced parameter options is the implication on the additive fixed-effects terms, where not all additive terms are automatically projected with each of these reduction methods.