Linear multidimensional regression with interactive fixed-effects *

Hugo Freeman

August 27, 2024

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 control unobserved heterogeneity. First, the model is embedded within the standard two-dimensional panel framework and restrictions are formed under which the factor structure methods in [Bai (2009)] lead to consistent estimation of model parameters, but at slow rates of convergence. The second approach develops a kernel weighted fixed-effects method that is more robust to the multidimensional nature of the problem and can achieve the parametric rate of consistency under certain conditions. Theoretical results and simulations show some benefits to standard two-dimensional panel methods when the structure of the interactive fixed-effect term is known, but also highlight how the kernel weighted method performs well without knowledge of this structure. The methods are implemented to estimate the demand elasticity for beer.

1 Introduction

Models of multidimensional data – data with more than two dimensions – are becoming increasingly popular in econometric analysis as large data sets with a multidimensional structure become available. For example, consider studying demand elasticities with

*I would like to greatly thank Martin Weidner. I would also like to thank Lars Nesheim. Additionally, I would like to thank Ivan Fernández-Val, Andrei Zeleneev, Tim Christensen, and Matthew Read for their helpful suggestions along with the audience at IPDC 2022, IAAE Conference 2022, the Bristol Econometrics Study Group 2022, the MEG 2022/2023, and the NASMES 2024. This research was supported by the European Research Council grant ERC-2018-CoG-819086-PANEDA.
consumption data that varies by product, $i$, store, $j$, repeated over time, $t$. In this context, analysts may be concerned with shifts in taste preferences unobserved by the econometrician related to unobserved characteristics in each dimension - like a sporting event that impacts prices and demand heterogeneously over product, store, and time. Whilst these are not explicitly observed by the econometrician, fixed-effects are a useful tool to infer unobserved heterogeneity and control for such variation. Additive fixed-effects in multidimensional data can at most accommodate variation in unobserved heterogeneity over a subset of dimensions with any of the fixed-effects terms. For example, in the three-dimensional model additive effects only control for variation over $ij$, $it$ and $jt$, but not jointly over all $ijt$. Hence, if heterogeneity is across all dimensions interactively, additive fixed-effects cannot sufficiently control for unobserved heterogeneity. This paper develops tools to control for unobserved heterogeneity in the form of interactive fixed-effects that controls for variation that interacts over all dimensions of the data.

Consider $\beta$ estimation in the interactive fixed-effects model with three dimensions:

$$Y_{ijt} = X'_{ijt}\beta + \sum_{r=1}^{L} \varphi_{i}^{(r)} \varphi_{j}^{(r)} \varphi_{t}^{(r)} + \varepsilon_{ijt},$$

where all terms in $\sum_{r=1}^{L} \varphi_{i}^{(r)} \varphi_{j}^{(r)} \varphi_{t}^{(r)}$ are unobserved and $L$ is bounded and fixed. The object $L$ is often difficult to calculate, see Håstad (1989). However, for purposes of this analysis it is only necessary to know what is called the multilinear rank, defined in Section 2.1. The multilinear rank is the set of matrix-ranks of the different matrices the array can be transformed into along each indices. For example, the multilinear rank of the interactive fixed-effects in three dimensions is a vector of three different ranks - the first being the rank when the first dimensions are the rows, and the second and third dimensions are jointly the columns, and so on for the other multilinear rank entries. In the presence of covariates, estimation of interactive fixed-effect rank is still an open area of research in the standard panel setting. Simulations suggest estimation with more than the required components performs well. Reducing the problem to three dimensions is without loss of generality for the methods considered herein. Additive fixed-effects are omitted for brevity but are subsumed by the interactive fixed-effect term or can be removed with a simple within transformation. Let $X_{ijt}$ be arbitrarily correlated with the unobserved interactive fixed-effects term, $\sum_{r=1}^{L} \varphi_{i}^{(r)} \varphi_{j}^{(r)} \varphi_{t}^{(r)}$, but uncorrelated with the noise term, $\varepsilon_{ijt}$. The challenge to estimating $\beta$ is isolating variation in $X_{ijt}$ that is not correlated with the

---

1A non-exhaustive list of related examples can be found in the introduction of Matyas (2017).
interactive fixed-effects term. This paper develops the multidimensional kernel weighted transformations to project out this unobserved heterogeneity and also shows settings where standard factor methods work well. The kernel weighted within transformation is a novel contribution. Group fixed-effects from Bonhomme, Lamadon and Manresa (2022), and in Freeman and Weidner (2023) are also possible.

This paper makes two main contributions to the literature. The first is to show that the three or higher dimensional model can be couched in a standard two-dimensional panel data model, and to derive sufficient conditions for consistency using factor model methods from Bai (2009); Moon and Weidner (2015), albeit at slow rates of convergence. The second contribution is to introduce kernel weighted fixed-effects methods. The asymptotic results show that under certain conditions the kernel weighted fixed-effects can achieve the parametric rate of convergence, which are yet to be proven in the three dimensional case using existing methods. The simulation results corroborate these theoretical findings and an empirical application that estimates the demand elasticity of beer demonstrates how these methods work in practice.

The novel estimators proposed in this paper can be described as extensions to the usual within transformation. Consider additive fixed-effects of the form \( a_{ij} + b_{it} + c_{jt} \).

These can be projected out using the transformation,

\[
\hat{Y}_{ijt} = Y_{ijt} - \bar{Y}_{jt} - \bar{Y}_{it} + \bar{Y}_{..} - \bar{Y}_{ij},
\]

applied equivalently to \( X_{ijt} \), where the bar variables denote the average taken over the “dotted” index for the entire sample. That is, \( \bar{Y}_{jt} := \frac{1}{N_t} \sum_{i=1}^{N_i} Y_{ijt} \), \( \bar{Y}_{..} := \frac{1}{N_1 N_2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} Y_{ijt} \), etc. where means are taken with equal weights over all observations. In the presence of interactive fixed-effects, the simple within transformation demeans each fixed-effect term to leave,

\[
\sum_{\ell=1}^{L} \left( \varphi_{i\ell}^{(1)} - \bar{\varphi}_{\ell}^{(1)} \right) \left( \varphi_{j\ell}^{(2)} - \bar{\varphi}_{\ell}^{(2)} \right) \left( \varphi_{t\ell}^{(3)} - \bar{\varphi}_{\ell}^{(3)} \right)
\]

which is clearly not controlled for if the \( \varphi \) terms admit heterogeneity across observations.

Consider instead an extension of the within transformation that uses weighted means instead of uniform means. With an abuse of notation, consider,

\[
\bar{Y}_{ijt} = Y_{ijt} - \bar{Y}_{jt} - \bar{Y}_{it} + \bar{Y}_{..} - \bar{Y}_{ij},
\]

where the bar variables combined with the star indices denote weighted means for that observation. For example, \( \bar{Y}_{i'jt} = \sum_{i'} w_{i'i'} Y_{i'jt} \) for weights across \( i' \) for each \( i \). In turn,
the remainder from the interactive fixed-effects term is,

$$\sum_{\ell=1}^{L} \left( \varphi_{it}^{(1)} - \sum_{i'} w_{i,i'} \varphi_{it}^{(1)} \right) \left( \varphi_{j\ell}^{(2)} - \sum_{j'} w_{j,j'} \varphi_{j\ell}^{(2)} \right) \left( \varphi_{t\ell}^{(3)} - \sum_{t'} w_{t,t'} \varphi_{t\ell}^{(3)} \right).$$

If appropriate weights are used, the weighted within transformation can project out the interactive fixed-effects. The same can be true if the weighted means are replaced with cluster means for appropriate cluster assignments, similar to a group fixed-effects estimator. Hence, with a relatively small change to how the within transformation is performed, much more general fixed-effects can be considered in the linear model.

The model for interactive fixed-effects has precedent in the standard two-dimensional panel data setting, for example in Bai (2009); Pesaran (2006),

$$Y_{it} = X_{it}' \beta + \sum_{\ell=1}^{L} \lambda_{i\ell} f_{it} + e_{it}.\quad (4)$$

The interactive term $$\sum_{\ell=1}^{L} \lambda_{i\ell} f_{it}$$ also sufficiently captures variation in additive individual and time effects without the need to specify these separately. For multidimensional applications the problem in (1) can be transformed to a two dimensional problem and estimated as (4) directly using the transformed data. Indeed, Section 3.1 displays useful preliminary convergence rates for this approach. However, convergence rates can suffer severely from the over-parametisation implied by (4). Without strong assumptions on the sparsity of the fixed-effects, only a slow rate of convergence can be guaranteed for this approach. However, when this approach is used to construct preliminary estimators of the fixed-effects, the parametric rate of consistency is possible when these are combined with the novel weighted within transformations. Kuersteiner and Prucha (2020) consider a similar dimension specific projection in the time dimension in the presence of error correlations, in particular see Supplement D.5.2.

On top of this slow convergence rate for the two dimensional transformation, finite sample bias may also arise when $$L$$ is large and only a subset of the unobserved heterogeneity parameters are low-dimensional. For unbiased estimates of $$\beta$$, transforming the multidimensional array to a matrix then estimating (4) requires either: (a) all the fixed-effects are low-dimensional; or (b) that a known subset of the fixed effects are low-dimensional, which is a strong assumption. Alternatively, whilst the kernel weighted transformations also require that a subset of the fixed-effect parameters are low-dimensional, the analyst does not need to know which ones are. In this sense, the novel kernel weighted estimator
is robust. A concrete example is considered in a simulation exercise, and there is evidence in the empirical application of differences in dimensionality of each fixed-effect.

The beer demand elasticity application uses Dominick’s supermarket data for Chicago, 1991-1995, where price and quantity vary over product, store, and fortnight. For comparison, barley price is used as an instrument to estimate elasticities. IV estimates show demand is strongly negatively elastic (-2.15), however, estimates are very imprecise. This could be because the instrument only varies over one dimension, time, or, e.g., because prices of other inputs are also highly variable so the instrument does not explain much price variation. Estimates from the weighted within transformation also show demand is downward sloping and elastic (-2.83), but with much better precision. Factor model estimates are highly sensitive to how the data is transformed into two-dimensions. The estimates from the weighted-within transformation are similar to the own-price elasticities estimated in Hausman, Leonard and Zona (1994).

The technical component of this paper is related to the numerical analysis literature on low-rank approximations of multidimensional arrays. As pointed out in De Silva and Lim (2008), the low-rank approximation problem in the tensor setting is not well-posed, hence poses technical difficulties. 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). Therefore, 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 two-dimensional sub-problems for use in nuisance parameter applications. This application has the advantage that they only require the fixed-effects to be projected out at a sufficient asymptotic rate, and do not attempt to directly solve the low-rank tensor problem.

These methods can also alleviate convergence issues in two-dimensional panel data models when \( N \) and \( T \) are not proportional. Interactive fixed effects estimation usually requires \( N, T \) grow at the same rate. Consider splitting the larger of the \( N \) or \( T \) dimension into sub-dimensions. For example, if \( N >> T \), then split \( N \) into as many sub-dimensions as necessary to make the dimensions proportional. If \( N = O(T^\alpha) \) for \( \alpha \in (1, M) \), then

---

5 Elden and Savas (2011), along with related papers, suggest a reformulation of the low multilinear rank problem that may have promising applications in econometrics.
split $N$ into roughly $\alpha$ equal parts. For example, if $\alpha = 2$,

$$\sum_{\ell=1}^{L} \lambda_{ij\ell} f_{i\ell} = \sum_{\ell=1}^{L} \left( \sum_{\ell' =1}^{L'} \lambda^{(1)}_{i\ell'\ell} \lambda^{(2)}_{j\ell'\ell} \right) f_{i\ell},$$

where $\lambda^{(1)}_{i\ell'\ell}$ and $\lambda^{(2)}_{j\ell'\ell}$ denote the superimposed fixed-effect for the new sub-dimensions. If there is sufficient sparsity in the first dimension fixed-effect parameter, $\lambda_{ij\ell}$, then $L'$ can be bounded and fixed, and the interactive fixed-effect term is synonymous with the multidimensional interactive fixed-effects in (1).

The paper is organised as follows: Section 2 introduces the model; Section 3 details the estimators and convergence results; Section 4 discusses the convergence results and some practical considerations; Section 5 displays the simulation results; Section 6 shows the beer demand estimation empirical application; and Section 7 concludes.

2 Model

Let $\beta^0$ denote the true parameter value for the slope coefficients. The model in full dimensional generality is,

$$Y = \sum_{k=1}^{K} X_k \beta^0_k + A + \epsilon,$$

where $Y, X_k, \epsilon \in \mathbb{R}^{N_1 \times N_2 \times \ldots \times N_d}$. $A = \sum_{\ell=1}^{L} \varphi^{(1)}_{\ell} \circ \ldots \circ \varphi^{(d)}_{\ell}$ where $\varphi^{(n)}_{\ell} \in \mathbb{R}^{N_n}$ for each $n = 1, \ldots, d$ and “$\circ$” is the outer product. $L$ is bounded and fixed. $\epsilon$ is a noise term uncorrelated with all $X_k$ 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, where $N_n$ is the sample size of dimension $n$. The regressors $X_k$ may be arbitrarily correlated with $A$. Throughout this paper all dimensions are considered to grow asymptotically, that is $N_n \to \infty$ for all $n$, however, for the asymptotic theory only a subset of two dimensions need to grow asymptotically.

Model (5) can be seen as a natural extension of the Bai (2009) model to three (or more) dimensions with $A$ interpreted as a “higher-dimensional” factor structure. Similar

---

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

$$Y_{i_1,i_2,\ldots,i_d} = \sum_{k=1}^{K} X_{i_1,i_2,\ldots,i_d; k} \beta^0_k + A_{i_1,i_2,\ldots,i_d} + \epsilon_{i_1,i_2,\ldots,i_d}$$

with $A_{i_1,i_2,\ldots,i_d} = \sum_{\ell=1}^{L} \varphi^{(1)}_{i_1\ell} \ldots \varphi^{(d)}_{i_d\ell}$.
to this strand of the literature, all terms in \( \mathbf{A} \) 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.

\( \mathbf{A} \) incorporates additive fixed effects that vary in any strict subset of dimensions. For example, in the three dimensional setting one may want to control for the additive terms, \( a_{ij} + b_{it} + c_{jt} \). These can be controlled for using \( L = \min\{N_1, N_2\} + \min\{N_1, N_3\} + \min\{N_2, N_3\} \), with the first \( \min\{N_1, N_2\} \) terms \( \sum_{\ell=1}^{\min\{N_1, N_2\}} \varphi_{i\ell}^{(1)} \varphi_{j\ell}^{(2)} = a_{ij} \) by setting \( \varphi_{i\ell}^{(3)} = 1 \) for \( \ell = 1, \ldots, \min\{N_1, N_2\} \), and so on for the \( b_{it} \) and \( c_{jt} \). These could also be controlled for directly using the standard within-transformation before considering the model in (5).

This paper comprises of two main modelling approaches. The first is to embed the multidimensional model into a standard panel data model by simply flattening all arrays into matrices. The second approach uses weighted differences across each dimension to reduce each \( \varphi^{(n)} \) component of \( \mathbf{A} \) separately for each \( n \). For this reason the model assumptions are split out and stated in Section 3 alongside each estimation approach.

### 2.1 Notation and preliminaries

For a \( d \)-order tensor, \( \mathbf{A} \), a factor-\( n \) flattening, denoted as \( \mathbf{A}_{(n)} \), is the rearrangement of the tensor into a matrix with dimension \( n \) varying along the rows and the remaining dimensions simultaneously varying over the columns. That is, \( \mathbf{A}_{(n)} \in \mathbb{R}^{N_n \times N_{n+1}N_{n+2} \ldots N_1} \).

The Frobenius norm, \( \| \cdot \|_F \), of a matrix or tensor is the entry-wise norm, \( \| \mathbf{A} \|_F^2 = \sum_{i=1}^{N_1} \cdots \sum_{i_d=1}^{N_d} A_{i_1 \ldots i_d}^2 \). The spectral norm, denoted \( \| \cdot \| \), is the largest singular value of a matrix. For a \( d \)-order tensor, \( \mathbf{A} \), the multilinear rank, denoted \( \mathbf{r} \), is a vector of matrix ranks after factor-\( n \) flattening in each dimension, with each component of this vector \( r_{n} = \text{rank}(\mathbf{A}_{(n)}) \). Tensor rank, different to multilinear rank, is defined as the least number of outer products of vectors to replicate the tensor. That is, for tensor \( \mathbf{A} \) and vectors \( u_{\ell}^{(n)} \in \mathbb{R}^{N_n} \), tensor rank is the smallest \( L \) such that \( \mathbf{A} = \sum_{\ell=1}^{L} u_{\ell}^{(1)} \circ \cdots \circ u_{\ell}^{(d)} \), where \( \circ \) is the outer product of a vector. The notation \( a \preceq b \) means the asymptotic order of \( a \) is bounded by the asymptotic order of \( b \). The \( n \)-mode product of a tensor \( \mathbf{A} \) and matrix \( B \) is denoted \( \mathbf{A} \times_n B \) and has elements

\[
(\mathbf{A} \times_n B)_{i_1,\ldots,i_d} = \sum_{i_n=1}^{N_n} A_{i_1,\ldots,i_n,\ldots,i_d} B_{j,i_n},
\]

which is equivalent to saying the flattening \( (\mathbf{A} \times_n B)_{(n)} = \mathbf{B} \mathbf{A}_{(n)} \).
The singular value decomposition of a matrix, \( A \in \mathbb{R}^{N_1 \times N_2} \) is

\[
A = U \Sigma V' = \sum_{r=1}^{\min\{N_1, N_2\}} \sigma_r u_r v_r',
\]

where \( U \) is the matrix of left singular vectors, \( u_r \), \( V \) is the matrix of right singular vectors, \( v_r \), and \( \Sigma \) is a diagonal matrix of singular values, \( \sigma_r \), with values running in descending order down the diagonal. For a rank-\( r \) matrix, the first \( r \) entries on the diagonal of \( \Sigma \) are strictly positive and the remaining entries are zero.

Take the approximation problem,

\[
\min_{A'} \|A - A'\|_F \text{ such that } \text{rank}(A') = k.
\]

It is well known from the Eckart-Young-Mirsky theorem that the solution to this approximation problem is the first \( k \) terms of the singular value decomposition, i.e. \( \sum_{r=1}^{k} \sigma_r u_r v_r' \).

The Eckart-Young-Mirsky theorem effectively picks out the row and column subspaces that best explain variation in the matrix \( A \) as the leading columns of the matrix \( U \), respectively of \( V \). The sum of squared error at the minimiser is thus \( \sum_{r=k+1}^{\min\{N_1, N_2\}} \sigma_r^2 \). This is commonly called a low-rank approximation and forms the cornerstone for estimation of unobserved heterogeneity in the factor model and interactive fixed-effects models in Bai and Ng (2002); Bai (2009); Moon and Weidner (2015) amongst others.

The Eckart-Young-Mirsky theorem, however, does not extend to the three or higher dimensional setting, see De Silva and Lim (2008) for details. To avoid this complication, the multidimensional problem can be translated to the two-dimensional setting to utilise the Eckart-Young-Mirsky theorem, or the fixed-effects parameters need to be shrunk separately, as is done with the weighted-within transformation.

### 3 Estimation

This section details the two estimation approaches used. The first subsection details how to apply standard two dimensional estimators to the problem and consistency results. The second subsection detail the weighted-within transformation approach and consistency result. The third subsection describes an inference procedure.
3.1 Matrix low-rank approximation estimator

This section provides a description of some matrix methods that can be applied directly to the multidimensional model and stipulates the assumptions required for consistency. Kapetanios, Serlenga and Shin (2021) employ a similar approach for three-dimensional arrays in conjunction with the Pesaran (2006) common correlated effects estimator. Babii, Ghysels and Pan (2022) employ a similar matricisation procedure as that detailed below, but are interested in inference on the fixed-effect parameters.

Consider recasting the multidimensional array problem into a two dimensional panel problem by flattening \( Y \) and \( X \) in the \( n \)-th dimension, \( Y(n) = X'(n)\beta^0 + \varphi(n)\Gamma'_n + \varepsilon(n) \)

where \( Y(n), X(n), \varepsilon(n) \in \mathbb{R}^{N_n \times \prod_{n' \neq n} N_{n'}} \), \( \varphi(n) \) is an \( N_n \times r_n \) matrix and \( \Gamma_n \) is an \( \prod_{n' \neq n} N_{n'} \times r_n \) matrix that accounts for variation in the remaining \( \varphi(n') \) for all \( n' \neq n \). The term \( r_n \) is indexed by the dimension \( n \) because it may vary non-trivially according to the flattened dimension. This is then the model described in (4), that is, the standard linear model with factor structure unobserved heterogeneity as studied in Bai (2009).

The two-dimensional fixed-effects estimator for a given flattening, \( n \), optimises

\[
R(\beta, \hat{r}_n, n) = \min_{\varphi(n) \in \mathbb{R}^{N_n \times \prod_{n' \neq n} N_{n'}}, \Gamma_n \in \mathbb{R}^{\prod_{n' \neq n} N_{n'} \times r_n}} \left\| Y(n) - X'(n)\beta - \varphi(n)\Gamma'_n \right\|_F^2.
\]

Then \( \hat{\beta}^{2D}_{(n)} = \arg\min_{\beta} R(\beta, \hat{r}_n, n) \) is the slope estimate for the two-dimensional setup. The analyst must choose both the dimension to flatten in, \( n \), and the rank of the estimated interactive fixed-effects term, \( \hat{r}_n \). It is well known that the minimum in (8) is achieved using the leading \( \hat{r}_n \) terms from the singular value decomposition of the error term, \( Y(n) - X'(n)\beta \). This gives \( \hat{\varphi}^{(n)} \) as the first \( \hat{r}_n \) columns of \( \hat{U}\hat{\Sigma} \) and \( \hat{\Gamma}_n \) as the first \( \hat{r}_n \) columns of \( \hat{V} \) where \( \hat{U}, \hat{\Sigma} \) and \( \hat{V} \) are the terms from (6) of the singular value decomposition of \( Y(n) - X'(n)\beta \). Because this error term is a function of \( \beta \), an iteration is required between estimating \( \beta \) and finding the singular value decomposition of the error term. This is a well studied iteration; for details see Bai (2009) or Moon and Weidner (2015).

In the following assumptions let \( \hat{r}_n \) be the estimated number of factors for the \( (n) \)-flattening of the regression line when applying the least square methods in (8). Also, let \( \mathcal{L} \subset \{1, \ldots, d\} \) be a non-empty subset of the dimensions. In the following, the multilinear rank of \( \mathcal{A} \) is restricted such that it is low-rank along at least one of the flattenings.
Assumption 1 (Bounded norms of covariates and exogenous error).

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

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

Assumption 2 (Weak exogeneity). \( \text{vec}(X_k)'\text{vec}(\varepsilon) = O_p \left( \prod_{n=1}^d \sqrt{N_n} \right) \) for each \( k \)

Assumption 3 (Low multilinear rank). For some positive integer, \( c, r_{n^*} < c \) for all \( n^* \in \mathcal{L} \), where \( r_n \) is the \( n^{th} \) component of the multilinear rank of \( \mathcal{A} \).

Assumption 4 (Non-singularity). Let \( \sigma_s(A) \) be the \( s^{th} \) singular value for a matrix \( A \). Consider linear combinations \( \delta_{n^*} \cdot X_{(n^*)} = \sum_k \delta_{n^*,k} X_{(n^*)} \). For each dimension \( n^* \in \mathcal{L} \) that satisfies Assumption 3, then for \( K \times 1 \) unit vector \( \delta_{n^*} \),

\[
\min_{\{\delta_{n^*} \in \mathbb{R}^K, \|\delta_{n^*}\| = 1\}} \sum_{s=r_{n^*}+r_{n^*}+1}^{\min\{N_{n^*}, \prod_{m \neq n^*} N_m\}} \sigma^2_s \left( \frac{\delta_{n^*} \cdot X_{(n^*)}}{\prod_n \sqrt{N_n}} \right) > b > 0 \quad \text{wpa} 1.
\]

Assumptions 1, 2 and 4 are standard regularity assumptions already well established in the literature, e.g. see [Moon and Weidner (2015)]. Assumption 1.(i) ensures that the covariates have bounded norms, for example having bounded second moments. Assumption 1.(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 2 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 4 states that, after factor projection, the set of covariates still collectively admit full-rank variation.

Assumption 3 is new and asserts that there exists at least one flattening of the interactive term, \( \mathcal{A} \), that is low-dimensional or simply low-rank. This requires that at least one of the unobserved terms \( \varphi^{(n)} \) is low dimensional. Note that not all dimensions must satisfy Assumption 3 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 \( \widehat{\beta}^{2D}_{(n^*)} \) be the estimator from [Bai (2009)] after first flattening along dimension \( n^* \in \mathcal{L} \). If Assumptions 2, 3, 4 hold, the subset \( \mathcal{L} \) is non-empty, and the estimated
number of factors $\hat{r}_{n^*} \geq r_{n^*}$, then, for each $n^* \in \mathcal{L}$ satisfying Assumption 3,

$$\left\| \tilde{\beta}^{2D}_{(n^*)} - \beta^0 \right\| = O_p \left( \frac{1}{\sqrt{\min\{N_{n^*}, \prod_{n \neq n^*} N_n\}}} \right).$$

(9)

Proposition 1 follows directly from Moon and Weidner (2015) since the flattening procedure reduces the problem to the standard linear interactive fixed-effects model. This result only applies to estimates in the dimension(s) that satisfy the low-rank assumption in Assumption 3, i.e., the analyst has chosen the correct dimension to flatten. The constraint $b_r n^* \geq r_{n^*}$ can also be changed to $b_r n^* \geq c$; however, this is more conservative than required for the statement of the result. This constraint does not require knowledge of $r_{n^*}$, just that the number of estimated factors is greater than or equal the true number. The estimation procedure from Proposition 1 can also be augmented to flatten over multiple indices, but makes Assumption 3 harder to satisfy. For example, take the tensor $\mathcal{A}$ flattened over the first two indices as $\mathcal{A}_{(1,2)} \in \mathbb{R}^{N_1 N_2 \times \prod_{n \in \{1,2\}} N_n}$. If the parameters $\varphi^{(n)}$ for $n = 3, \ldots, d$ are high-dimensional, Assumption 3 is only satisfied when both $\varphi^{(1)}$ and $\varphi^{(2)}$ and their product space is low-dimensional. However, flattening along multiple dimensions can improve the convergence rate in Proposition 1 to $O_p \left( \min\{N_1 N_2, \prod_{n \notin \{1,2\}} N_n\} \right)^{-1/2}$.

3.2 Kernel weighted fixed-effects

Let $\tilde{\varphi}_{i_n}^{(n)} \in \tilde{\Phi}_n$ generically denote a known or estimated proxy for fixed-effect $\varphi_{i_n}^{(n)}$. For an example of when this is known, the analyst may have access to exogenous variables that meaningfully explain unobserved heterogeneity in dimension $n$. The use of this notation will become clear in the statement of Proposition 2 and in discussion of how to estimate these proxy measures in Section 4.2. Let $\mathcal{W}$ be an ordered set of weight matrices, where the $n^*$th item $W_{n} \in \mathbb{R}^{N_n \times N_n}$ has elements,

$$W_{n,i_n j_n} := k \left( \frac{1}{h_n} \left\| \tilde{\varphi}_{i_n}^{(n)} - \varphi_{j_n}^{(n)} \right\| \right),$$

(10)

where $k$ is a kernel function, and $h_n$ is a bandwidth parameter. The weighted-within transformation in (3) can be generalised with the following series of $n$-mode products, $Y \times M_1 \times M_2 \times \cdots \times M_d$, likewise also on each $X_k$, where $M_n = \mathbb{I}_{N_n} - W_{n}$ and $\times_n$ is the $n$-mode product defined in Section 2.1. This projection may not optimise any specific objective function - it is designed to sufficiently project out variation in the fixed-effect.
term. In this sense, the following theoretical analysis is viewed in the light of an asymptotic bias problem. Let \( \hat{\beta}_{KER} \) be the pooled OLS estimator after using weights \( W \) in the weighed-within transformation on each observed variable. All technical assumptions and results in this section relate to this estimator.

For the remainder of the paper, in particular Section 3.3 for inference, weight proxies are estimated from Section 3.1 matrix method estimates, see Section 4.2. Alternatively, if proxies are estimated directly from the error term \( Y - \sum_{k=1}^{K} X_k \hat{\beta}_k \), then there is an iteration between slope estimation and estimation of kernel weights. This can also be optimised over a grid space for \( \beta \) if \( \beta \) is low-dimensional.

**Assumption 5 (Kernels).** The kernel function, \( k(\cdot) \) is, (i). Non-negative, non-increasing, (ii). \( \int k(u)du = 1 \) (iii). \( \int u^2 k(u)du < \infty \) (iv). \( k(u) \lesssim u^{-\delta} \) for \( \delta > 2 \).

Assumption 5(i)-(iii) are standard kernel function restrictions. Assumption 5(iv) regularises the kernel function to decay sufficiently quickly. This is a weak condition on the decay, and can be satisfied very easily with, e.g., \( k(u) = \exp(-u^2) \) or \( k(u) = 1\{u < 1\} \).

**Assumption 6 (Regularity of proxies).** Let \( \hat{\varphi}_{in}^{(n)} \in \hat{\Phi}_n \) and \( B_{he}(x) \) be an he-neighbourhood around \( x \). Then, each dimension \( n \), for any \( e > 0 \), as \( h \to 0 \), such that \( N_nh \to \infty \),

\[
\frac{1}{N_n} \sum_{i_n=1}^{N_n} \sum_{i'_n \neq i_n} 1(\hat{\varphi}_{i'_n}^{(n)} \in B_{he}(\hat{\varphi}_{in}^{(n)})) > 0 \quad \text{wpa1, (11)}
\]

Equation (11) ensures that there are order \( N_n \) observations such that \( W_{i_n,i'_n} \) do not degenerate to 1 when \( i_n = i'_n \). Compact support over the distribution of \( \hat{\varphi}_{in}^{(n)} \) would satisfy the assumption, but is not necessary. For example, whilst the condition is not satisfied for \( h \to 0 \) if \( \hat{\varphi}_{in}^{(n)} = i_n \); it can be that \( \hat{\varphi}_{in}^{(n)} = i_n/N_n^\rho \), and the condition is satisfied for \( N_n^\rho h \to \infty \), which restricts the speed of decay \( h \to 0 \). The restriction in Assumption 6 on the proxies can be used to show a higher-level condition on the sampling of the kernel function, see Lemma A.1 in the Appendix. The condition in Assumption 6 ensures that, asymptotically, neighbourhoods around each proxy are on average non-empty, and avoids the projection \( (\mathbb{I}_{N_n} - W_n) \) degenerating to 0 for too many observations\(^4\).

Assumption 6 is stated for vector-valued \( \hat{\varphi}_{in}^{(n)} \in \hat{\Phi}_n \), which presents a curse of dimensionality, since it scales poorly with \( \text{dim}(\hat{\varphi}_{in}^{(n)}) \). However, the weighted-within transformation can be implemented iteratively, so the restriction can be relaxed to an element-wise

\(^4\)Some observations can degenerate, which amounts to removing observations according to outlier fixed-effects proxies. The condition ensures not too many observations are removed.
restriction on each component of $\varphi_{i_n}^{(n)} \in \hat{\Phi}_n$, but leads to a significantly more complicated theory. See Appendix A.3 for a treatment of the iterative estimator, which includes consistency results. Hence, Assumption 6 is viewed as a simplifying assumption.

**Assumption 7** (Regularity conditions). Let $\hat{T}_{i_1, \ldots, i_d}$ be the entries of tensor $T$ after the kernel weighted fixed-effects are differenced out. Then,

(i). \[ \left( \frac{1}{\prod_n N_n} \sum_{i_1} \cdots \sum_{i_d} \hat{X}_{i_1, \ldots, i_d} \hat{X}'_{i_1, \ldots, i_d} \right) \] converges in probability to a nonrandom positive definite matrix as $N_1, \ldots, N_d \to \infty$.

(ii). \[ \frac{1}{\prod_n N_n} \sum_{i_1} \cdots \sum_{i_d} \hat{X}_{i_1, \ldots, i_d} \varepsilon_{i_1, \ldots, i_d} = O_p \left( \frac{1}{\sqrt{\prod_n N_n}} \right). \]

Assumption 7(i) is very similar to Assumption 4 except here full rank is required after the weighted-within projection rather than the factor projection. Assumption 6 helps justify Assumption 7(i) but does not guarantee it. Assumption 7(ii) is an exogeneity condition that requires weak exogeneity in the covariates after the weighted-within transformation, which can be viewed as similar to Assumption 2. Assumption 7(ii) is required because the noise term $\varepsilon$ can foreseeably impact weights if weights are estimated from functionals of a residual term. This is alleviated by, e.g., making sure weights are based on variables extraneous to the regression line, hence independent of $\varepsilon$; or, e.g., with a sample split, such as that proposed in Freeman and Weidner (2023).

**Proposition 2** (Upper bound on kernel weighted estimator). Let Assumptions 5-7 hold. Let \[ \frac{1}{N_n} \sum_{i_n} \left\| \varphi_{i_n}^{(n)} - \hat{\varphi}_{i_n}^{(n)} \right\|^2 = O_p(C_{n}^{-2}) \text{ for } n^* \in M \text{ with } C_{n}^{-2} \to 0, \] and \[ \frac{1}{N_n} \sum_{i_{n'}} \left\| \varphi_{i_{n'}}^{(n')} - \hat{\varphi}_{i_{n'}}^{(n')} \right\|^2 = O_p(1) \text{ for } n' \notin M, \] where $M$ is a non-empty subset of dimensions. Then,

\[ \left\| \hat{\beta}_{KER,W} - \beta \right\| = \sqrt{LO_p} \left( \prod_{n^* \in M} O_p \left( C_{n^*}^{-1} \right) + O_p \left( h_{n^*} \right) \right) + O_p \left( \prod_{n=1}^d \frac{1}{\sqrt{N_n}} \right). \]

For $h_n \lesssim O(C_{n}^{-1})$ this reduces to

\[ \left\| \hat{\beta}_{KER,W} - \beta \right\| = \sqrt{LO_p} \left( \prod_{n^* \in M^{'}} O_p \left( C_{n^*}^{-1} \right) \right) + O_p \left( \prod_{n=1}^d \frac{1}{\sqrt{N_n}} \right). \]

---

5Assumption 6 weaken some conditions for Assumption 7(i) to hold, but cannot fully rule out regressor degeneracy.
Proposition 2 shows convergence for the kernel estimator is bounded by convergence of the proxy estimates. Discussions in Section 4.2 suggest $O_p(C_n^{-1})$ can be $O_p(1/\sqrt{N_n^*})$ under regularity conditions imposed in Bai (2009). This shows the parametric rate is attainable if $\mathcal{M}' = \{1, \ldots, d\}$ and $L$ is fixed and bounded. Implicit in the statement of Proposition 2 is that the correct multilinear rank of the tensor of interactive fixed-effects is known, at least for the dimensions $\mathcal{M}'$. This can likely be relaxed to the case where the upper bound on the multilinear rank is known, but this is left for further research. This would follow from the results in Moon and Weidner (2015), also used in Proposition 1.

3.2.1 Optimal kernel weighted fixed-effects

In the previous section, the proposed estimator was designed to sufficiently project out interactive fixed-effects, without explicitly optimising an objective function. As a diversion, consider an estimator that does optimise an objective function with respect to a weighted fixed-effects model, but adds significant complexity to the technical analysis. This estimator is considered for discussion, and no formal results are derived. Let $\Delta$ be a $d$-list of $\times_{n=1}^d N_n$ tensors $\delta_n \in \mathbb{R}^{N_n \times \cdots \times N_n}$. For a given set of proxy measures and kernel function, the kernel weighted fixed-effects estimator optimises

$$S(\beta, W) = \min_{\delta \in \Delta} \left\| Y - \sum_{k=1}^K X_k \beta_k - \sum_{n=1}^d \delta_n \times_n W_n' \right\|_F^2.$$  \hspace{1cm} (12)

Then, $\hat{\beta}_{KEROpt,W} := \arg\min_{\beta \in \mathbb{R}^K} S(\beta, W)$, where $KEROpt$ stands for kernel optimum. The notation $\times_n$ is the $n$-mode product defined in Section 2.1. Each of the terms in $\sum_{n=1}^d \delta_n \times_n W_n'$ can be interpreted as weighted fixed-effects. Take $M_n$ to be an $N_n \times N_n$ matrix defined as $\mathbb{I}_{N_n} - W_n (W_n' W_n)^\dagger W_n'$, where $\dagger$ is the Moore-Penrose generalised inverse. Then, the estimator can be formed by pooled OLS after the following series of $n$-mode products, $Y \times_1 M_1 \times_2 M_2 \times_3 \cdots \times_d M_d$, likewise also on each $X_k$. This sequentially differences out the weighted means from each dimension separately. Then the Frisch-Waugh-Lovell theorem straightforwardly applies.

Projection of these weighted fixed-effects is equivalent to using the weighted-within transformation in (3), similar to the previous section, but with different weights. To obtain the kernel optimum estimator, input weights $W_n (W_n' W_n)^\dagger W_n'$ to the weighted-within estimator. Using $W_n$ as per the earlier $\hat{\beta}_{KER,W}$ estimator normalises weights to $[0, 1]$, which turns out to make technical proofs much simpler. It is expected, however, that the $\hat{\beta}_{KEROpt,W}$ estimator will have similar asymptotic properties.
3.3 Asymptotic Distribution

This section establishes an inference procedure for the kernel weighted fixed-effects model. Proposition 2 establishes an upper bound on the kernel weighted fixed-effect estimator convergence rate that can be refined to exactly the parametric rate for the fixed-effect asymptotic bias component. To ensure the bias from the fixed-effect term converges sufficiently quickly a further correction is required. Below is an additional orthogonalisation as well as an asymptotic bias component. To ensure the bias from the fixed-effect term converges sufficiently quickly a further correction is required. Below is an additional orthogonalisation as well as an asymptotic bias component. To ensure the bias from the fixed-effect term converges sufficiently quickly a further correction is required.

Consider the conditional expectation, \( \mathbb{E}(Y|A) = \Gamma_Y, \mathbb{E}(X|A) = \Gamma_X \), such that,

\[
X = \Gamma_X + \eta, \quad \Gamma_Y = \Gamma_X \cdot \beta^0 + A, \quad Y - \Gamma_Y = (X - \Gamma_X) \cdot \beta^0 + \varepsilon, \tag{13}
\]

with \( \mathbb{E}(\eta|A) = 0, \mathbb{E}(\varepsilon|X, A) = 0 \). The display for \( X \) is general, in that if \( X \) is unrelated to \( A, \Gamma_X = 0 \). Hence, the display in (13) is a representation, not an imposed model. If the correlation between \( X \) and \( A \) is weak in the sense that the empirical mean of \( \Gamma_X^2 \) converges to zero at arbitrary rate, then the following inference correction is not required and the convergence result from Proposition 2 strengthens to a sufficient rate.

The (infeasible) inference corrected estimator, \( \beta_{IC}^{(infeasible)} \), is

\[
\hat{\beta}_{IC}^{(infeasible)} = (\text{vec}_K(X - \Gamma_X)' \text{vec}_K(X - \Gamma_X))^{-1} (\text{vec}_K(X - \Gamma_X)' \text{vec}(Y - \Gamma_Y)),
\]

where \( \text{vec}_K(X - \Gamma_X) \in \mathbb{R}^{\prod_n N_n \times K} \) is shorthand for the matrix of vectorised covariates, with each column the vectorised \( (X_k - \Gamma_{X_k}) \). The vec notation for \( (Y - \Gamma_Y) \) is the standard vectorisation. This is so far infeasible because \( \Gamma_X \) and \( \Gamma_Y \) need to be estimated.

Consider estimates for \( \Gamma_X \), and \( \Gamma_Y \) as \( \hat{\Gamma}_X \), and \( \hat{\Gamma}_Y \), respectively, and \( \hat{\Omega}_X = N^{-1} \text{vec}_K(X - \hat{\Gamma}_X)' \text{vec}_K(X - \hat{\Gamma}_X) \). Then, with \( \tilde{\eta} := X - \hat{\Gamma}_X \),

\[
\hat{\Omega}_X^{-1} N^{-1} \text{vec}_K(\tilde{\eta})' \text{vec}(Y - \hat{\Gamma}_Y) = \hat{\Omega}_X^{-1} N^{-1} \text{vec}_K(\tilde{\eta})' \text{vec}(X \cdot \beta^0 + A + \varepsilon - \hat{\Gamma}_X \cdot \beta - \hat{A})
\]

\[
= \beta^0 + \hat{\Omega}_X^{-1} N^{-1} B + \hat{\Omega}_X^{-1} N^{-1} \text{vec}_K(\tilde{\eta})' \text{vec}(\varepsilon), \tag{14}
\]

where \( \beta \), and \( \hat{A} \) are preliminary estimates of \( \beta^0 \), and \( A \), respectively, used to form the estimate \( \hat{\Gamma}_Y \). \( B \) is a bias term that must be sufficiently small order for standard asymptotic normality arguments. Specifically, \( \hat{\Omega}_X^{-1} N^{-1} B = o_p\left( \prod_{n=1} N_n^{-1/2} \right) \) is required. Hence, preliminary convergence rates are required on \( (\Gamma_X - \hat{\Gamma}_X), (\beta^0 - \beta), \) and \( (A - \hat{A}) \) to ensure bias is bounded appropriately. From Section 3.2, with the kernel weighted estimator,
(β₀ − ̂β) can be bounded by $O_p(\prod_n h_n)$ if $N_n^{-1/2} \lesssim h_n$, where $h_n$ is the bandwidth used for dimension $n$. Likewise, $\left(\prod_{n=1}^{N_n} \frac{1}{h_n} \mathbf{A} - \hat{\mathbf{A}}\right)' \mathbf{A}^{-1/2} = O_p(\prod_n h_n)$. The bound for $(\mathbf{X} - \hat{\mathbf{X}})$ is incidentally also useful to bound $\mathbf{vec}(\hat{\mathbf{n}})$.

Let $\frac{1}{N} \mathbf{vec}_K(\mathbf{X} - \hat{\mathbf{X}})' \mathbf{vec}_K(\mathbf{X} - \hat{\mathbf{X}}) = O_p(\xi_X^2)\mathbb{I}_K$, where $\xi_X$ is the rate of consistency for the estimator of $\hat{\mathbf{X}}$. Also, let $(β₀ − ̂β) = O_p(\xi_β)\mathbb{I}_K$, and $\frac{1}{N} \mathbf{vec}(\mathbf{A} - \hat{\mathbf{A}})' \mathbf{vec}(\mathbf{A} - \hat{\mathbf{A}}) = O_p(\xi_A^2)$. Under Assumption 8, and assumptions in Proposition 1 hold; reliance on regularity conditions it can be shown, 

$$\sqrt{N}(\hat{\beta}_{IC} - \beta_0) = \sqrt{N} \left( O_p(\xi_X\xi_β)\mathbb{I}_K + O_p(\xi_X\xi_A)\mathbb{I}_K + O_p(\xi_βN^{-1/2})\mathbb{I}_K + \Omega^{-1} \mathbf{vec}_K(\hat{\mathbf{n}})' \mathbf{vec}(\varepsilon) \right).$$

If $\xi_X\xi_β = o_p(N^{-1/2})$, $\xi_β → 0$, and $\xi_X\xi_A = o_p(N^{-1/2})$, asymptotic bias is of order $o_p(N^{-1/2})$. Proposition 2 establishes the upper bound on convergence for $\xi_β$ and $\xi_A$ can be $O_p(N^{-1/2})$, such that for any $\xi_X = o_p(1)$, the $o_p(N^{-1/2})$ convergence rate for asymptotic bias is achievable. Dependence between either $\mathbf{n}$, respectively $\varepsilon$, and $\hat{\mathbf{X}}$ or $\hat{\mathbf{A}}$ can occur because the estimator $\hat{\mathbf{X}}$, and/or $\hat{\mathbf{A}}$, may be functions of $\mathbf{n}$, respectively $\varepsilon$. To break this dependence, a simple sample splitting procedure can be implemented, e.g., as outlined in Freeman and Weidner (2023), which is easily transferable to this setting.

Below regularity conditions from Bai (2009), Bai and Ng (2002) ensure estimates of the fixed-effects used as proxies in the kernel weights converge at the sufficient rate.

**Assumption 8** (Bai (2009) conditions). Let $N := \prod_{n=1}^{N_n} N_n$. (i). $\mathbb{E}\|X_{i_1...i_d}\|^4$ bounded. 
(ii). For each $n = 1, . . . , d$, $\mathbb{E}\|\varphi_{r_n}^{(n)}\|^4$ bounded, and $\frac{1}{N_n} \sum_{i_n=1}^{N_n} \varphi_{r_n}^{(n)}\varphi_{r_n}^{(n)'},$ converges to an $r_n \times r_n$ p.d. matrix as $N_n → \infty$. (iii). $\mathbb{E}\varepsilon_{i_1...i_d} = 0$, $\mathbb{E}(\varepsilon_{i_1...i_d})^8$ bounded.
(iv). $\mathbb{E}\varepsilon_{i_1...i_d}\varepsilon_{i'_1...i'_d} = \tau_{i_1...i_d,i'_1...i'_d}$, and $N^{-1} \sum_{i_1...i_d} \sum_{i'_1...i'_d} |\tau_{i_1...i_d,i'_1...i'_d}|$ bounded. (v). $\varepsilon_{i_1...i_d}$ independent of $X_{i_1...i_d}$ and $\varphi_{r_n}^{(n)}$ for all $i_1...i_d$ and $i'_1...i'_d$, and $n = 1, . . . , d$.
(vi). For $σ_{i_n,j_n,k_n,m_n}^{(n)} = \text{cov}(\varepsilon_{i_1,...,i_d,\varepsilon_{i_1,...,j_n},...},\varepsilon_{i_1,...,k_n,...,\varepsilon_{i_1,...,m_n,...,i_d}})$, for all $n = 1, . . . , d$;

$$N_{n}^{-2} \prod_{n' \neq n} N_{n'}^{-1} \sum_{i_n,j_n,k_n,m_n} \sum_{i_{n'},n' \neq n} |σ_{i_n,j_n,k_n,m_n}^{(n)}| \leq M < \infty$$

**Corollary 1.** If Assumption 8 and assumptions in Proposition 7 hold; $C_n^{-2} = 1/\min\{N_n, \prod_{n' \neq n} N_{n'}\}$ for each $n = 1, . . . , d$ if the Bai (2009) estimator is used in the matrix low-rank setting from Section 3.1 separately for each $n = 1, . . . , d$.

Below are sampling assumptions to achieve asymptotic normality.

---

7See Proposition A.1 in Bai (2009) appendix
Assumption 9. Let $\xi_X$, $\xi_\beta$, and $\xi_A$ be sequences bounded in probability as $N_n \to \infty$ for each $n = 1, \ldots, d$. Maintain $N := \prod_{n=1}^d N_n$. As $N_n \to \infty$ for each $n = 1, \ldots, d$, $\{\xi_X, \xi_\beta\} \to 0$, $\xi_X\xi_\beta = o_p(N^{-1/2})$ and $\xi_X\xi_A = o_p(N^{-1/2})$, where: (i). $\left(\|\hat{\Gamma}_{X_k} - \Gamma_{X_k}\|_F^2/N\right)^{1/2} = O_P(\xi_X)$ for each $k = 1, \ldots, K$, (ii). $\left(\|\hat{\mathcal{A}} - \mathcal{A}\|_F^2/N\right)^{1/2} = O_P(\xi_A)$, (iii). $\|\beta - \hat{\beta}\| = O_p(\xi_\beta)$.

Assumption 9(i) implies there exists a consistent estimate of $\mathbb{E}(X|A)$, with an arbitrary convergence rate. For implementation, $X$ can be modelled as a pure interactive fixed-effect model with regularity conditions on how $\mathcal{A}$ enters $X$’s data generating process. The convergence rate $\xi_X$ can be arbitrarily slow as long as $\xi_A$ and $\xi_\beta$ converge to zero sufficiently fast. Hence, in the context of Proposition 2, this is not an overly strong assumption on $\hat{\Gamma}_{X_k}$. The convergence rates in Assumption 9(ii)-(iii) can be $O_p(1/\sqrt{N})$ by Proposition 2 and Corollary 1. Assumption 9 details a potential reprieve from requiring that $\xi_A = O_p(1/\sqrt{N})$, and $\xi_\beta = O_p(1/\sqrt{N})$ if indeed it is possible for $\xi_X$ to converge at a faster rate.

The next assumption restricts $\eta$ and $\epsilon$ to be independent of $\hat{\Gamma}_X$ and $\hat{\mathcal{A}}$. As mentioned already, a sample splitting device can achieve this with regularity in $\eta$.

Assumption 10. Let $\hat{\Gamma}_X$ be the estimate for $\Gamma_X$ and $\hat{\mathcal{A}}$ the estimate for $\mathcal{A}$ in (13). Then, (i). $\eta_{k,i1...id} \perp \mathcal{A}_{i1...id}$, $\hat{\mathcal{A}}_{i1...id}$ for all $k = 1, \ldots, K$; (ii). $\eta_{k,i1...id} \perp \hat{\Gamma}_{X_{k'},i1...id}$ for all $k, k'$; (iii). $\epsilon_{i1...id} \perp \hat{\Gamma}_{X_{k'},i1...id}$ for all $k = 1, \ldots, K$.

Independence between $\eta_{k,i1...id}$ and $\mathcal{A}_{i1...id}$ in Assumption 10(i) is made for convenience for a cleaner derivation, but can be relaxed under further regularity conditions on the moments of $\eta_{k,i1...id}$. Assumption 10(iii) implies that in (14), $\hat{\Omega}^{-1}_X \text{vec}_K(\eta)'\text{vec}(\epsilon) = \hat{\Omega}^{-1}_X \text{vec}_K(\eta)'\text{vec}(\epsilon) = o_p(N^{-1/2})$.

Assumption 11. $\mathbb{E} [\eta_{i1...id}\eta'_{i1...id}]$ is non–singular and $\mathbb{E} [\epsilon_{i1...id}|\eta_{i1...id}] = 0$.

With Proposition 2 assumptions, Assumptions 8–11 and with $C_n^{-1} h_n = O_p(1/\sqrt{N_n})$ for all $n = 1, \ldots, d$ and $\mathcal{M} = \{1, \ldots, d\}$, maintaining that $N := \prod_{n=1}^d N_n$,

$$\sqrt{N}(\hat{\beta}_{IC} - \beta^0) = o_p(1) + \left(\frac{1}{N} \hat{\Omega}_X\right)^{-1} \frac{1}{\sqrt{N}} \text{vec}_K(\eta)'\text{vec}(\epsilon)$$

$$= o_p(1) + (\mathbb{E} [\eta_{i1...id}\eta'_{i1...id}] + o_p(1))^{-1} \frac{1}{\sqrt{N}} \text{vec}_K(\eta)'\text{vec}(\epsilon)$$

In the presence of both heteroskedasticity and correlation in all dimensions, the most general specification to be considered here, a general central limit theorem is required for
the term, $(N)^{-1/2} \text{vec}_K(\eta)\text{vec}(\varepsilon)$. In this case the variance is,

$$
\text{var}\left(\frac{1}{\sqrt{N}} \sum_{i_1\ldots i_d} \eta_{i_1\ldots i_d} \varepsilon_{i_1\ldots i_d}\right) = \frac{1}{N} \sum_{i_1\ldots i_d} \cdots \sum_{i_d'\ldots i_d'} \mathbb{E}(\eta_{i_1\ldots i_d} \eta_{i_1'\ldots i_d'}') \mathbb{E}(\varepsilon_{i_1\ldots i_d} \varepsilon_{i_1'\ldots i_d'}').
$$

Stronger independent and identically distributed assumptions are considered first, which are sequentially relaxed for stronger results later in the section.

**Assumption 12.**

(i). $\{\varepsilon_{i_1\ldots i_d}, \eta_{i_1\ldots i_d}\}$, are i.i.d. across $i_1\ldots i_d$,

(ii). $\mathbb{E}(\varepsilon_{i_1\ldots i_d}^2 | \eta_{i_1\ldots i_d}) = \mathbb{E}(\varepsilon_{i_1\ldots i_d}^2) =: \sigma_{\varepsilon}^2 \leq M < \infty$.

**Theorem 1** (Asymptotic distribution under homoskedasticity). Let the assumptions in Proposition 2 hold. Additionally, let Assumptions 8–12 hold. Then, for $N_n \rightarrow \infty$, with $N_n \lesssim \prod_{N'} N_n'$, for all $n = 1, \ldots, d$,

$$
\sqrt{N}(\hat{\beta}_{IC} - \beta^0) \rightarrow_d \mathcal{N}(0, \sigma_{\varepsilon}^2 \Omega_X^{-1}), \quad \Omega_X := \text{plim}_{N \rightarrow \infty} \frac{1}{N} \sum_{i_1\ldots i_d} \eta_{i_1\ldots i_d} \eta_{i_1\ldots i_d}'.
$$

Consistent estimation of the variance term is possible under the same set of assumptions. Estimate $\hat{\varepsilon} = Y - X \hat{\beta}_{IC} - \hat{\mathbf{A}}$. Then, $\frac{1}{N} \sum_{i_1\ldots i_d} \hat{\varepsilon}_{i_1\ldots i_d}^2 = \sigma_{\varepsilon}^2 + o_p(1)$ follows from the consistency of $\hat{\Gamma}_Y$, $\hat{\Gamma}_X$, and $\hat{\beta}_{IC}$. Likewise, $\mathbb{E}[\eta_{i_1\ldots i_d} \eta_{i_1'\ldots i_d}']$ can be estimated consistently from the sample analog $\frac{1}{N} \sum_{i_1\ldots i_d} (X_{i_1\ldots i_d} - \hat{\Gamma}_{X_{i_1\ldots i_d}}) (X_{i_1\ldots i_d} - \hat{\Gamma}_{X_{i_1\ldots i_d}}')$.

Assumptions on the error can be used to satisfy Lyapunov conditions. Instead, a central limit theorem is assumed, as is common in the interactive fixed-effects literature.  

**Assumption 13.**

(i). $\varepsilon_{i_1\ldots i_d}$ independent across $i_1\ldots i_d$, $\eta_{i_1\ldots i_d}$ are i.i.d. $i_1\ldots i_d$.

(ii). $\sigma_{i_1\ldots i_d}^2 := \mathbb{E}(\varepsilon_{i_1\ldots i_d}^2 | \eta)$ is bounded.

(iii). For nonrandom positive definite $\Sigma$,

$$
\text{plim}_{N \rightarrow \infty} \frac{1}{N} \sum_{i_1\ldots i_d} \sigma_{i_1\ldots i_d}^2 \eta_{i_1\ldots i_d} \eta_{i_1\ldots i_d}' = \Sigma, \quad \frac{1}{\sqrt{N}} \sum_{i_1\ldots i_d} \eta_{i_1\ldots i_d} \varepsilon_{i_1\ldots i_d} \rightarrow_d \mathcal{N}(0, \Sigma).
$$

**Theorem 2** (Asymptotic distribution under heteroskedasticity). Let the assumptions in Proposition 2 hold. Additionally, let Assumptions 8–12 and Assumption 13 hold. Then, for $N_n \rightarrow \infty$, with $N_n \lesssim \prod_{N'} N_n'$, for all $n = 1, \ldots, d$,

$$
\sqrt{N}(\hat{\beta}_{IC} - \beta^0) \rightarrow_d \mathcal{N}(0, \Omega_X^{-1} \Sigma \Omega_X^{-1}), \quad \Omega_X := \text{plim}_{N \rightarrow \infty} \frac{1}{N} \sum_{i_1\ldots i_d} \eta_{i_1\ldots i_d} \eta_{i_1\ldots i_d}'.
$$

\footnote{See for example Assumption E in Section 5 in Bai (2009).}
Assumption 14. For nonrandom positive definite $\tilde{\Sigma}$,

$$\lim_{N \to \infty} \frac{1}{N} \sum_{i_1 \ldots i_d} \sum_{i'_1 \ldots i'_d} \varepsilon_{i_1 \ldots i_d} \varepsilon_{i'_1 \ldots i'_d} \eta_{i_1 \ldots i_d} \eta_{i'_1 \ldots i'_d} = \tilde{\Sigma}, \quad \frac{1}{\sqrt{N}} \sum_{i_1 \ldots i_d} \eta_{i_1 \ldots i_d} \varepsilon_{i_1 \ldots i_d} \xrightarrow{d} \mathcal{N}(0, \tilde{\Sigma}).$$

The following restricts how much correlation is allowed in the error term.

Assumption 15. Define $\sigma_{i_1 \ldots i_d; i'_1 \ldots i'_d} := \mathbb{E}(\varepsilon_{i_1 \ldots i_d} \varepsilon_{i'_1 \ldots i'_d})$ and $\sigma_{k;i_1 \ldots i_d; i'_1 \ldots i'_d} := \mathbb{E}(\eta_{k;i_1 \ldots i_d} \eta_{k;i'_1 \ldots i'_d})$.

For finite $M < \infty$ such that,

$$\lim_{N \to \infty} \xi^2_{\sigma,N} \leq M \lim_{N \to \infty} \xi^2_{\sigma_k,N} \leq M$$

where,

$$\xi^2_{\sigma,N} := \frac{1}{N} \sum_{i'_1 \ldots i'_d \neq i_1 \ldots i_d} \sum_{i_1 \ldots i_d} \sigma^2_{i_1 \ldots i_d; i'_1 \ldots i'_d}, \quad \lim_{N \to \infty} \xi^2_{\sigma_k,N} := \frac{1}{N} \sum_{i'_1 \ldots i'_d \neq i_1 \ldots i_d} \sum_{i_1 \ldots i_d} \sigma^2_{k;i_1 \ldots i_d; i'_1 \ldots i'_d}.$$

For $\xi_X, \xi_A$ defined in Assumption 9, $\xi^2_{X} \xi_{\sigma,N} N^{1/2} = o_p(1)$, $\xi^2_{A} \xi_{\sigma_k,N} N^{1/2} = o_p(1)$ for all $k$.

Assumption 15 can be achieved if, e.g. $\{\xi_X, \xi_A\} = o_p(N^{-1/4})$, without further restrictions on $\xi_{\sigma,N}$. If $\varepsilon_{i_1 \ldots i_d}$ are i.i.d., or i.n.i.d., $\xi_{\sigma,N} = 0$, satisfying the assumption.

Theorem 3 (Asymptotic distribution under heteroskedasticity and correlation). Let the assumptions in Proposition 2 hold. Additionally, let Assumptions 11-13, and 14-15 hold.

Then, for $N_n \to \infty$, with $N_n \leq \prod_{n' \neq n} N_{n'}$, for all $n = 1, \ldots, d$,

$$\sqrt{N} (\hat{\beta}_{1C} - \beta^0) \xrightarrow{d} \mathcal{N} \left(0, \Omega_X^{-1} \hat{\Sigma} \Omega_X^{-1} \right).$$

From Bai (2009), $\hat{\Sigma}$ can be estimated with the Newey and West (1987) kernel approach for time dimension, in conjunction with a partial sample estimator for cross-sections.

4 Discussion of estimators

This section serves to discuss the results in Section 3, motivate further some of the chosen methods, and provide some methods to estimate proxies for kernel weights.

4.1 Matrix method results

Proposition 1 takes for granted the dimension to flatten across admits a low rank interactive fixed-effect term for the least square method in Bai (2009). 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 in the pure factor model, which could be repeated.
along the different flattenings in the multidimensional setting. However, establishing the interactive fixed-effect rank in the presence of covariates is an open area of research.

Consider the factor model applied to a flattening that may not be low-rank. First, assume $\varphi^{(1)}$ varies in a high-dimensional parameter space, e.g. with $N_1 < N_2 N_3$, $\varphi^{(1)} \in \mathbb{R}^{N_1 \times N_1}$ and $\Gamma \in \mathbb{R}^{N_2 N_3 \times N_1}$ with each column mutually orthogonal for both these matrices. Then $\varphi^{(1)} \Gamma'$ is full-rank and the factor model will not control for this. 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. This situation is exemplified in simulations in Section 5. The matrix method requires knowledge of this low rank dimension - the kernel weighted-within does not.

Knowing the multilinear rank, Moon and Weidner (2015) show a factor model with at least $r_n$ factors results in consistent $\beta$ estimates. However, with generic heteroskedasticity and weak correlations, the factor model can only be bounded by a slow rate of convergence. In the three dimensional setting with proportional dimensions, this bound is as slow as $N^{-1/6}$, where $N$ is total sample size. This is too slow to use for standard inference procedures.

4.2 Estimating kernel weight proxies

Discussed here are some functionals of multidimensional arrays that are useful for estimating proxies to form kernel weights. For this purpose, it is important to find proxies that isolate variation in each dimension since weights are formed separately for each index.

Consider for any of the dimensions $n$ the corresponding matrix of left singular vectors from above, $U^{(n)}$, estimated with noise $\varepsilon_{ijt}$. That is, each $U^{(n)}$ are calculated from the matrix $V^{(n)} = A^{(n)} + \varepsilon^{(n)}$. Under regularity conditions on the noise term $\varepsilon$, the left singular vectors from this decomposition consistently estimate $U^{(n)}$, up to rotations. In the three dimensional case, define the $r_n$-vector $\tilde{U}_i^{(n)}$ as the $i$-th row of the left singular matrix of $A + \varepsilon$ flattened in the $n$th dimension. Bai and Ng (2002) show the following “up-to-rotation” consistency result:

Lemma 1 (Theorem 1 from Bai and Ng (2002)). For any fixed integer $k \geq 1$, there exists an $(r_n \times k)$ matrix $H_k^k$ with rank($H_k^k$) = min\{$k, r_n$} and $C_n = \min \{\sqrt{N_n}, \prod_{n' \neq n} \sqrt{N_{n'}}\}$

---

9This actually requires knowing an upper bound on multilinear rank, not the actual multilinear rank.

10Debias do exist, but would achieve at best $N^{-1/3}$ convergence.
such that for each \( n \) under some regularity conditions

\[
C_n^2 \left\| \widehat{U}_{in}^{(n)} - H_n^{k'} \varphi_{in}^{(n)} \right\|^2 = O_p(1).
\]

Hence, if the true error, \( \mathbf{V} = \mathbf{A} + \mathbf{\varepsilon} \), is observed, this establishes consistency of \( \widehat{U}_{in}^{(n)} \) as proxies. The rotation matrices, \( H_n^k \), in Lemma 1 can be ignored since these do not change relative distances used in the kernel weights, see Lemma A.2. However, the true error, \( \mathbf{V} = \mathbf{A} + \mathbf{\varepsilon} \), is not observed. Instead, the observed error \( \hat{V}_{ijt} = Y_{ijt} - X_{ijt} \hat{\beta} = X_{ijt}(\beta^0 - \hat{\beta}) + A_{ijt} + \varepsilon_{ijt} \), depends on the estimate \( \hat{\beta} \), hence should be accounted for. Proposition A.1 in Bai (2009) provides,

\[
\frac{1}{N_n} \sum_{i_n=1}^{N_n} \left\| \widehat{U}_{in}^{(n)} - H_n^{k'} \varphi_{in}^{(n)} \right\|^2 = O_p(\|\beta^0 - \hat{\beta}\|^2) + O_p\left( \frac{1}{\min\{N_n, \prod_{n' \neq n} N_{n'}\}} \right),
\]

if \( \widehat{U}_{in}^{(n)} \) come from least squares in Bai (2009). Ignoring \( H_n^k \), \( \frac{1}{N_n} \sum_{i_n=1}^{N_n} \left\| \widehat{U}_{in}^{(n)} - \varphi_{in}^{(n)} \right\|^2 = O_p(N_n^{-1}) \) if each dimension sample size grows at a roughly comparable rate and the convergence rate for \( \|\beta^0 - \hat{\beta}\| \) in Proposition 1 is used.\(^{11}\) Then, \( C_n^{-2} \) from Proposition 2 is \( 1/N_n \) and the parametric rate for the kernel weighted estimator can be established. Again, the rotation matrices, \( H_n^k \), can be ignored from Lemma A.2.

5 Simulations

Two simulation exercises are performed. The first analyses the performance of the matrix methods compared to the kernel weighted methods when sample size is allowed to increase. These results are summarised in Figure 1 and Table 1. The second simulation exercise analyses some fixed sample properties of the set of estimators when Assumption 3 is violated in some dimensions. This assumption governs the rank of the fixed-effect term after it is transformed into a matrix. These results are summarised in Table 2.

\(^{11}\)To be precise, the comparable rate restriction for any dimension \( n = 1, \ldots, d \) is \( N_n \lesssim \prod_{n' \neq n} N_{n'} \). The comparable rate commonly used in the two-dimensional panel literature is \( N \approx T \). Hence, in the multidimensional problem, the relative rate requirement is milder than the two-dimensional case.
5.1 Growing sample exercise

Data is generated as,

\[ Y_{ijt} = X_{ijt} + A_{ijt} + \varepsilon_{ijt} \]  

\[ X_{ijt} = 2A_{ijt} - \rho^2 \sum_{\ell=1}^{2} (\lambda_{i\ell} + \lambda_{i-1\ell}) (\gamma_{j\ell} + \gamma_{j-1\ell}) (f_{t\ell} + f_{t-1,\ell}) + \eta_{ijt} \]

\[ \varepsilon_{ijt} = \left(\frac{1}{\sqrt{2}}\right) (\nu_{ijt} + \nu_{ijt-1} + \nu_{ij-1t} + \nu_{ij-1t-1} + \nu_{i-1jt} + \nu_{i-1j-1t} + \nu_{i-1j-1t-1}) \]

with, \( A_{ijt} = \sum_{\ell=1}^{2} \lambda_{i\ell}\gamma_{j\ell}f_{t\ell} \). Also,

\[ \eta_{ijt} \overset{i.i.d.}{\sim} N(0, 1), \nu_{ijt} \overset{i.n.i.d.}{\sim} N(0, \min\{4, \eta_{ijt}^2\}) \] and for each \( \ell \), \( \lambda_{i\ell}, \gamma_{j\ell}, f_{t\ell} \overset{i.i.d.}{\sim} N(0, 1) \).

Error \( \varepsilon_{ijt} \) admits heteroskedasticity, and correlation in each dimension of the data. The order of \( i \) and \( j \), are randomised to simulate an unknown cross correlation structure.

The results depicted in Figure 1 show two specifications for \( \rho \). The left panel depicts \( \rho = 1 \), and the right panel depicts \( \rho = 1/3 \). Estimators considered use 2 estimated factors such that the multilinear rank is correctly predicted. The left panel shows that whilst the matrix methods, labelled Factor, look consistent, they converge at a rate much slower than the empirical bounds. This leads to spurious inference. The right panel shows more egregious convergence when the bias problem is made worse with \( \rho = 1/3 \). Indeed, for the sample sizes considered, the empirical bounds do not ever cover the true parameter value in either panel.

This contrasts with the results for the kernel weighted fixed-effect projection. Finite sample bias is negligible compared to variance in both the left and right panel. This estimator has roughly the same variance as the matrix methods, but for the sample sizes considered, the empirical bounds are always valid.

Table 1 displays coverage and compares the kernel weighted estimator without and with the inference correction from Section 3.3, and the factor model in any dimension. The Ker Inf coverage, which is coverage for the inference corrected estimator, is always better than the uncorrected estimator. This follows directly from the theoretical arguments made in Section 3.3. The factor model undercovers with progressively worse coverage as the sample size grows.
5.2 Fixed sample exercise

Table 2 shows simulation results for the following DGP,

\[ Y_{ijt} = X_{ijt}\beta + A_{ijt} + \varepsilon_{ijt} \]

\[ X_{ijt} = A_{ijt} + \sum_{\ell=1}^{N_1} \left( \lambda_i + \lambda_{i-1} \right) \left( \gamma_j + f_{t-1} \right) + \eta_{ijt} \]

with \( A_{ijt} = \sum_{\ell=1}^{N_1} \lambda_i \gamma_j f_{t-1} \) and all other parameters the same as (15). \( A_{ijt} \) is normalised to have unit variance. \( A \) is specified such that it is rank 1 when flattened in the first dimension and rank \( N_1 \) when flattened in either dimension two or three. That is, the multilinear rank is \( r = (1, N_1, N_1) \). This comes directly from the data generating process for each fixed-effect, where the matrix \( \lambda \) is designed to be rank-1 and the matrices \( \gamma \) and \( f \) are designed to be rank-\( N_1 \).

In Table 2, 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 have bias. 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 bias is close to zero when
the correct dimension is flattened over (the first dimension in this case) and very poor bias when the incorrect dimension is used (the second and third dimensions). Lastly, the kernel differencing estimator is estimated with Gaussian kernel function with various bandwidths; which are standardised to be equivalent to standard deviations of the proxy measures. Kernel estimators show a clear bias-variance trade-off, but all with bias of the same order as the correct factor model.

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

### Table 1: Estimated coverage for %5 nominal; test 10,000 Monte Carlo rounds

| Dim | Ker | Ker Inf | Factor 1 | Factor 2 | Factor 3 |
|-----|-----|---------|----------|----------|----------|
| 30  | 0.48| 0.85    | 0.04     | 0.05     | 0.12     |
| 40  | 0.56| 0.92    | 0.03     | 0.03     | 0.09     |
| 50  | 0.57| 0.93    | 0.02     | 0.02     | 0.08     |
| 60  | 0.58| 0.95    | 0.01     | 0.01     | 0.06     |
| 70  | 0.58| 0.95    | 0.00     | 0.00     | 0.05     |
| 80  | 0.60| 0.95    | 0.00     | 0.00     | 0.04     |

Ker stands for the kernel weighted estimator without inference correction. Ker Inf stands for the kernel weighted estimator with the inference correction. Factor # is the matrix method making dimension # the rows, and the other dimensions jointly the columns. Dim stands for the size of each dimension, such that the total sample size is the cube of this. Coverage is measured for heteroskedasticity and correlation robust standard errors.

6 Empirical application - demand estimation for beer

The methods proposed in this paper are applied to estimate 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 fortnights \(t\). Fixed-effects that interact across all three dimensions can control for temporal taste
| 3-D               | Bias  | St. dev. | RMSE  | 4-D               | Bias  | St. dev. | RMSE  |
|-------------------|-------|----------|-------|-------------------|-------|----------|-------|
| OLS               | 0.3655| 0.0038   | 0.3655|                   | 0.1190| 0.0211   | 0.1209|
| Fixed-effects     | 0.3708| 0.0040   | 0.3708|                   | 0.1321| 0.0265   | 0.1347|
| Kernel Inf (h = 0.05) | 0.0017| 0.0083   | 0.0085|                   | -0.0002| 0.0068   | 0.0068|
| Kernel Inf (h = 0.1) | 0.0035| 0.0060   | 0.0069|                   | 0.0003| 0.0040   | 0.0040|
| Kernel Inf (h = 0.2) | 0.0090| 0.0050   | 0.0103|                   | 0.0027| 0.0031   | 0.0041|
| Factor (dim = 1)  | -0.0028| 0.0041   | 0.0050|                   | 0.0013| 0.0016   | 0.0021|
| Factor (dim = 2)  | 0.3603| 0.0064   | 0.3604|                   | -0.0013| 0.0016   | 0.0021|
| Factor (dim = 3)  | 0.3603| 0.0064   | 0.3604|                   | 0.1199| 0.0256   | 0.1226|
| Factor (dim = 4)  | NA    | NA       | NA    |                   | 0.1205| 0.0269   | 0.1234|

Table 2: Fixed-sample size simulation with heterogeneous multilinear rank

3D model \((N_1 = N_2 = N_3 = 40)\), with 10,000 Monte Carlo rounds.

4D model \((N_1 = N_2 = N_3 = N_4 = 20)\), with 1,000 Monte Carlo rounds. Bandwidth scaled up by \(\sqrt{2}\) to account for smaller per dimension sample. All results relate to \(\beta\) estimation.

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 \((i)\). Analysts may want some way to control for such heterogeneity that interacts over all dimensions, even if they do not explicitly observe such sponsorship variables. 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 a 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.

Models for demand estimation ideally account for endogenous variation in prices and quantity. The classic instrumental variable approach is to find a supply shifter that shifts the supply curve, allowing the econometrician to trace out the slope of the demand curve. 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 fixed-effects and noise term, \( \varepsilon \).

For valid inference, the instrument is also required to be strong, in the sense that it is strongly correlated with the price of beer. In this dataset, correlation between the price of barley, which varies over only month, every second \( t \), and price of beer depends on how beer price is first aggregated. If beer price is first integrated over \( i, j \), and to the monthly level, such that it only varies over every second \( t \), then it is highly correlated with the price of barley, at 0.79. However, if beer price is not aggregated at all it is only correlated at 0.001. This heuristic suggests there are important product and store level price drivers for beer that are not accounted for by fluctuations in the price of barley, or indeed by price fluctuations only over time. Standard errors for the IV estimator below are much larger than other estimators, which may be explained partly by this loss in effective sample size.\(^{12}\)

Table 3 refers to the estimates the following model,

\[
\log(\text{quantity}_{ijt}) = \log(\text{price}_{ijt}) \beta + A_{ijt} + \varepsilon_{ijt}
\]

where \( A_{ijt} \) is the interactive fixed-effects term. No additional controls are included here since they are low-dimensional and subsumed by additive fixed-effects. Estimates for pooled OLS are positive, a contradiction that demand curves are downward sloping. IV estimates are negative, but standard errors are large. As noted above, the effective sample size for the second stage of the instrumental variable approach is orders of magnitude smaller because the instrument only varies over time, specifically every second fortnight. Estimates under the additive model for fixed-effects are also negative, with much smaller standard errors than the IV estimate.

The matrix method is implemented in all three dimensions, i.e., the three-dimensional array is transformed into three different matrices - with products as rows, with stores as rows, and time as rows. Five factors are estimated for each. The results suggest the factor model for fixed-effects can be sensitive to how the array is organised into a two-dimension problem. Simulation results from Section 5 suggest a potential explanation - sparsity of the dimension specific fixed-effects in the interaction term can be heterogeneous, and imply a different factor model rank when the tensor of data is flattened, or matricised, in

\(^{12}\)The effective sample size for the IV estimator drops from \( N_1N_2T \) to just \( T \).
different dimensions. The kernel weighted fixed-effects estimates elasticities similar to the IV point estimates, but with much smaller standard errors. The kernel weighted estimates are similar to the own-price elasticity estimates from Table 1 in Hausman, Leonard and Zona (1994), which average around $-2.5$.

Figure 2 displays estimates from the factor model with products as rows, and the weighted-within estimator, with estimated confidence bands, allowing the estimated number of interactive fixed-effects terms to increase. The figure displays two things. First, there is evidence that up to approximately 9-10 interactive terms may be appropriate. Second, at each number of estimated interactive terms, there is a substantial shift in estimate when the weighted-within transformation is applied. Taking the model proposed for granted, this implies a persistent debias with the weighted-within transformation.

| Estimator                          | $\hat{\beta}$ (St. err.) |
|------------------------------------|----------------------------|
| Pooled OLS                         | 1.304 (0.007)              |
| Additive Fixed-effects             | -0.069 (0.040)             |
| Pooled IV                          | -2.148 (1.158)             |
| Factor (Product (i) as rows)       | -2.787 (0.055)             |
| Factor (Store (j) as rows)         | -0.098 (0.042)             |
| Factor (Time (t) as rows)          | -0.033 (0.003)             |
| Weighted-within                    | -3.317 (0.053)             |

Table 3: Log-log demand elasticities (45 products, 48 stores, 110 fortnights).

Estimates for the standard errors are included in brackets. The factor model estimator uses five factors in all three cases.

7 Conclusion

This paper develops methods to generalise the interactive fixed-effect to multidimensional datasets 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 and generally have slow convergence rates. Nonetheless, these provide useful preliminary estimates. The multiplicative interactive fixed-effect error from the kernel weighted method show an improvement on the convergence rate of slope coefficient estimates to the parametric rate, and suggest a more robust approach to projecting fixed-
Figure 2: Elasticity estimates for increasing number of estimated factors

Effects. Simulations show finite sample properties when the sample size is allowed to grow and when it is fixed. These simulations exemplify how sensitive the two-dimensional methods are to model specification issues as simple as how to organise the dataset. They also show the robustness of the kernel weighted fixed-effects estimators without having to make these same specifications. A method for inference with the weighted fixed-effects estimator is also introduced.

The model is applied to a demand model for beer consumption. The application demonstrates that simply applying the two-dimensional factor model approach is sensitive to how dimensions are arranged to suit these estimators. The weighted-within transformation estimates elasticities close to an instrumental variable point estimate, but with substantially better precision.

References

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

Babii, A., E. Ghysels, and J. Pan (2022). Tensor principal component analysis. *arXiv preprint arXiv:2212.12981*. 
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 (2022). Discretizing unobserved heterogeneity. *Econometrica* **90**(2), 625–643.

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.

Freeman, H. and D. Kristensen (2024). Inference on linear regressions with two-way unobserved heterogeneity. Unpublished Manuscript.

Freeman, H. and M. Weidner (2023). Linear panel regressions with two-way unobserved heterogeneity. *Journal of Econometrics* **237**(1), 105498.

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.

Hástad, J. (1989). Tensor rank is np-complete. In *International Colloquium on Automata, Languages, and Programming*, pp. 451–460. Springer.

Hausman, J., G. Leonard, and J. D. Zona (1994). Competitive analysis with differentiated products. *Annales d’Economie et de Statistique*, 159–180.

Kapetanios, G., L. Serlenga, and Y. Shin (2021). Estimation and inference for multidimensional 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.

Kuersteiner, G. M. and I. R. Prucha (2020). Dynamic spatial panel models: Networks, common shocks, and sequential exogeneity. *Econometrica* **88**(5), 2109–2146.

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

Matyas, L. (2017). The econometrics of multi-dimensional panels. *Advanced studies in
Appendix: Proofs

A.1 Proofs of main paper results

Proof of Proposition 2. In the following, let $\text{vec}_K(\tilde{X})$ 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,

$$
\beta_{KFE,c} = \left( \text{vec}_K(\tilde{X})' \text{vec}_K(\tilde{X}) \right)^{(-1)} \text{vec}_K(\tilde{X})' \text{vec}(\tilde{Y}) \\
= \beta^0 + \left( \text{vec}_K(\tilde{X})' \text{vec}_K(\tilde{X}) \right)^{(-1)} \text{vec}_K(\tilde{X})' \left( \text{vec}(\tilde{A}) + \text{vec}(\tilde{e}) \right),
$$
such that,
\[
\| \beta_{KFE,C} - \beta^0 \| = \left\| \left( \text{vec}_K(\bar{X})' \text{vec}_K(\bar{X}) \right)^{(-1)} \text{vec}_K(\bar{X})' \left( \text{vec}(\bar{A}) + \text{vec}(\bar{e}) \right) \right\|
\leq \| \kappa_N \| + \| \omega_N \|
\]
where
\[
\| \kappa_N \| := \left\| \left( \text{vec}_K(\bar{X})' \text{vec}_K(\bar{X}) \right)^{(-1)} \text{vec}_K(\bar{X})' \text{vec}(\bar{A}) \right\|;
\| \omega_N \| := \left\| \left( \text{vec}_K(\bar{X})' \text{vec}_K(\bar{X}) \right)^{(-1)} \text{vec}_K(\bar{X})' \text{vec}(\bar{e}) \right\|.
\]
\| \omega_N \| is bounded at the parametric rate by standard arguments, so \| \kappa_N \| is the focus here.

Notice,
\[
\| \kappa_N \| \leq \left\| \left( \text{vec}_K(\bar{X})' \text{vec}_K(\bar{X}) \right)^{(-1)} \right\| \left\| \text{vec}_K(\bar{X})' \text{vec}(\bar{A}) \right\|
\]
Focus on the right hand part, and let \( \langle \cdot, \cdot \rangle_F \) be the Frobenius inner product,
\[
\left\| \text{vec}_K(\bar{X})' \text{vec}(\bar{A}) \right\| = \left\| \begin{bmatrix} \langle \bar{X}_1, \bar{A} \rangle_F \\ \vdots \\ \langle \bar{X}_L, \bar{A} \rangle_F \end{bmatrix} \right\| \leq \left\| \begin{bmatrix} \sum_{i=1}^N \bar{X}_{i,1} \bar{A}_i \\ \vdots \\ \sum_{i=1}^N \bar{X}_{i,K} \bar{A}_i \end{bmatrix} \right\| \tag{A.1}
\]
Let \( \tilde{\varphi}_{i_n,\ell}^{(n)} \) denote now the estimate of \( \varphi_{i_n,\ell}^{(n)} \) according the the kernel weighted estimator.
The entries for each \( k \) in (A.1) can be written as,
\[
\sum_{i=1}^N \bar{X}_{i,k} \bar{A}_i = \frac{1}{N} \sum_{i_1,\ldots,i_d} \bar{X}_{i_1,\ldots,i_d,k} \sum_{\ell=1}^L \prod_{n} W_{i_n,j_n}^{(n)} \left( \varphi_{i_n,\ell}^{(n)} - \varphi_{j_n,\ell}^{(n)} \right)
\leq \left( \frac{1}{N} \sum_{\ell=1}^L \sum_{i_1,\ldots,i_d} \bar{X}_{i_1,\ldots,i_d,k}^2 \sum_{j_1,\ldots,j_d} \prod_{n} W_{i_n,j_n}^{(n)} \right)^{1/2} \times \ldots
\]
\begin{align*}
&\cdots \times \left( \frac{1}{N} \sum_{i_1,\ldots,i_d, j_1,\ldots,j_d} \prod_{n} W_{i_n,j_n}^{(n)} \left\| \varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)} \right\|^2 \right)^{1/2} \\
&= \sqrt{LO_p(1)} \left( \frac{1}{N} \sum_{i_1,\ldots,i_d, j_1,\ldots,j_d} \prod_{n} W_{i_n,j_n}^{(n)} \left\| \varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)} \right\|^2 \right)^{1/2}
\end{align*}
\]
The last line comes from bounded second moments in X and from weights summing to 1. The final term can be treated separately for each $n$.

$$\frac{1}{N_n} \sum_{i_n} \sum_{j_n} W^{(n)}_{i_n,j_n} \left\| \varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)} \right\|^2$$

Use the triangle inequality and $(a + b + c)^2 = 3(a^2 + b^2 + c^2)$ to bound this as

$$\frac{1}{N_n} \sum_{i_n} \sum_{j_n} W^{(n)}_{i_n,j_n} O \left( \left\| \varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)} \right\|^2 + \left\| \varphi_{j_n}^{(n)} - \varphi_{j_n}^{(n)} \right\|^2 + \left\| \varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)} \right\|^2 \right).$$

Take the first of these terms,

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

$$= O_p(C_n^{-2})$$

Likewise, the second term is $O_p(C_n^{-2})$ since weights are symmetric and sum to 1.

The third term, with the shorthand $K_{ij,h}^{(n)} = k \left( \frac{1}{h_n} \left\| \varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)} \right\| \right)$, and $1(a,b)_{ij} = 1\{a < \left\| \varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)} \right\| < b\}$,

$$\frac{1}{N_n} \sum_{i_n} \sum_{j_n} W^{(n)}_{i_n,j_n} \left\| \varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)} \right\|^2 = \frac{1}{N_n} \sum_{i_n} \sum_{j_n} 1(0, h_n)_{ij} \frac{K_{ij,h}^{(n)} \left\| \varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)} \right\|^2}{\sum_{j_n} K_{ij,h}^{(n)}}$$

$$+ \frac{1}{N_n} \sum_{i_n} \sum_{j_n} \sum_{m=1}^{\infty} 1(mh_n, (m+1)h_n)_{ij} \frac{K_{ij,h}^{(n)} \left\| \varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)} \right\|^2}{\sum_{j_n} K_{ij,h}^{(n)}}$$

$$\leq O(h_n^2) \frac{1}{N_n} \sum_{i_n} \sum_{j_n} 1(h_n)_{ij} K_{ij,h}^{(n)}$$

$$+ \frac{h_n^2}{N_n} \sum_{i_n} \sum_{m=1}^{\infty} (m+1)^2 \sum_{j_n} 1(mh_n, (m+1)h_n)_{ij} K_{ij,h}^{(n)}$$

The first term after the last inequality is $O(h_n^2)$ since the denominator always dominates the numerator. For the last term, for each $i_n$, the term

$$\sum_{m=1}^{\infty} (m+1)^2 \sum_{j_n} 1(mh_n, (m+1)h_n)_{ij} K_{ij,h}^{(n)} \leq \sum_{j_n} 2^2 K_{ij,h}^{(n)}$$

32
because for each $i_n,j_n$ the indicator function can only be 1 for one $m \in \{1,\ldots,\infty\}$, and since $(m+1)^2 R_{ij,h}^{(n)}$ is decreasing in $m$, from Assumption 5.(iv), across the indicator functions the term is maximised at $m = 1$. Hence, the last line is bounded $O(h_n^2)$.

This makes the whole term bounded $O_p(h_n^2)$. Hence, for each $n$,

$$
\frac{1}{N_n} \sum_{i_n} \sum_{j_n} W_{i_n,j_n}^{(n)} \left\| \varphi_{i_n}^{(n)} - \varphi_{j_n}^{(n)} \right\|^2 \leq \left( O_p(C_{n}^{-2}) + O_p(h_n^2) \right).
$$

Taking the product of these terms over dimensions forms the statement of the result. $\blacksquare$

**Proof of Theorem 3.** Begin with (14),

$$
\hat{\beta}_{IC} - \beta^0 = \hat{\Omega}_X^{-1} B + \hat{\Omega}_X^{-1} \text{vec}(\hat{\eta})' \text{vec}(\varepsilon)
$$

where, the bias term, $B = B_1 + B_2$. Let $N := \prod_{n=1}^N N_n$. Then,

$$
B_1 := \text{vec}(\hat{\eta})' \text{vec}(\hat{\Gamma}_X \cdot (\beta^0 - \bar{\beta})) ; \quad B_2 := \text{vec}(\hat{\eta})' \text{vec}(\mathcal{A} - \hat{\mathcal{A}}).
$$

To proceed, firstly, need to show $N^{-1/2} B_1$ and $N^{-1/2} B_2$ are $o_p(1)$.

$$
\frac{1}{\sqrt{N}} B_1 = \frac{1}{\sqrt{N}} \text{vec}(\Gamma_X - \hat{\Gamma}_X)' \text{vec}(\hat{\Gamma}_X \cdot (\beta^0 - \bar{\beta})) + \frac{1}{\sqrt{N}} \text{vec}(\eta)' \text{vec}(\hat{\Gamma}_X \cdot (\beta^0 - \bar{\beta}))
$$

$$
\leq O_p(\sqrt{N} \xi_\beta) + O_p(\xi_\beta) \frac{1}{\sqrt{N}} \sum_{k=1}^K \text{vec}(\eta)' \text{vec}(\hat{\Gamma}_{X_k}).
$$

Boundedness of $N^{-1/2} \left\| \hat{\Gamma}_{X_k} \right\|_F$ comes from Assumption 1.(i). From Assumption 9, the first term is then $o_p(1)$. Assumptions 9, 10 bound the second term as $O_p(\xi_\beta) = o_p(1)$.

$$
\frac{1}{\sqrt{N}} B_2 = O_p(\sqrt{N} \xi_\beta) + \frac{1}{\sqrt{N}} \text{vec}(\eta)' \text{vec}(\mathcal{A} - \hat{\mathcal{A}}).
$$

The first term is $o_p(1)$ by Assumption 9. For the second term,

$$
\text{var} \left( \frac{1}{\sqrt{N}} \text{vec}(\eta)' \text{vec}(\mathcal{A} - \hat{\mathcal{A}}) \right) = o_p(1)
$$

and $E \left( N^{-1/2} \text{vec}(\eta)' \text{vec}(\mathcal{A} - \hat{\mathcal{A}}) \right) = 0$, where the zero variance derivation is analogous to the derivation below for $\text{vec}(\Gamma_X - \hat{\Gamma}_X)' \text{vec}(\varepsilon)$ term from Assumption 15.

Next,

$$
\frac{1}{\sqrt{N}} \text{vec}(\hat{\eta})' \text{vec}(\varepsilon) = \frac{1}{\sqrt{N}} \text{vec}(\Gamma_X - \hat{\Gamma}_X)' \text{vec}(\varepsilon) + \frac{1}{\sqrt{N}} \text{vec}(\eta)' \text{vec}(\varepsilon).
$$
The second term is asymptotically normally distributed by Assumption 14. The first term
is mean zero and has variance,

$$\mathbb{E}\left[ \frac{1}{N} \sum_{i_1, \ldots, i_d} \left( \Gamma_{X,i_1,\ldots,i_d} - \hat{\Gamma}_{X',i_1,\ldots,i_d} \right) \left( \Gamma_{X,i_1,\ldots,i_d} - \hat{\Gamma}_{X',i_1,\ldots,i_d} \right)' \varepsilon_{i_1,\ldots,i_d}^2 + \ldots \right]$$

$$\ldots + \frac{1}{N} \sum_{i_1' \ldots i_d' \neq i_1 \ldots i_d} \sum_{i_1, \ldots, i_d} \left( \Gamma_{X,i_1,\ldots,i_d} - \hat{\Gamma}_{X',i_1,\ldots,i_d} \right) \left( \Gamma_{X,i_1',\ldots,i_d'} - \hat{\Gamma}_{X',i_1',\ldots,i_d'} \right)' \varepsilon_{i_1,\ldots,i_d} \varepsilon_{i_1',\ldots,i_d'}$$

The first term is $o_p(1)$ from Assumption 9 and bounded $\mathbb{E} \varepsilon_{i_1,\ldots,i_d}^2$. Use shorthand index notation $i$ in place of $i_1, \ldots, i_d$. From Assumption 10 for each $k, k'$,

$$\frac{1}{N} \sum_{j \neq i} \sum_{i} \mathbb{E} \left[ \left( \Gamma_{X,k,i} - \hat{\Gamma}_{X,k,i} \right) \left( \Gamma_{X',j,i} - \hat{\Gamma}_{X',j,i} \right) \right] \mathbb{E} \left[ \varepsilon_i \varepsilon_j \right]$$

$$\leq \left( \frac{1}{N} \sum_{j \neq i} \sum_{i} \mathbb{E} \left[ \left( \Gamma_{X,k,i} - \hat{\Gamma}_{X,k,i} \right) \left( \Gamma_{X',j,i} - \hat{\Gamma}_{X',j,i} \right) \right]^2 \right)^{1/2} \left( \frac{1}{N} \sum_{j \neq i} \sum_{i} \mathbb{E} \left[ \varepsilon_i \varepsilon_j \right]^2 \right)^{1/2}$$

$$\leq \left( \frac{1}{N} \sum_{j \neq i} \sum_{i} \mathbb{E} \left[ \left( \Gamma_{X,k,i} - \hat{\Gamma}_{X,k,i} \right)^2 \right] \mathbb{E} \left[ \left( \Gamma_{X',j,i} - \hat{\Gamma}_{X',j,i} \right)^2 \right] \right)^{1/2} \left( \frac{1}{N} \sum_{j \neq i} \sum_{i} \sigma_i^2 \right)^{1/2}$$

$$= O_p \left( N^{1/2} \varepsilon^2 X \varepsilon_{\sigma,N} \right).$$

Notation comes from Assumption 15 which also implies $O_p \left( N^{1/2} \varepsilon^2 X \varepsilon_{\sigma,N} \right) = o_p(1)$.

Finally, $\hat{\Omega}_X^{-1} = (\mathbb{E} \varepsilon_{i_1,\ldots,i_d}^2 \eta_{i_1,\ldots,i_d}) + o_p(1)^{-1}$. Assumption 14 gives the asymptotic distribution for $N^{-1/2} \text{vec}_K(\eta') \text{vec}(\varepsilon)$, which completes the proof.

\[ \blacksquare \]

### A.2 Proofs of supplementary results

**Lemma A.1** (Regularity of proxy measures). Let $\varphi_{kn}^{(n)} \in \hat{\Phi}_n$ be the proxy space for the fixed-effects and let $k(\cdot)$ be a bounded kernel function from Assumption 6. If Assumption 6 holds, then as $h_n \to 0$ such that $N_n h_n \to \infty$,

$$\lim_{N_n \to \infty} \frac{1}{N_n} \sum_{n=1}^{N_n} \sum_{i_n=1}^{N_n} k \left( \frac{1}{h_n} \| \varphi_{kn}^{(n)} - \varphi_{kn}^{(n)} \| \right) > 0 \quad \text{wpa}. \quad (A.3)$$

This implies

$$N_n^{-1} \text{Tr} \left\{ (I_{N_n} - W_n)' (I_{N_n} - W_n) \right\} > 0 \quad \text{wpa}. \quad (A.4)$$
Proof of Lemma A.1. The supposition is that when
\[ \frac{1}{N_n} \sum_{i_n=1}^{N_n} \sum_{i_n' \neq i_n} I \left( \tilde{\varphi}^{(n)}_{i_n} \in B_{he \left( \varphi^{(n)}_{i_n} \right)} \right) > 0 \quad \text{wpal.} \quad (A.5) \]
in Assumption 6 holds, then (A.3) holds. Take
\[
\frac{1}{N_n} \sum_{i_n=1}^{N_n} \sum_{i_n' \neq i_n} k \left( \frac{1}{h_n} \left\| \varphi^{(n)}_{i_n} - \varphi^{(n)}_{i_n'} \right\| \right) = \frac{1}{N_n} \sum_{i_n=1}^{N_n} \sum_{i_n' \in B_{he \left( \varphi^{(n)}_{i_n} \right)}} k \left( \frac{1}{h_n} \left\| \varphi^{(n)}_{i_n} - \varphi^{(n)}_{i_n'} \right\| \right)
+ \frac{1}{N_n} \sum_{i_n=1}^{N_n} \sum_{i_n' \notin B_{he \left( \varphi^{(n)}_{i_n} \right)}} k \left( \frac{1}{h_n} \left\| \varphi^{(n)}_{i_n} - \varphi^{(n)}_{i_n'} \right\| \right) =: A_1 + A_2
\]
If the probability limit of $A_1$ or $A_2$ is strictly positive, then the sum is bounded below since both terms are weakly positive.

\[
A_1 \geq \frac{1}{N_n} \sum_{i_n=1}^{N_n} \sum_{i_n' \neq i_n} I \{ \tilde{\varphi}^{(n)}_{i_n} \in B_{he \left( \varphi^{(n)}_{i_n} \right)} \} k \left( \varepsilon \right).
\]
Set $\varepsilon$ low enough such that $k \left( \varepsilon \right) > 0$. Then plim $A_1 > 0$ by (A.5). $A_2 \geq 0$, hence no tighter lower bound needs to be shown for this term.

Since weights are in $[0, 1]$ and sum to 1 only the trace of $N_n^{-1} (\mathbb{I} - W_n)^2$ needs to be bounded away from 0 for (A.4) to hold. Using shorthand $K_{i_n', h}^{(n)} = k \left( \frac{1}{h_n} \left\| \varphi^{(n)}_{i_n} - \varphi^{(n)}_{i_n'} \right\| \right)$,
\[
\frac{1}{N_n} Tr \{(\mathbb{I} - W_n)^2\} = \frac{1}{N_n} \sum_{i_n} (1 - W_{n,i_n,i_n})^2 > 0 \quad \text{wpal.}
\]
if $\frac{1}{N_n} \sum_{i_n} \sum_{i_n' \neq i_n} W_{n,i'_n,i_n} > 0 \quad \text{wpal}$, which is equivalent to $\frac{1}{N_n} \sum_{i_n} \sum_{i_n' \neq i_n} K_{i_n', h}^{(n)} > 0 \quad \text{wpal}$.

The first part of the Lemma showed this, which shows (A.4) holds.

Lemma A.2 (Linear Combinations Bound). For the linear operator $\tilde{H}_n \in \mathbb{R}^{r_n \times L}$, with bounded eigenvalues, define linear transformations of $\varphi^{(n)}_{i_n}$ as $\tilde{\varphi}^{(n)}_{i_n} = \tilde{H}_n \varphi^{(n)}_{i_n}$. When $\text{rank} (\tilde{H}_n) = r_n < L$,
\[
\| \varphi^{(n)}_{i_n} - \tilde{\varphi}^{(n)}_{i_n} \|^2 \leq \lambda_{\text{max}} (\tilde{H}_n) \left( \| \varphi^{(n)}_{i_n} - \tilde{\varphi}^{(n)}_{i_n} \|^2 + \| \varphi^{(n)}_{i_n} - \tilde{\varphi}^{(n)}_{i_n} \|^2 + \| \tilde{\varphi}^{(n)}_{i_n} - \tilde{\varphi}^{(n)}_{i_n} \|^2 \right)
\]
where $\lambda_{\text{max}}$ is the maximum eigenvalue. When $\text{rank} (\tilde{H}_n) = r_n = L$ then $\tilde{H}_n$ is invertible and bounding $\| \varphi^{(n)}_{i_n} - \tilde{\varphi}^{(n)}_{i_n} \|^2$ by $\| \tilde{\varphi}^{(n)}_{i_n} - \tilde{\varphi}^{(n)}_{i_n} \|^2$ is straightforward.
A.3 Iterative Estimator

Assumption 6 can be significantly relaxed with an iterative fixed-effects projection. The following allows for a higher dimensional set of proxy measures used to form projection weights. The sampling restriction can be entry-wise, i.e. on scalars, rather than on the neighbourhoods of entire vectors. The procedure is general, and can use many weights, or even clusters, in place of the kernel weights used to describe the estimator below.

Every tensor has a higher-order singular value decomposition (HOSVD),

$$\mathcal{A} = \mathcal{S} \times (U^{(1)}, \ldots, U^{(d)})$$

where $\mathcal{S} \in \mathbb{R}^{n \times 1}$ and $U^{(n)} \in \mathbb{R}^{n \times n}$, where $r_n$ is the $n$-th component of the multilinear rank. The iterative estimator works as follows.

1. Obtain proxies for $U^{(n)}$, called $\hat{U}^{(n)}$ for each $n = 1, \ldots, d$. Use these to form weights, for $m_n = 1, \ldots, r_n$:

$$W_{nm_n, in_j} := \frac{k \left( \frac{1}{h_n} (\hat{U}^{(n)}_{in, m_n} - \hat{U}^{(n)}_{jn, m_n})^2 \right)}{\sum_{i_n=1}^{N_n} k \left( \frac{1}{h_n} (\hat{U}^{(n)}_{i_n, m_n} - \hat{U}^{(n)}_{i_n, m_n})^2 \right)} \quad \text{for each } n = 1, \ldots, d. \quad (A.6)$$

2. For each $m_n = 1, \ldots, r_n$ project out weighted means from $Y$ and $X_k$ for each $k = 1, \ldots, K$ along each dimension, $n = 1, \ldots, d$ as follows,

$$\tilde{A} = \mathcal{A} \times \left( \prod_{m_1=1}^{r_1} (I - W_{1m_1}), \ldots, \prod_{m_d=1}^{r_d} (I - W_{dm_d}) \right) \quad (A.7)$$

3. Perform pooled OLS of $\tilde{Y}$ on $\tilde{X}$. Call the estimator $\hat{\beta}_{IK}$, for iterative kernel.

Proof of Lemma A.2.

$$\|\varphi_{i_n}^{(n)} - \varphi_{i_n}^{(n)}\|^2 \leq 5 \|\varphi_{i_n}^{(n)} - \hat{H}'_{i_n} \varphi_{i_n}^{(n)}\|^2 + 5 \|\varphi_{i_n}^{(n)} - \hat{H}'_{i_n} \varphi_{i_n}^{(n)}\|^2 + 5 \|\hat{H}'_{i_n} \varphi_{i_n}^{(n)} - \hat{H}'_{i_n} \varphi_{i_n}^{(n)}\|^2$$

The first term from the right hand side, $\|\varphi_{i_n}^{(n)} - \hat{H}'_{i_n} \varphi_{i_n}^{(n)}\|^2 = \| (I - \hat{H}'_{i_n} \hat{H}_{i_n}) (\varphi_{i_n}^{(n)} - \varphi_{i_n}^{(n)}) \|^2 = 0$, from $(I - \hat{H}'_{i_n} \hat{H}_{i_n}) (\varphi_{i_n}^{(n)} - \varphi_{i_n}^{(n)}) \|^2 \leq \lambda_{\max}(I - \hat{H}'_{i_n} \hat{H}_{i_n}) \|\varphi_{i_n}^{(n)} - \varphi_{i_n}^{(n)}\|^2 = 1 \cdot \|\varphi_{i_n}^{(n)} - \varphi_{i_n}^{(n)}\|^2$. Likewise, the second term is 0. Finally, use the matrix norm bound to bound the three remaining terms by the maximum eigenvalue of $\hat{H}_{i_n}$ factored out from each term.

$\blacksquare$
From the representation in (A.7) the proof techniques from previous results follow straightforwardly. As is proposed in the main text, proxies $\hat{U}^{(n)}$ can be used to bound the estimation error of these proxies with respect to the true $U^{(n)}$. Greater care is needed for the fact Proposition A.1 in Bai (2009) is an up-to-rotations result that is not required in previous results, hence this is taken care of specifically in the proof.

**Assumption A.1** (Regularity of proxy measures). Define $M_n^h \left( \hat{U}_{i,n,m_n}^{(n)} \right)$ as

$$M_n^h \left( \hat{U}_{i,n,m_n}^{(n)} \right) := \sum_{j_n=1}^{N_n} \mathbb{I} \left( \hat{U}_{j,n,m_n}^{(n)} \in B_{he} \left( \hat{U}_{i,n,m_n}^{(n)} \right) \right),$$

where $B_{he}(x)$ is an he-neighbourhood around $x$. For each dimension $n$, for any $e > 0$, as $h \rightarrow 0$ such that $N_n h \rightarrow \infty$, $N_n^{-1} \sum_{i_n=1}^{N_n} M_n^h \left( \hat{U}_{i,n,m_n}^{(n)} \right) > 0$ wpa. 1.

**Proposition A.1** (Upper bound on iterative estimator). Let Assumption 5, Assumption 7 specifically for the iterative transformation, and Assumption A.1 hold. For the HOSVD of fixed-effects, let $\frac{1}{N_n} \sum_{i_n} \left\| U_{i,n}^{(n)} - H^{(n')} \hat{U}_{i,n}^{(n')} \right\|^2 = O_p(C_{n^*}^{-2})$ for $n^* \in \mathcal{M}$ with $C_{n^*}^{-2} \rightarrow 0$, and $\frac{1}{N_n} \sum_{i_n} \left\| U_{i,n'}^{(n)} - H^{(n')} \hat{U}_{i,n'}^{(n')} \right\|^2 = O_p(1)$ for $n' \notin \mathcal{M}$, where $\mathcal{M}$ is a non-empty subset of dimensions. $H^{(n)}$ is an invertible $r_n \times r_n$ matrix. Then,

$$\left\| \hat{\beta}_{IK} - \beta^0 \right\| = O_p \left( \prod_{n} r_n^{1/2} (C_n^{-1} + r_n^{1/2} h_n) \right) + O_p \left( \prod_{n=1}^{d} \frac{1}{\sqrt{N_n}} \right).$$

For $h_n \lesssim O(C_n^{-1} r_n^{-1/2})$ this reduces to

$$\left\| \hat{\beta}_{IK} - \beta^0 \right\| = \prod_{n} r_n^{1/2} O_p \left( \prod_{n^* \in \mathcal{M}} O_p \left( C_{n^*}^{-1} \right) \right) + O_p \left( \prod_{n=1}^{d} \frac{1}{\sqrt{N_n}} \right).$$

**Proof of Proposition A.1.**

$$\frac{1}{N} \sum_{i_1, \ldots, i_d} \sum_{m_1, \ldots, m_d} \sum_{s_{m_1, \ldots, m_d}^{r_1}} \cdots \sum_{s_{m_1, \ldots, m_d}^{r_d}} \prod_{m_1' = 1}^{r_1} \left( U_{i_1, m_1' - W_{1m_1', i_1}^{(1)} U_{m_1}^{(1)}} \right) \cdots \prod_{m_d' = 1}^{r_d} \left( U_{i_d, m_d' - W_{dm_d', i_d}^{(d)} U_{m_d}^{(d)}} \right)$$

$$= \frac{1}{N} \sum_{i_1, \ldots, i_d} \sum_{m_1, \ldots, m_d} \sum_{s_{m_1, \ldots, m_d}^{r_1}} \cdots \sum_{s_{m_1, \ldots, m_d}^{r_d}} \prod_{m_1' = 1}^{r_1} W_{1m_1', i_1}^{(1)} \left( U_{i_1, m_1' - U_{j_1, m_1}^{(1)}} \right) \cdots$$

$$= \frac{1}{N} \sum_{i_1, \ldots, i_d} \sum_{m_1, \ldots, m_d} \prod_{m_1' = 1}^{r_1} \sqrt{W_{1m_1', i_1j_1}} \cdots \prod_{m_d' = 1}^{r_d} \sqrt{W_{dm_d', i_dj_d}} \left( U_{i_d, m_d' - U_{j_d, m_d}^{(d)}} \right) \cdots$$

37
By Cauchy-Schwarz, this can be bound by

\[
\left( \frac{1}{N} \sum_{i_1, \ldots, i_d} \sum_{m_1} \cdots \sum_{m_d} \tilde{X}_{i_1, \ldots, i_d}^2 \sum_{j_1, \ldots, j_d} \prod_{m_1' = 1}^{r_1} W_{1,m_1',i_1,j_1} \cdots \prod_{m_d' = 1}^{r_d} W_{d,m_d',i_d,j_d} \right)^{1/2} \cdots (A.8)
\]

\[
\cdots \left( \frac{1}{N} \sum_{i_1, \ldots, i_d} \sum_{j_1, \ldots, j_d} \prod_{m_1'}^{r_1} S_{m_1',m_d',i_1,j_1} \prod_{m_d'}^{r_d} W_{1,m_1',i_1,j_1} (U_{i_1,m_1}^{(1)} - U_{j_1,m_1}^{(1)})^2 \cdots \right)^{1/2}
\]

Focus on the first term in (A.8) and note

\[
\sum_{j_1, \ldots, j_d} \prod_{m_1'}^{r_1} W_{1,m_1',i_1,j_1} \cdots \prod_{m_d'}^{r_d} W_{d,m_d',i_d,j_d} = \sum_{j_1} \prod_{m_1'}^{r_1} W_{1,m_1',i_1,j_1} \sum_{j_d} \prod_{m_d'}^{r_d} W_{d,m_d',i_d,j_d}
\]

such that for any \( n \),

\[
\sum_{j_n} \prod_{m_n'}^{r_1} W_{n,m_n',i_n,j_n} \leq \left( \sum_{j_n} W_{n,i_n,j_n}^2 \right)^{1/2} \left( \sum_{j_n} \prod_{m_n'} \frac{W_{n,m_n',i_n,j_n}^2}{W_{n,m_n',i_n,j_n}} \right)^{1/2} \]

(weights in \([0,1]\))

\[
\leq \left( \sum_{j_n} W_{n,i_n,j_n} \right)^{1/2} \left( \sum_{j_n} \prod_{m_n'} \frac{W_{n,m_n',i_n,j_n}}{W_{n,m_n',i_n,j_n}} \right)^{1/2} \]

(weights sum to 1)

\[
\leq 1 \cdot \left( \sum_{j_n} W_{n,i_n,j_n}^2 \right)^{1/2} \left( \sum_{j_n} \prod_{m_n' \neq \{1,2\}} \frac{W_{n,m_n',i_n,j_n}^2}{W_{n,m_n',i_n,j_n}} \right)^{1/2} \cdots
\]

which is bound by 1. Bounded second moments of \( \tilde{X} \) bounds the first term by \( O_p \left( \prod_n r_n^{1/2} \right) \).

For the second term in (A.8), singular values \( S_{m_1, \ldots, m_d} \) bounded in probability gives,

\[
\max_{m_1, \ldots, m_d} S_{m_1, \ldots, m_d}^2 \prod_{n=1}^{d} \frac{1}{N_n} \sum_{m_n} \sum_{j_n} \prod_{m_n' = 1}^{r_n} W_{nm_n',i_n,j_n} (U_{i_n,m_n}^{(n)} - U_{j_n,m_n}^{(n)})^2
\]

\[
= O_p(1) \prod_{n=1}^{d} \frac{1}{N_n} \sum_{m_n} \sum_{j_n} \prod_{m_n' = 1}^{r_n} W_{nm_n',i_n,j_n} (U_{i_n,m_n}^{(n)} - U_{j_n,m_n}^{(n)})^2
\]

38
Focus for each $n$, where the superscript notation in $H^{(n)}$ is suppressed for brevity,

$$
\sum_{m_n}^{r_n} \frac{1}{N_n} \sum_{i_n,j_n}^{r_n} \prod_{m_n' = 1}^{r_n} W_{nm_n', i_n j_n} (U_{i_n, m_n}^{(n)} - U_{j_n, m_n}^{(n)})^2 \leq \ldots
$$

$$
\ldots \leq \sum_{m_n}^{r_n} \frac{3}{N_n} \sum_{i_n,j_n}^{r_n} \prod_{m_n' = 1}^{r_n} W_{nm_n', i_n j_n} (U_{i_n, m_n}^{(n)} - H_{m_n}^{(n)} \hat{U}_{i_n}^{(n)})^2
$$

$$
+ \sum_{m_n}^{r_n} \frac{3}{N_n} \sum_{i_n,j_n}^{r_n} \prod_{m_n' = 1}^{r_n} W_{nm_n', i_n j_n} (U_{j_n, m_n}^{(n)} - H_{m_n}^{(n)} \hat{U}_{j_n}^{(n)})^2
$$

$$
+ \sum_{m_n}^{r_n} \frac{3}{N_n} \sum_{i_n,j_n}^{r_n} \prod_{m_n' = 1}^{r_n} W_{nm_n', i_n j_n} (H_{m_n}^{(n)} \hat{U}_{i_n}^{(n)} - H_{m_n}^{(n)} \hat{U}_{j_n}^{(n)})^2
$$

by Jensen’s inequality. The first term,

$$
\frac{3}{N_n} \sum_{i_n}^{r_n} \sum_{m_n}^{r_n} (U_{i_n, m_n}^{(n)} - H_{m_n}^{(n)} \hat{U}_{i_n}^{(n)})^2 \sum_{j_n}^{r_n} \prod_{m_n' = 1}^{r_n} W_{nm_n', i_n j_n} \leq \frac{3}{N_n} \sum_{i_n}^{r_n} ||U_{i_n}^{(n)} - H^{(n)} \hat{U}_{i_n}^{(n)}||^2 \cdot 1
$$

is $O_p(C_n^{-2})$. Likewise, the second term is $O_p(C_n^{-2})$.

The final term,

$$
\frac{3}{N_n} \sum_{i_n,j_n}^{r_n} \prod_{m_n' = 1}^{r_n} W_{nm_n', i_n j_n} \left( H_{m_n}^{(n)} (\hat{U}_{i_n}^{(n)} - \hat{U}_{j_n}^{(n)}) \right)^2 \leq \ldots
$$

$$
\ldots \leq \frac{3}{N_n} \sum_{i_n,j_n}^{r_n} \prod_{m_n' = 1}^{r_n} W_{nm_n', i_n j_n} ||\hat{U}_{i_n}^{(n)} - \hat{U}_{j_n}^{(n)}||^2 \sum_{m_n}^{r_n} ||H_{m_n}||^2
$$

$$
= O_p(r_n) \frac{1}{N_n} \sum_{i_n,j_n}^{r_n} \prod_{m_n' = 1}^{r_n} W_{nm_n', i_n j_n} \left( \hat{U}_{i_n, m_n}^{(n)} - \hat{U}_{j_n, m_n}^{(n)} \right)^2;
$$

from $H$ invertible. Since $\sum_{m_n}^{r_n} ||H_{m_n}||^2 = \sum_{m_n}^{r_n} \sum_{m_n'}^{r_n} H_{m_n, m_n'}^2 = \sum_{r=1}^{r_n} \sigma_r^2(H)$, where $\sigma_r^2(H)$ are singular values of $H$ (bounded from invertibility). Without loss, assume the same ordering of $m_n$ and $m_n'$ in the sum and product. Then, for each $m_n$,

$$
\frac{1}{N_n} \sum_{i_n,j_n}^{r_n} \prod_{m_n' = 1}^{r_n} W_{nm_n', i_n j_n} (\hat{U}_{i_n, m_n}^{(n)} - \hat{U}_{j_n, m_n}^{(n)})^2 = \frac{1}{N_n} \sum_{i_n,j_n}^{r_n} \prod_{m_n' = 1}^{r_n} W_{nm_n', i_n j_n} (\hat{U}_{i_n, m_n}^{(n)} - \hat{U}_{j_n, m_n}^{(n)})^2 \prod_{m_n' \neq m_n} W_{nm_n', i_n j_n}
$$

$$
\leq \frac{1}{N_n} \sum_{i_n,j_n}^{r_n} \prod_{m_n' = 1}^{r_n} W_{nm_n', i_n j_n} (\hat{U}_{i_n, m_n}^{(n)} - \hat{U}_{j_n, m_n}^{(n)})^2 \cdot 1
$$

As in the proof of Proposition 2, this final term is bound as $O_p(h_n^2)$, hence, the second term of (A.8) is $O_p(r_n h_n^2)$.

Thus, consistency is at the rate $O_p \left( \prod_n r_n^{1/2} \left( C_n^{-1} + r_n^{1/2} h_n \right) \right)$.