Main Effect Factor Models in High-Dimensional Matrix Time Series: Identification and Sparsity

Zetai Cen[*], Kaixin Liu[†], and Clifford Lam[‡]

Abstract

We propose a general identification framework for main effect factor models for matrix-valued time series. The classical sum-to-zero restriction on the row and column main effects is replaced by a broad class of shift-varying functions, which includes weighted averages, quantiles, and reference-unit anchors as special cases. We prove that the shift-varying property is necessary and sufficient for parameter identification, and we derive closed-form estimators under minimal assumptions. Asymptotic convergence rates of all estimated components are derived under weak, heterogeneous factor strengths. Building on this flexible identification, we address sparsity in the main effects by selecting a minimum-based shift-varying function that sets the smallest main effect to zero, and we introduce a doubly adaptive fused Lasso estimator that consistently recovers the true sparse and dense blocks. A modified Mallows's statistic is developed for tuning parameter selection, and the entire procedure is computationally practical via an equivalent generalized lasso formulation. Simulation experiments are performed under a variety of settings, showing our proposed methods work well. Two real applications on employment growth and economic indices are demonstrated to illustrate the empirical value of our approach.

Keywords and phrases: Parameter identification; Shift-varying functions; Sparse main effects; Matrix factor models; Weak factors.

[*]Equal contribution; corresponding author. School of Mathematics, University of Bristol. Email: zetai.cen@bristol.ac.uk. Supported by Engineering and Physical Sciences Research Council (EP/Z531327/1).
[†]Equal contribution. Department of Statistics, London School of Economics. Email: K.Liu31@lse.ac.uk.
[‡]Department of Statistics, London School of Economics. Email: C.Lam2@lse.ac.uk.

1 Introduction

With the rapid development of computational hardware and technology for big data, researchers obtain and analyze datasets that are ever larger in size and complexity. This leads to significant advances in the area of factor analysis, which is a useful tool in multivariate analysis with a wide range of applications in psychology (McCrae and John, 1992), biology (Hirzel et al., 2002; Hochreiter et al., 2006), economics and finance (Fama and French, 1993; Stock and Watson, 2002a,b), to name but a few areas. In particular, since the early work of Chamberlain and Rothschild (1983), approximate factor models have been well studied over the past few decades; see e.g. Bai and Ng (2002), Pan and Yao (2008), Lam et al. (2011), and the references therein.

Related literature and motivation. To facilitate interpretation of the estimated factor structure, one of the major solutions is through sparsity, which is not new in time series analysis. Sparsity can be imposed on the data covariance matrix (Bickel and Levina, 2008) and the noise covariance matrix (Fan et al., 2013), among many other methods. See also Section 1.4 in Uematsu and Yamagata (2023a) for a discussion on sparse principal components. More recently, researchers are interested in studying the sparsity in factor loadings (Freyaldenhoven, 2022; Uematsu and Yamagata, 2023a), which is a concept closely related to factor strength as discussed in Section 7 in Barigozzi and Hallin (2024). In fact, various forms of sparsity in factor loadings are discussed in the literature. An example is a multilevel/group factor model (Wang, 2008; Hu et al., 2025), where each observed vector $\mathbf{x}_t^s \in \mathbb{R}^{p_s}$ $(t = 1, \ldots, T,\ s = 1, \ldots, S)$ can be represented by $\mathbf{x}_t^s = \mathbf{A}^s \mathbf{g}_t + \mathbf{B}^s \mathbf{f}_t^s + \mathbf{e}_t^s$, with $\mathbf{g}_t$ and $\mathbf{f}_t^s$ called the global and group-specific factors respectively, $\mathbf{A}^s$, $\mathbf{B}^s$ their corresponding factor loading matrices, and $\mathbf{e}_t^s$ the noise. The model can be rewritten as

$$\begin{pmatrix} \mathbf{x}_t^1 \\ \mathbf{x}_t^2 \\ \vdots \\ \mathbf{x}_t^S \end{pmatrix} = \begin{pmatrix} \mathbf{A}^1 & \mathbf{B}^1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{A}^2 & \mathbf{0} & \mathbf{B}^2 & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}^S & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{B}^S \end{pmatrix} \begin{pmatrix} \mathbf{g}_t \\ \mathbf{f}_t^1 \\ \vdots \\ \mathbf{f}_t^S \end{pmatrix} + \begin{pmatrix} \mathbf{e}_t^1 \\ \mathbf{e}_t^2 \\ \vdots \\ \mathbf{e}_t^S \end{pmatrix},$$

so that the factor loading matrix has a specific sparse structure, which is based on a priori data grouping. Another similar example of sparse loadings is models of the above kind, except that the membership of the data is unknown; see Ando and Bai (2017) and Zhang et al. (2024). In addition to block sparsity due to grouping, Uematsu and Yamagata (2023a,b) studies a sparsity-induced weak factor model under rather restricted conditions to address the identification issue that sparse factor loadings are generally not rotation-invariant. Wei and Zhang (2024) further investigates the near-sparsity preservation property of the estimated loadings. Other related examples include Fan et al. (2023) which relates sparsity

and factor models through factor-augmented regression, and Mosley et al. (2024) which focuses on sparsity in the loading matrix of dynamic factor models. For Bayesian methods on sparsity in factor structures, see Zhang et al. (2025) and the references therein.

Despite the growing literature from the above discussion, all of them focus only on vector-valued time series. As first proposed by Wang et al. (2019) and later studied in broader literature (e.g. Yu et al., 2022; He et al., 2024), matrix factor models have become powerful tools for studying economic and financial data, leading to more insightful interpretations and improved estimations. In particular, a series of matrices $\mathbf{X}_t$, $t \in \{1, \dots, T\} =: [T]$, is observed and admits the decomposition $\mathbf{X}_t = \mathbf{A}_r \mathbf{F}_t \mathbf{A}_c^\intercal + \mathbf{E}_t$, where $\mathbf{A}_r, \mathbf{A}_c$ are the row and column factor loading matrices respectively, and $\mathbf{F}_t$ is the core factor matrix. Although one may write $\mathbf{vec}(\mathbf{X}_t) = (\mathbf{A}_c \otimes \mathbf{A}_r)\mathbf{vec}(\mathbf{F}_t) + \mathbf{vec}(\mathbf{E}_t)$, where $\mathbf{vec}(\mathbf{X}_t)$ is the stacking of the columns of $\mathbf{X}_t$ into a vector, and thus estimate a factor model for the vectorized data with sparsity, such an approach is neither efficient nor appropriate. For instance, the global inference step in Uematsu and Yamagata (2023b) is now inapplicable as it neglects the Kronecker product structure in the factor loading matrix.

This paper provides an alternative solution to address sparsity in matrix factor models. It is worth pointing out that it remains a challenge to address the interaction between rows and columns in factor modeling for matrix-valued time series, for which very limited work has been done. An example is He et al. (2023) where a specification test is proposed. Another attempt is by Lam and Cen (2025), which generalizes the Tucker-decomposition factor model to a Main Effect Factor Model (MEFM), significantly improving model interpretability. In particular, given a sequence of mean-zero matrix observations $\mathbf{X}_t \in \mathbb{R}^{p \times q}$ for $t = 1, \dots, T$, MEFM decomposes each $\mathbf{X}_t$ as

$$\mathbf{X}_t = \mu_t \mathbf{1}_p \mathbf{1}_q^\intercal + \boldsymbol{\alpha}_t^* \mathbf{1}_q^\intercal + \mathbf{1}_p \boldsymbol{\beta}_t^{*\intercal} + \mathbf{C}_t + \mathbf{E}_t, \tag{1.1}$$

where $\mu_t \in \mathbb{R}$ is the grand mean, $\boldsymbol{\alpha}_t^* \in \mathbb{R}^p$ and $\boldsymbol{\beta}_t^* \in \mathbb{R}^q$ are respectively the row and column main effects with potentially many zero entries, $\mathbf{C}_t = \mathbf{A}_r \mathbf{F}_t \mathbf{A}_c^\intercal$ is the common component, $\mathbf{A}_r \in \mathbb{R}^{p \times k_r}$ and $\mathbf{A}_c \in \mathbb{R}^{q \times k_c}$ are the row and column factor loading matrices, $\mathbf{F}_t$ is the core factor matrix, and $\mathbf{E}_t$ is the noise. In particular, the identification of parameters, e.g. main effects and factor loadings, relies heavily on the condition

$$\mathbf{1}_p^\intercal \boldsymbol{\alpha}_t^* = 0, \quad \mathbf{1}_q^\intercal \boldsymbol{\beta}_t^* = 0, \quad \mathbf{1}_p^\intercal \mathbf{A}_r = \mathbf{0}, \quad \mathbf{1}_q^\intercal \mathbf{A}_c = \mathbf{0}. \tag{1.2}$$

Lam and Cen (2025) shows that any Tucker-decomposition matrix factor model can be rewritten into MEFM, whereas the converse generally requires far more factors and can still be empirically inferior. Note that for any fixed row (resp. column), the contribution of

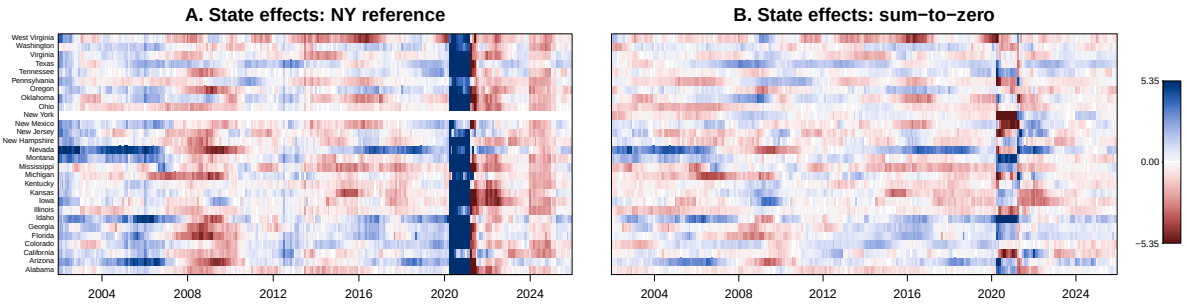

Figure 1: QCEW state effects under identifications of *NY reference* (panel A) and *sum-to-zero* (panel B). Color scales are symmetrically clipped at the 98th percentile of absolute effect magnitudes for readability.

$\boldsymbol{\alpha}_t^*$ (resp. $\boldsymbol{\beta}_t^*$) towards the observed $\mathbf{X}_t$ is constant across columns (resp. rows). Hence row or column structures are solely featured by the main effects, and the interaction between rows and columns by the common component.

However, the identification condition (1.2) gives rise to complications when sparsity of the main effects is desired. For instance, suppose $\mathbf{X}_t$ records the values of economic indicators with countries indexed by rows and indicators indexed by columns. If a small group of countries form an economic entity with pervasive effects on each indicator of the member countries, then the row main effects vector $\boldsymbol{\alpha}_t^*$ is sparse with non-zero entries corresponding to those member countries. Yet (1.2) implies that either some member countries have opposite effects which are unnatural, or some non-member countries have non-zero effects which are not realistic under this scenario. To resolve this, we propose a flexible class of identification conditions based on *shift-varying* functions. Such a function satisfies that adding a constant to its argument strictly changes its value; typical examples include weighted sums (yielding the sum-to-zero condition as a special case) and quantile functions (e.g., anchoring the main effects at their minimum or median). By choosing a shift-varying function that sets a particular quantile to zero, one can naturally accommodate sparse and interpretable main effects, while the sum-to-zero condition becomes only one of many possibilities. This new identification significantly enhances the interpretability of the model and is the foundation of our sparsity recovery methodology.

As an empirical example, consider the monthly employment data from the U.S. Bureau of Labor Statistics Quarterly Census of Employment and Wages (QCEW). The dataset contains $T = 288$ months from 2002-M01 to 2025-M12, and at each month t, $\mathbf{X}_t \in \mathbb{R}^{28\times 17}$ denotes the employment growth for 28 states and 17 non-overlapping private-sector industries; see Section 5.2 for detailed investigations and we only highlight the value of flexible identifications here. The row main effects $\boldsymbol{\alpha}_t^* \in \mathbb{R}^{28}$, namely the state effects, are estimated under two identifications for all $t \in [288]$: *NY reference* where the element of $\boldsymbol{\alpha}_t^*$

corresponding to New York is zero; *sum-to-zero* such that $\mathbf{1}_{28}^{\intercal}\boldsymbol{\alpha}_t^* = 0$ as in Lam and Cen (2025). While the interpretation of *sum-to-zero* is not straightforward, the choice of *NY reference* is already meaningful such that the state effects represent the *excess* employment growth of each state with respect to the growth in New York. Later, we will show that *NY reference* is a valid identification condition for the main effects. We also show in Figure 1 the estimated effects under these two settings. The comparison is sheer around the pandemic period from 2020: *NY reference* clearly shows excess employment growth in all other states, reflecting the significant employment crash in New York, which is vague under *sum-to-zero*. From *NY reference*, the labor market in New York also seems to experience a post-pandemic rebound during 2024, which again shows the benefits of considering such identification.

Contributions and organization. Built upon the existing MEFM framework, we develop a general theory for its identification that allows researchers to anchor the row and column main effects via any shift-varying function. This class includes the sum-to-zero restriction of Lam and Cen (2025) as a special case, as well as quantile-based and reference-unit (viz. a particular cross-sectional unit's effect is fixed to be zero) identifications that are far better flexible and interpretable in different scenarios, and are suited to possibly sparse or tail-heavy data. We show that the shift-varying condition is necessary and sufficient for parameter identification, derive closed-form estimators, and establish their asymptotic properties under weak, heterogeneous factor strengths. As a first in the literature, we consider MEFM with sparsity in the main effects and develop methods to recover the structure for more natural and better interpretation of the main effects over time. A special (and sparsest) case is the traditional matrix factor model where all main effects are zero. A period of non-zero main effects in relation to other periods of zero main effects can indicate a significant change in circumstances for the rows or columns in question. By employing a quantile-based identification (e.g., zero-minimum constraint), we can directly interpret zero main effects as the absence of row- or column-specific deviations, and the sparsity pattern itself becomes a quantity of interest.

The rest of this paper is organized as follows. The new model identification is detailed in Section 2, followed by parameter estimation. Section 3 presents the main assumptions and theoretical results of the estimators. Section 4 develops regularized estimation for sparse main effects and studies the asymptotic properties of the proposed method for sparsity recovery. We also discuss the implementation of the algorithm and hyperparameter tuning based on a modified C_p statistic. Lastly, numerical results are shown in Section 5, where two real data applications are also analyzed. Section 6 concludes the paper. All theoretical proofs and additional information are relegated to the supplement.

Notations. We use the lower-case or capital letter, bold lower-case letter, and bold capital letter, i.e., a or A, $\mathbf{a}$, $\mathbf{A}$, to denote a scalar, a vector, and a matrix respectively. We also

use $a_i, A_{i,j}, \mathbf{A}_{i\cdot}, \mathbf{A}_{\cdot i}$ to denote, respectively, the i-th element of $\mathbf{a}$, the (i,j)-th element of $\mathbf{A}$, the i-th row vector (as a column vector) of $\mathbf{A}$, and the i-th column vector of $\mathbf{A}$. We use $\circ$ to denote the Hadamard product and $\otimes$ the Kronecker product; $a \asymp b$ represents $a = O(b)$ and $b = O(a)$; $a \vee b := \max(a, b)$, and $a \wedge b := \min(a, b)$. For a vector $\mathbf{a}$, we write $\min\{\mathbf{a}\}$, $\mathrm{med}\{\mathbf{a}\}$, and $\max\{\mathbf{a}\}$ for its minimum, median, and maximum, respectively. A random variable X is sub-Gaussian with variance proxy σ^2, denoted as $X \sim \mathrm{subG}(\sigma^2)$, if $\mathbb{E}[\exp\{s(X-\mathbb{E}[X])\}] \leq \exp(s^2\sigma^2/2)$ for all $s \in \mathbb{R}$. A random variable X is sub-exponential with parameter λ, denoted as $X \sim \mathrm{subE}(\lambda)$, if $\mathbb{E}[\exp\{s(X - \mathbb{E}[X])\}] \leq \exp(s^2\lambda^2/2)$ for all $|s| \leq 1/\lambda$. Given a positive integer a, define $[a] := \{1, \ldots, a\}$. We define $\mathbf{1}_a$ as a vector of ones with length a, and $\mathbf{M}_a = \mathbf{I}_a - a^{-1}\mathbf{1}_a\mathbf{1}_a^\intercal$. Given a matrix $\mathbf{A}$, $\lambda_i(\mathbf{A})$ denotes its i-th largest eigenvalue, $\mathbf{A}^\intercal$ denotes its transpose, and without loss of generality, the eigenvalues of a matrix are arranged by descending orders, and so are their corresponding eigenvectors. We use $\mathrm{diag}(\{\mathbf{A}_1, \ldots, \mathbf{A}_n\})$ to represent the block diagonal matrix with $\{\mathbf{A}_1, \ldots, \mathbf{A}_n\}$ on the diagonal. We denote by $|\cdot|$ the cardinality of a set, $\|\cdot\|$ the spectral norm of a matrix or the L_2 norm of a vector, and $\|\cdot\|_F$ the Frobenius norm of a matrix. Lastly, $\|\cdot\|_1$ and $\|\cdot\|_\infty$ denote the L_1 and L_∞ norms of a matrix or a vector, respectively.

2 Model and Estimation

2.1 Identification of main effect factor models

For a matrix time series admitting the MEFM representation depicted in (1.1), we investigate thoroughly in this subsection the identification of main effects. To start with, we say a function $g(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}$ is *shift-varying* at c_0 if for any vector $\mathbf{a} \in \mathbb{R}^d$ with $g(\mathbf{a}) = c_0$, we have $g(\mathbf{a} + c_1\mathbf{1}_d) \neq c_0$ for any constant $c_1 \neq 0$. We write in short $g(\cdot)$ is shift-varying if $g(\cdot)$ is shift-varying at any real number. The following condition serves to identify parameters in model (1.1).

(IC1) (Identification). *We assume: (i)* $\mathbf{A}_r^\intercal\mathbf{1}_p = \mathbf{0}$ *and* $\mathbf{A}_c^\intercal\mathbf{1}_q = \mathbf{0}$*; (ii) Given functions* $g_\alpha(\cdot) : \mathbb{R}^p \mapsto \mathbb{R}$ *and* $g_\beta(\cdot) : \mathbb{R}^q \mapsto \mathbb{R}$*, we have* $g_\alpha(\boldsymbol{\alpha}_t^*) = 0$ *and* $g_\beta(\boldsymbol{\beta}_t^*) = 0$ *for any* $t \in [T]$*; and (iii) The functions* $g_\alpha(\cdot)$ *and* $g_\beta(\cdot)$ *are shift-varying at zero.*

Note immediately that for any constant c, the function $f(\cdot) = g(\cdot) + c$ is shift-varying at $c_0 + c$ if and only if $g(\cdot)$ is shift-varying at c_0. Hence in parts (ii)–(iii) of (IC1), it is equivalent to assume $g_\alpha(\boldsymbol{\alpha}_t^*) = c_\alpha$ and $g_\beta(\boldsymbol{\beta}_t^*) = c_\beta$, with $g_\alpha(\cdot)$ and $g_\beta(\cdot)$ being shift-varying at some constants c_α and c_β, respectively. Hence Assumption (IC1) is stated with $c_\alpha = c_\beta = 0$, without loss of generality. In fact, even the shift-varying functions and the corresponding constants can be time-varying at the cost of heavier notations, and we do not pursue this line of analysis.

Notably, the class of shift-varying functions encapsulates a large, useful set of functions, thus making the identified MEFM framework sufficiently general. The functions g_α and g_β are pre-specified by researchers to reflect any prior knowledge on the matrix time series. In what follows, we discuss two important types of shift-varying functions in terms of g_α, together with examples of the practical interpretation of the identified row main effects by Assumption (IC1).

1. (Most linear functions). Given $\mathbf{x} \in \mathbb{R}^p$, define the set of functions

$$\mathcal{L} := \Big\{ g(\mathbf{x}) = \mathbf{v}^\intercal \mathbf{x} - c : \mathbf{v} \in \mathbb{R}^p \text{ with } \mathbf{1}_p^\intercal \mathbf{v} \neq 0,\ c \in \mathbb{R} \Big\}. \tag{2.1}$$

   Then all functions in $\mathcal{L}$ are shift-varying. With non-negative entries in $\mathbf{v}$, the condition in Assumption (IC1) can be viewed as imposing a level c on the weighted average of the main effects. The identification strategy in Lam and Cen (2025) is a special case of this with $\mathbf{v} = \mathbf{1}_p$ and $c = 0$. Note that an extreme case can be $g_\alpha(\boldsymbol{\alpha}_t^*) = \alpha_{t,i}^*$, that is, the identification of the row main effects is anchored at the i-th cross-sectional unit. This can be of particular interest when researchers wish to fix a reference unit and measure all other units relative to this baseline, and the main effects can be interpreted as deviations from the chosen benchmark.

2. (All quantile functions). With $q_u(\mathbf{x})$ representing the u-th quantile of elements in a vector $\mathbf{x}$, define the set of functions

$$\mathcal{Q} := \Big\{ g(\mathbf{x}) = q_u(\mathbf{x}) - c : u \in [0, 1], c \in \mathbb{R} \Big\}. \tag{2.2}$$

   Then all functions in $\mathcal{Q}$ are shift-varying. For $u = 0$, 0.5, and 1, the identification on the row main effects can be read as $\min\{\boldsymbol{\alpha}_t^*\} = c$, $\text{med}\{\boldsymbol{\alpha}_t^*\} = c$, and $\max\{\boldsymbol{\alpha}_t^*\} = c$, respectively. The two extrema setups naturally accommodate the empirical range of the observed time series, such as non-negative values for metric reading in monitoring river flows in ecological application. On the other hand, main effects identified based on median function could serve as a tail-robust framework allowing for outliers, which are typical for financial return data.

From the above discussion, our model design can be significantly flexible compared to MEFM in Lam and Cen (2025). While the above examples hold similarly for the column main effects, it is worth pointing out that the identification design for the column main effects can be different from that for the row main effects, depending upon the data sets in practice and the purpose of analysis. Moreover, regardless of how the main effects are identified in (IC1), they always imply that, e.g. from $\boldsymbol{\alpha}_t^* \mathbf{1}_q^\intercal$, the contribution of $\boldsymbol{\alpha}_t^*$ towards

$\mathbf{X}_t$ only varies over different rows and remains the same over columns. Hence the main effects specialize in row-wise and column-wise contribution, while the common component $\mathbf{C}_t$ picks up the interaction between rows and columns. It is worth pointing out that the grand mean μ_t should be understood as the remainder from row and column main effects under the specified identification.

In all previous works on MEFM, the grand mean μ_t is first identified via the identification condition (1.2), so that the row and column main effects can be subsequently identified. The same argument is inadmissible under our Assumption (IC1), so we take another route by identifying the main effects first. In fact, by our new identification strategy, we can show that part (iii) of Assumption (IC1) is a necessary and sufficient condition to identify the main effects. Additionally, for a certain class of functions specified in Assumption (IC1), model (1.1) is a more general framework than the Tucker-decomposition matrix factor model. We formalize all such results in the following theorem.

Theorem 1. *For model (1.1), we have the following.*

(i) Under Assumption (IC1) (i)–(ii), the parameters μ_t, $\boldsymbol{\alpha}_t^*$, $\boldsymbol{\beta}_t^*$ and $\mathbf{C}_t$ can be identified if and only if (IC1) (iii) holds.

(ii) Let $g_\alpha(\cdot)$ and $g_\beta(\cdot)$ be the functions from Assumption (IC1). Further assume $g_\alpha(\cdot)$ satisfies that for any $\mathbf{a} \in \mathbb{R}^p$, there exists some scalar c such that $g_\alpha(\mathbf{a} - c\mathbf{1}_p) = 0$; and assume similarly for $g_\beta(\cdot)$. If $\mathbf{X}_t$ follows a Tucker-decomposition matrix factor model such that $\mathbf{X}_t = \acute{\mathbf{A}}_r \mathbf{F}_t \acute{\mathbf{A}}_c^\intercal + \mathbf{E}_t$ with arbitrary factor loading matrices $\acute{\mathbf{A}}_r$ and $\acute{\mathbf{A}}_c$, then $\mathbf{X}_t$ also follows model (1.1) with the resulting parameters satisfying Assumption (IC1).

Theorem 1 (i) shows that shift-varying functions fully characterize the identifiability of the particular structure in MEFM. In particular, the grand mean is identified on top of the main effects, which is different from Lam and Cen (2025) where the grand mean can be directly identified. Part (ii) of the above theorem generalizes their Remark 1. We emphasize that the additional requirement on functions g_α and g_β in Theorem 1 (ii) is very mild, e.g. it can be fulfilled by all the functions in (2.1) and (2.2).

2.2 Parameter estimation

In this subsection, we discuss the procedure of estimating parameters in model (1.1). First, we state another condition to facilitate estimation.

(IC2) (Closed-form solution). *Let $g_\alpha(\cdot)$ and $g_\beta(\cdot)$ be the shift-varying functions from Assumption (IC1). For any $\mathbf{x} \in \mathbb{R}^p$, the equation $g_\alpha(\mathbf{x} - c\mathbf{1}_p) = 0$ has a closed-form solution $c = \widetilde{g}_\alpha(\mathbf{x})$ for some known function $\widetilde{g}_\alpha(\cdot)$; and assume similarly for $g_\beta(\cdot)$, with the solution function denoted as $\widetilde{g}_\beta(\cdot)$.*

Assumption (IC2) requires the existence of solution for the equation in Theorem 1 (ii), and that the solution can be computed explicitly. While not required, note that the solution must be unique by the definition of shift-varying functions and Assumption (IC1) (iii). It is also straightforward to check that all functions in $\mathcal{L}$ and $\mathcal{Q}$ fulfill Assumption (IC2); we present their solution functions at the end of this section.

Estimation of grand mean and main effects. Fix any $t \in [T]$. We first detail the estimation procedure for row main effects. Right-multiplying $\mathbf{1}_q$ on $\mathbf{X}_t$ in (1.1), by Assumption (IC1) (i) we have

$$\mathbf{X}_t\mathbf{1}_q = \mathbf{1}_p(q\mu_t + \mathbf{1}_q^\intercal\boldsymbol{\beta}_t^*) + q\boldsymbol{\alpha}_t^* + \mathbf{E}_t\mathbf{1}_q,$$

which, together with Assumption (IC1) (ii), gives

$$0 = g_\alpha\left(\boldsymbol{\alpha}_t^*\right) = g_\alpha\left(q^{-1}\mathbf{X}_t\mathbf{1}_q - q^{-1}\mathbf{E}_t\mathbf{1}_q - \mathbf{1}_p(\mu_t + q^{-1}\mathbf{1}_q^\intercal\boldsymbol{\beta}_t^*)\right).$$

Then by Assumption (IC2), we have $(\mu_t + q^{-1}\mathbf{1}_q^\intercal\boldsymbol{\beta}_t^*) = \widetilde{g}_\alpha(q^{-1}\mathbf{X}_t\mathbf{1}_q - q^{-1}\mathbf{E}_t\mathbf{1}_q)$, and hence

$$\boldsymbol{\alpha}_t^* = q^{-1}\mathbf{X}_t\mathbf{1}_q - \mathbf{1}_p \cdot \widetilde{g}_\alpha(q^{-1}\mathbf{X}_t\mathbf{1}_q - q^{-1}\mathbf{E}_t\mathbf{1}_q) - q^{-1}\mathbf{E}_t\mathbf{1}_q.$$

Therefore, a feasible estimator for the row main effect at t can be

$$\widetilde{\boldsymbol{\alpha}}_t := q^{-1}\mathbf{X}_t\mathbf{1}_q - \mathbf{1}_p \cdot \widetilde{g}_\alpha(q^{-1}\mathbf{X}_t\mathbf{1}_q). \tag{2.3}$$

Analogous arguments yield the estimator for the column main effect

$$\widetilde{\boldsymbol{\beta}}_t := p^{-1}\mathbf{X}_t^\intercal \mathbf{1}_p - \mathbf{1}_q \cdot \widetilde{g}_\beta(p^{-1}\mathbf{X}_t^\intercal \mathbf{1}_p). \tag{2.4}$$

Next, we estimate the grand mean at time $t \in [T]$. This is straightforward using the row and column main effect estimators. To this end, by left-multiplying $\mathbf{1}_p^\intercal$ and right-multiplying $\mathbf{1}_q$ on $\mathbf{X}_t$, we have

$$\mathbf{1}_p^\intercal \mathbf{X}_t \mathbf{1}_q = pq\mu_t + q\mathbf{1}_p^\intercal \boldsymbol{\alpha}_t^* + p\mathbf{1}_q^\intercal \boldsymbol{\beta}_t^* + \mathbf{1}_p^\intercal \mathbf{E}_t \mathbf{1}_q.$$

This leads to the grand mean estimator

$$\widetilde{\mu}_t := (pq)^{-1}\mathbf{1}_p^\intercal \mathbf{X}_t \mathbf{1}_q - p^{-1}\mathbf{1}_p^\intercal \widetilde{\boldsymbol{\alpha}}_t - q^{-1}\mathbf{1}_q^\intercal \widetilde{\boldsymbol{\beta}}_t. \tag{2.5}$$

Estimation of factor structure. Note that due to rotational indeterminacy, we cannot identify the loading matrices $\mathbf{A}_r$ and $\mathbf{A}_c$ exactly, but only their column spaces. To take into account potentially heterogeneous weak factors (e.g. Lam and Yao, 2012; Zhang et al., 2024), we normalize the loadings and the core factor as $\mathbf{Q}_r = \mathbf{A}_r \mathbf{Z}_r^{-1/2}$, $\mathbf{Q}_c = \mathbf{A}_c \mathbf{Z}_c^{-1/2}$, and $\mathbf{F}_{Z,t} = \mathbf{Z}_r^{1/2}\mathbf{F}_t\mathbf{Z}_c^{1/2}$, with $\mathbf{Z}_r$ and $\mathbf{Z}_c$ from Assumption (L1); see more details in Section 3.1. Since $\mathbf{C}_t = \mathbf{Q}_r \mathbf{F}_{Z,t} \mathbf{Q}_c^\intercal$, we may equivalently estimate the normalized parameters. To this end, define the matrix

$$\widetilde{\mathbf{L}}_t := \mathbf{X}_t - \widetilde{\mu}_t \mathbf{1}_p \mathbf{1}_q^\intercal - \widetilde{\boldsymbol{\alpha}}_t \mathbf{1}_q^\intercal - \mathbf{1}_p \widetilde{\boldsymbol{\beta}}_t^\intercal. \tag{2.6}$$

Then the estimator for the normalized row loading matrix, denoted by $\widehat{\mathbf{Q}}_r$, is defined as the eigenvector matrix corresponding to the k_r largest eigenvalues of the matrix $T^{-1}\sum_{t=1}^T \widetilde{\mathbf{L}}_t \widetilde{\mathbf{L}}_t^\intercal$. Similarly, the normalized column loading matrix estimator $\widehat{\mathbf{Q}}_c$ is the eigenvector matrix corresponding to the k_c largest eigenvalues of $T^{-1}\sum_{t=1}^T \widetilde{\mathbf{L}}_t^\intercal \widetilde{\mathbf{L}}_t$. Lastly, the core factor and the common component can be estimated by

$$\widehat{\mathbf{F}}_{Z,t} := \widehat{\mathbf{Q}}_r^\intercal \widetilde{\mathbf{L}}_t \widehat{\mathbf{Q}}_c, \quad \widehat{\mathbf{C}}_t := \widehat{\mathbf{Q}}_r \widehat{\mathbf{F}}_{Z,t} \widehat{\mathbf{Q}}_c^\intercal = \widehat{\mathbf{Q}}_r \widehat{\mathbf{Q}}_r^\intercal \mathbf{X}_t \widehat{\mathbf{Q}}_c \widehat{\mathbf{Q}}_c^\intercal,$$

where the last equality uses the fact that $\widehat{\mathbf{Q}}_r^\intercal \widetilde{\mathbf{L}}_t \widehat{\mathbf{Q}}_c = \widehat{\mathbf{Q}}_r^\intercal \mathbf{X}_t \widehat{\mathbf{Q}}_c$, since both $\widehat{\mathbf{Q}}_r$ and $\widehat{\mathbf{Q}}_c$ fulfill Assumption (IC1) (i) with $\mathbf{A}_r^\intercal \mathbf{1}_p = \mathbf{0}$ and $\mathbf{A}_c^\intercal \mathbf{1}_q = \mathbf{0}$ (see Lemma B.3 in the supplement).

Main effect estimators under identification of $\mathcal{L}$ and $\mathcal{Q}$, respectively. Lastly, we give explicit forms of the main effect estimators in (2.3) and (2.4), when the shift-varying functions in Assumption (IC1) are specified from $\mathcal{L}$ and $\mathcal{Q}$, respectively. To this end, first note that for any shift-varying functions in $\mathcal{L}$, using the notations in (2.1), their solution functions and thus the estimator in (2.3) take the form

$$\begin{aligned} \widetilde{g}(\mathbf{x}) &= (\mathbf{v}^\intercal \mathbf{1}_p)^{-1} (\mathbf{v}^\intercal \mathbf{x} - c), \\ \widetilde{\boldsymbol{\alpha}}_t &= q^{-1}\mathbf{X}_t \mathbf{1}_q - \mathbf{1}_p \cdot (\mathbf{v}^\intercal \mathbf{1}_p)^{-1} \left(q^{-1}\mathbf{v}^\intercal \mathbf{X}_t \mathbf{1}_q - c\right). \end{aligned} \tag{2.7}$$

For instance, when $\mathbf{v} = \mathbf{1}_p$ and $c = 0$, the above main effect estimator simplifies to

$$\widetilde{\boldsymbol{\alpha}}_t = q^{-1}\mathbf{X}_t \mathbf{1}_q - (pq)^{-1}\mathbf{1}_p \mathbf{1}_p^\intercal \mathbf{X}_t \mathbf{1}_q = q^{-1}(\mathbf{I}_p - p^{-1}\mathbf{1}_p \mathbf{1}_p^\intercal)\mathbf{X}_t \mathbf{1}_q = q^{-1}\mathbf{M}_p \mathbf{X}_t \mathbf{1}_q,$$

which matches Equations (3.2)–(3.3) in Lam and Cen (2025).

On the other hand, consider shift-varying functions from $\mathcal{Q}$. With the notations in (2.2), we can write the solution function and the corresponding estimator respectively as

$$\begin{aligned} \widetilde{g}(\mathbf{x}) &= q_u(\mathbf{x}) - c, \\ \widetilde{\boldsymbol{\alpha}}_t &= q^{-1}\mathbf{X}_t \mathbf{1}_q - q^{-1}\mathbf{1}_p \cdot q_u(\mathbf{X}_t \mathbf{1}_q) + c\mathbf{1}_p. \end{aligned} \tag{2.8}$$

Suppose $c = 0$. When $u = 0.5$, i.e., $q_u(\cdot)$ takes the median of a vector, the main effect estimator reads as $\widetilde{\boldsymbol{\alpha}}_t = q^{-1}\mathbf{X}_t \mathbf{1}_q - q^{-1}\mathbf{1}_p \cdot \text{med}\{\mathbf{X}_t \mathbf{1}_q\}$. Similarly, if we set $u = 0$, then the main effect estimator is $\widetilde{\boldsymbol{\alpha}}_t = q^{-1}\mathbf{X}_t \mathbf{1}_q - q^{-1}\mathbf{1}_p \cdot \min\{\mathbf{X}_t \mathbf{1}_q\}$. Markedly, given $c = 0$, model identification based on any shift-varying functions from $\mathcal{Q}$ indicates that the main effects vanish at a certain quantile. This allows practitioners to study the sparsity pattern on main effects, which can be of particular interest; we defer the detailed discussion in Section 4.

3 Assumptions and Theoretical Results

3.1 Assumptions

We discuss the main theoretical results under general identification based on Assumptions (IC1) and (IC2). For simplicity, throughout this paper, we assume the shift-varying

functions for the main effects have the same functional form which is generically denoted by $g(\cdot)$. To begin with, we present additional assumptions in the following to characterize model (1.1). The explanations to each assumption are deferred to the end of this subsection.

(L1) (Factor strength). *We assume that* $\mathbf{A}_r$ *and* $\mathbf{A}_c$ *are of full rank and independent of* $\{\mathbf{F}_t\}$ *and* $\{\mathbf{E}_t\}$. *Furthermore, as* $p, q \to \infty$,

$$\mathbf{Z}_r^{-1/2}\mathbf{A}_r^\intercal\mathbf{A}_r\mathbf{Z}_r^{-1/2} \to \mathbf{\Sigma}_{A,r}, \quad \mathbf{Z}_c^{-1/2}\mathbf{A}_c^\intercal\mathbf{A}_c\mathbf{Z}_c^{-1/2} \to \mathbf{\Sigma}_{A,c}, \tag{3.1}$$

where $\mathbf{Z}_r = \mathrm{diag}(\mathbf{A}_r^\intercal\mathbf{A}_r)$, $\mathbf{Z}_c = \mathrm{diag}(\mathbf{A}_c^\intercal\mathbf{A}_c)$, *and both* $\mathbf{\Sigma}_{A,r}$ *and* $\mathbf{\Sigma}_{A,c}$ *are positive definite with all eigenvalues bounded away from 0 and infinity. We assume* $(\mathbf{Z}_r)_{jj} \asymp p^{\delta_{r,j}}$ *for* $j \in [k_r]$ *and* $1/2 < \delta_{r,k_r} \leq \cdots \leq \delta_{r,2} \leq \delta_{r,1} \leq 1$. *Similarly, we assume* $(\mathbf{Z}_c)_{jj} \asymp q^{\delta_{c,j}}$ *for* $j \in [k_c]$, *with* $1/2 < \delta_{c,k_c} \leq \cdots \leq \delta_{c,2} \leq \delta_{c,1} \leq 1$.

(F1) (Time series in $\mathbf{F}_t$). *There is* $\mathbf{X}_{f,t}$ *the same dimension as* $\mathbf{F}_t$, *such that* $\mathbf{F}_t = \sum_{w\geq 0} a_{f,w}\mathbf{X}_{f,t-w}$. *The time series* $\{\mathbf{X}_{f,t}\}$ *has i.i.d. elements with mean* 0 *and variance* 1, *with uniformly bounded fourth order moments. The coefficients* $a_{f,w}$ *are such that* $\sum_{w\geq 0} a_{f,w}^2 = 1$ *and* $\sum_{w\geq 0} |a_{f,w}| \leq c$ *for some constant* c.

(E1) (Decomposition of $\mathbf{E}_t$). *We assume that*

$$\mathbf{E}_t = \mathbf{A}_{e,r}\mathbf{F}_{e,t}\mathbf{A}_{e,c}^\intercal + \mathbf{\Sigma}_\epsilon \circ \boldsymbol{\epsilon}_t, \tag{3.2}$$

where $\mathbf{F}_{e,t}$ *is a matrix of size* $k_{e,r} \times k_{e,c}$, *containing independent elements with mean* 0 *and variance* 1. *The matrix* $\boldsymbol{\epsilon}_t \in \mathbb{R}^{p\times q}$ *contains independent elements with mean 0 and variance 1, with* $\{\boldsymbol{\epsilon}_t\}$ *independent of* $\{\mathbf{F}_{e,t}\}$. *The matrix* $\mathbf{\Sigma}_\epsilon$ *contains the standard deviations of the corresponding elements in* $\boldsymbol{\epsilon}_t$, *and has elements uniformly bounded away from 0 and infinity. Moreover,* $\mathbf{A}_{e,r}$ *and* $\mathbf{A}_{e,c}$ *are (approximately) sparse matrices with sizes* $p \times k_{e,r}$ *and* $q \times k_{e,c}$ *respectively, such that* $\|\mathbf{A}_{e,r}\|_1, \|\mathbf{A}_{e,c}\|_1 = O(1)$, *with* $k_{e,r}, k_{e,c} = O(1)$.

(E2) (Time series in $\mathbf{E}_t$). *There is* $\mathbf{X}_{e,t}$ *the same dimension as* $\mathbf{F}_{e,t}$, *and* $\mathbf{X}_{\epsilon,t}$ *the same dimension as* $\boldsymbol{\epsilon}_t$, *such that* $\mathbf{F}_{e,t} = \sum_{w\geq 0} a_{e,w}\mathbf{X}_{e,t-w}$ *and* $\boldsymbol{\epsilon}_t = \sum_{w\geq 0} a_{\epsilon,w}\mathbf{X}_{\epsilon,t-w}$, *with* $\{\mathbf{X}_{e,t}\}$ *and* $\{\mathbf{X}_{\epsilon,t}\}$ *independent of each other, and each time series has independent elements with mean* 0 *and variance* 1 *with uniformly bounded fourth order moments.*

Both $\{\mathbf{X}_{e,t}\}$ *and* $\{\mathbf{X}_{\epsilon,t}\}$ *are independent of* $\{\mathbf{X}_{f,t}\}$ *from Assumption (F1). The coefficients* $a_{e,w}$ *and* $a_{\epsilon,w}$ *satisfy* $\sum_{w\geq 0} a_{e,w}^2 = \sum_{w\geq 0} a_{\epsilon,w}^2 = 1$ *and* $\sum_{w\geq 0} |a_{e,w}|, \sum_{w\geq 0} |a_{\epsilon,w}| \leq c$ *for some constant* c.

(E3) (Tail condition in $\mathbf{F}_t$ and $\mathbf{E}_t$). *Each element in the time series* $\{\mathbf{X}_{f,t}\}$ *from Assumption (F1),* $\{\mathbf{X}_{e,t}\}$ *and* $\{\mathbf{X}_{\epsilon,t}\}$ *from Assumption (E2) is sub-Gaussian.*

(R1) (Rate assumptions). *We assume that,*

$$T^{-1}p^{2(1-\delta_{r,k_r})}q^{1-2\delta_{c,1}} = o(1), \quad p^{1-2\delta_{r,k_r}}q^{2(1-\delta_{c,1})} = o(1),$$

$$T^{-1}q^{2(1-\delta_{c,k_c})}p^{1-2\delta_{r,1}} = o(1), \quad q^{1-2\delta_{c,k_c}}p^{2(1-\delta_{r,1})} = o(1).$$

(S1) (Lipschitz continuity). *The solution functions* $\widetilde{g}_\alpha(\cdot)$ *and* $\widetilde{g}_\beta(\cdot)$ *in Assumption (IC2) are Lipschitz continuous with respect to some vector norm* $\|\cdot\|_*$, *such that there exist constants* L_α *and* L_β, *that may depend on* T, p, q, *such that for all* $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^p$ *and* $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^q$:

$$|\widetilde{g}_\alpha(\mathbf{x}_1) - \widetilde{g}_\alpha(\mathbf{x}_2)| \leq L_\alpha\|\mathbf{x}_1 - \mathbf{x}_2\|_*, \quad |\widetilde{g}_\beta(\mathbf{y}_1) - \widetilde{g}_\beta(\mathbf{y}_2)| \leq L_\beta\|\mathbf{y}_1 - \mathbf{y}_2\|_*.$$

With Assumption (L1), we can define the normalized row and column loading matrices $\mathbf{Q}_r = \mathbf{A}_r\mathbf{Z}_r^{-1/2}$ and $\mathbf{Q}_c = \mathbf{A}_c\mathbf{Z}_c^{-1/2}$. Hence $\mathbf{Q}_r^\intercal\mathbf{Q}_r \to \mathbf{\Sigma}_{A,r}$ and $\mathbf{Q}_c^\intercal\mathbf{Q}_c \to \mathbf{\Sigma}_{A,c}$. Assumption (L1) allows for weak factors with heterogeneous factor strengths, which is also similarly seen in Remark 1 in Lam and Yao (2012), differing from traditional approximate factor models that consider either pervasive factors only (He et al., 2024) or weak factors with the same factor strength (Wang et al., 2019).

Assumptions (F1), (E1) and (E2) characterize the dynamics in the core factor and noise series, which are necessary in matrix factor models. Rather than directly presenting, e.g. the weak dependence in noise as in Assumptions (D) in Yu et al. (2022), our assumptions naturally specify the data generating process, allowing the noise and core factors to be

general linear processes. In particular, the core factor can be serially correlated under (F1) which also implies that for each $t \in [T]$, $\mathbb{E}[\mathbf{F}_t\mathbf{F}_t^\intercal] = k_c\mathbf{I}_{k_r}$ and $\mathbb{E}[\mathbf{F}_t^\intercal\mathbf{F}_t] = k_r\mathbf{I}_{k_c}$. This together with Assumption (L1) thus serve as an alternative set of identification conditions where the dependence among the latent dynamics driven by the core factors is featured by the loading matrices; see the discussion in Section 3 in Bai and Ng (2002) for instance. The decomposition by Assumption (E1) together with the general linear process in (E2) allow the noise to have both serial and cross-sectional dependence. Hence our Assumptions (E1) and (E2) are comparable to, for example, conditional independence as in Assumption 3.2 in He et al. (2024).

Assumption (E3) controls the tail of the random variables in model (1.1). It is arguably very mild since stronger assumptions are often needed in the regularized regression literature, such as Condition (a) in Zou (2006), Assumption (A1) in Huang et al. (2008), and Assumption (E) in Rinaldo (2009), to name but a few. In particular, as pointed out in Remark 3 in Fan et al. (2024), such sub-Gaussianity would not hold under highly correlated covariates in the regression problem, but the covariate matrix in our formulation is an identity matrix (see Section 4.3) and hence this issue is circumvented.

Assumption (R1) spells out the rates required on the dimensions and factor strengths, which directly hold if all factors are pervasive. Assumption (S1) imposes a smoothness condition on the solution functions, so as to study the convergence rate of the main effect estimators. Markedly, using different vector norms controls the deviation of the grand mean and main effect estimators from the true parameters to various extents, since the rates of convergence depend on some forms of the noise measured in $\|\cdot\|_*$. Note that the vector norms in Assumption (S1) can be specified differently over row and column main effects and timestamps, if the shift-varying functions in Assumption (IC1) are set differently. Throughout this paper, we assume all these are identical given the observed time series for simplicity. To further elaborate this assumption, we discuss some examples using the row main effects.

Suppose $\|\cdot\|_*$ is set as the Euclidean norm $\|\cdot\|$. Then for any shift-varying function in the set $\mathcal{L}$, its solution function takes the form (2.7) and hence

$$|\widetilde{g}_\alpha(\mathbf{x}_1) - \widetilde{g}_\alpha(\mathbf{x}_2)| = \left|(\mathbf{v}^\intercal\mathbf{1}_p)^{-1}\mathbf{v}^\intercal(\mathbf{x}_1 - \mathbf{x}_2)\right| \leq \frac{\|\mathbf{v}\|}{|\mathbf{v}^\intercal\mathbf{1}_p|} \cdot \|\mathbf{x}_1 - \mathbf{x}_2\|,$$

which indicates its Lipschitz constant as $L_\alpha = \|\mathbf{v}\|/|\mathbf{v}^\intercal\mathbf{1}_p|$. For instance, $L_\alpha = p^{-1/2}$ when $\mathbf{v} = \mathbf{1}_p$. The solution function of any shift-varying function in $\mathcal{Q}$ admits the form in (2.8),

which has the corresponding Lipschitz constant $L_\alpha = 1$, since

$$|\widetilde{g}_\alpha(\mathbf{x}_1) - \widetilde{g}_\alpha(\mathbf{x}_2)| = |q_u(\mathbf{x}_1) - q_u(\mathbf{x}_2)| \le \|\mathbf{x}_1 - \mathbf{x}_2\|_\infty \le \|\mathbf{x}_1 - \mathbf{x}_2\|.$$

On the other hand, suppose $\|\cdot\|_\infty$ is the L_∞ norm. Then for functions in $\mathcal{L}$, similar steps as above give $|\widetilde{g}_\alpha(\mathbf{x}_1) - \widetilde{g}_\alpha(\mathbf{x}_2)| \le (\|\mathbf{v}\|_1/|\mathbf{v}^\intercal \mathbf{1}_p|) \cdot \|\mathbf{x}_1 - \mathbf{x}_2\|_\infty$, thus implying $L_\alpha = \|\mathbf{v}\|_1/|\mathbf{v}^\intercal \mathbf{1}_p|$. For functions in $\mathcal{Q}$, their Lipschitz constants are again $L_\alpha = 1$ from the above display.

3.2 Theoretical results

We now study the statistical properties for the estimators in Section 2.2. First, we present the theoretical guarantee on estimating the main effects and grand mean, under identification using arbitrary shift-varying functions in Assumption (IC1).

Theorem 2 (Properties of $\widetilde{\boldsymbol{\alpha}}_t$, $\widetilde{\boldsymbol{\beta}}_t$, and $\widetilde{\mu}_t$, under arbitrary shift-varying functions)**.** *Let Assumptions (IC1), (IC2), (E1), (E2), and (S1) hold. Then the estimators for the main effects and the grand mean in Section 2.2 satisfy that for any $t \in [T]$, as $\min(T, p, q) \to \infty$,*

$$\begin{aligned}
\frac{1}{p}\|\widetilde{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}_t^*\|^2 &= O_P\Big(\frac{L_\alpha^2}{q^2}\,\|\mathbf{E}_t \mathbf{1}_q\|_*^2 + \frac{1}{q}\Big), \\
\frac{1}{q}\|\widetilde{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^*\|^2 &= O_P\Big(\frac{L_\beta^2}{p^2}\,\|\mathbf{E}_t^\intercal \mathbf{1}_p\|_*^2 + \frac{1}{p}\Big), \\
(\widetilde{\mu}_t - \mu_t)^2 &= O_P\Big(\frac{L_\alpha^2}{q^2}\,\|\mathbf{E}_t \mathbf{1}_q\|_*^2 + \frac{L_\beta^2}{p^2}\,\|\mathbf{E}_t^\intercal \mathbf{1}_p\|_*^2 + \frac{1}{pq}\Big).
\end{aligned}$$

Theorem 2 spells out the rates of convergence for the estimators defined in (2.3), (2.4), and (2.5). The theorem indicates that, for instance, the row (resp. column) main effect estimators at least carry the rate of $1/q$ (resp. $1/p$). Also notably, the convergence rate for the grand mean estimator has some identical terms as in the main effect estimators. While it is unsurprising from its construction in (2.5), the presented result only serves as a feasible bound for general identification in Assumption (IC1). The following theorem shows an exception of $\mathbf{v} = \mathbf{1}$, where the grand mean estimator has better rate than the main effect estimators; it also gives explicit results when the shift-varying functions in

Assumption (IC1) are specified in $\mathcal{L}$ or $\mathcal{Q}$.

Theorem 3 (Properties of $\widetilde{\boldsymbol{\alpha}}_t$, $\widetilde{\boldsymbol{\beta}}_t$, and $\widetilde{\mu}_t$, under shift-varying functions in $\mathcal{L}$ or $\mathcal{Q}$)**.** *In addition to all assumptions in Theorem 2, further let Assumption (E3) hold. Then for any* $t \in [T]$*, as* $\min(T, p, q) \to \infty$*, the following hold.*

(i) When the shift-varying functions are in $\mathcal{L}$*, i.e.,* $g(\mathbf{x}) = \mathbf{v}^\intercal \mathbf{x} - c$*, we have*

$$\begin{aligned}
\frac{1}{p}\|\widetilde{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}_t^*\|^2 &= O_P\left\{ \frac{(p\|\mathbf{v}\|^2) \bigwedge \{\log(p)\|\mathbf{v}\|_1^2\}}{q(\mathbf{v}^\intercal \mathbf{1}_p)^2} + \frac{1}{q} \right\}, \\
\frac{1}{q}\|\widetilde{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^*\|^2 &= O_P\left\{ \frac{(q\|\mathbf{v}\|^2) \bigwedge \{\log(q)\|\mathbf{v}\|_1^2\}}{p(\mathbf{v}^\intercal \mathbf{1}_q)^2} + \frac{1}{p} \right\}, \\
(\widetilde{\mu}_t - \mu_t)^2 &= O_P\left\{ \frac{(p\|\mathbf{v}\|^2) \bigwedge \{\log(p)\|\mathbf{v}\|_1^2\}}{q(\mathbf{v}^\intercal \mathbf{1}_p)^2} + \frac{(q\|\mathbf{v}\|^2) \bigwedge \{\log(q)\|\mathbf{v}\|_1^2\}}{p(\mathbf{v}^\intercal \mathbf{1}_q)^2} + \frac{1}{pq} \right\}.
\end{aligned}$$

In particular, when $\mathbf{v} = \mathbf{1}$ *(with dimensions* p *and* q *for the row and column main effects, respectively), we have* $p^{-1}\|\widetilde{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}_t^*\|^2 = O_P(1/q)$*,* $q^{-1}\|\widetilde{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^*\|^2 = O_P(1/p)$*, and markedly,* $(\widetilde{\mu}_t - \mu_t)^2 = O_P\{1/(pq)\}$*.*

(ii) When the shift-varying functions are in $\mathcal{Q}$*, i.e.,* $g(\mathbf{x}) = q_u(\mathbf{x}) - c$*, we have*

$$\begin{aligned}
\frac{1}{p}\|\widetilde{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}_t^*\|^2 &= O_P\left\{ \frac{\log(p)}{q} \right\}, \\
\frac{1}{q}\|\widetilde{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^*\|^2 &= O_P\left\{ \frac{\log(q)}{p} \right\}, \\
(\widetilde{\mu}_t - \mu_t)^2 &= O_P\left\{ \frac{\log(p)}{q} + \frac{\log(q)}{p} \right\}.
\end{aligned}$$

Notably, both parts of the theorem are agnostic to the value of c. From Theorem 3 (i), the choice of $\mathbf{v}$ guides the convergence rates. For instance, if $\mathbf{v}$ is sparse, then $p\|\mathbf{v}\|^2$ tends to dominate $\log(p)\|\mathbf{v}\|_1^2$ and the latter would then be the resulting term in the final presented rate. One such instance is when $\mathbf{v}$ is some standard basis vector, in which case we would have $p^{-1}\|\widetilde{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}_t^*\|^2 = O_P\{\log(p)/q\}$, etc. Setting $\mathbf{v} = \mathbf{1}$ illustrates a particular case where the grand mean estimator has better rate of convergence than the main effect estimators. This is essentially due to the cancellation of main effect estimators when contributed towards the estimation error of $\widetilde{\mu}_t$. We mark that part (i) also implies any non-zero scaling of $\mathbf{v}$ would not change the rates of convergence.

Theorem 3 (ii) is a consequence of Theorem 2 with L_∞ norm in Assumption (S1). The resulting rates carry some logarithmic terms of the dimensions, which is intuitive since the extrema of the noise series are inevitably involved due to the worst-case identification with u close to 0 or 1. However, when the shift-varying functions are central quantiles (e.g., median with $u = 0.5$), the Lipschitz constant in Assumption (S1) can be of smaller order if the main effects are sparse and the noise distribution has a smooth density, thus leading to an improved rate; we do not pursue this direction here.

Next, define the following square matrices

$$\mathbf{H}_r := \frac{1}{T}\widehat{\mathbf{D}}_r^{-1}\widehat{\mathbf{Q}}_r^\intercal\mathbf{Q}_r\sum_{t=1}^{T}(\mathbf{F}_{Z,t}\mathbf{Q}_c^\intercal\mathbf{Q}_c\mathbf{F}_{Z,t}^\intercal),$$
$$\mathbf{H}_c := \frac{1}{T}\widehat{\mathbf{D}}_c^{-1}\widehat{\mathbf{Q}}_c^\intercal\mathbf{Q}_c\sum_{t=1}^{T}(\mathbf{F}_{Z,t}^\intercal\mathbf{Q}_r^\intercal\mathbf{Q}_r\mathbf{F}_{Z,t}),$$

where $\widehat{\mathbf{D}}_r := \widehat{\mathbf{Q}}_r^\intercal(T^{-1}\sum_{t=1}^{T}\widetilde{\mathbf{L}}_t\widetilde{\mathbf{L}}_t^\intercal)\widehat{\mathbf{Q}}_r$ is the $k_r \times k_r$ diagonal matrix consisting of eigenvalues of $T^{-1}\sum_{t=1}^{T}\widetilde{\mathbf{L}}_t\widetilde{\mathbf{L}}_t^\intercal$, and $\widehat{\mathbf{D}}_c := \widehat{\mathbf{Q}}_c^\intercal(T^{-1}\sum_{t=1}^{T}\widetilde{\mathbf{L}}_t^\intercal\widetilde{\mathbf{L}}_t)\widehat{\mathbf{Q}}_c$ is the $k_c \times k_c$ diagonal matrix of eigenvalues of $T^{-1}\sum_{t=1}^{T}\widetilde{\mathbf{L}}_t^\intercal\widetilde{\mathbf{L}}_t$. Our next theorem then presents the asymptotic properties of $\mathbf{H}_r$, $\mathbf{H}_c$, and our estimators on the factor structure in Section 2.2.

Theorem 4 (Asymptotic rates for estimators on factor structure)**.** *Let Assumptions (IC1), (IC2), (L1), (F1), (E1), (E2), and (R1) hold. Then as* $\min(T, p, q) \to \infty$, $\mathbf{H}_r$ *and* $\mathbf{H}_c$ *are asymptotically invertible, and both the row and column factor loading estimators are*

consistent such that

$$\frac{1}{p}\big\|\widehat{\mathbf{Q}}_r - \mathbf{Q}_r\mathbf{H}_r^\intercal\big\|_F^2 = O_P\Big(T^{-1}p^{1-2\delta_{r,k_r}}q^{1-2\delta_{c,1}} + p^{-3\delta_{r,k_r}}q^{2(1-\delta_{c,1})}\Big),$$
$$\frac{1}{q}\big\|\widehat{\mathbf{Q}}_c - \mathbf{Q}_c\mathbf{H}_c^\intercal\big\|_F^2 = O_P\Big(T^{-1}q^{1-2\delta_{c,k_c}}p^{1-2\delta_{r,1}} + q^{-3\delta_{c,k_c}}p^{2(1-\delta_{r,1})}\Big).$$

Furthermore, the estimated core factor series and common components are consistent such that for any $t \in [T]$, $i \in [p]$, $j \in [q]$, we have

$$\big\|\widehat{\mathbf{F}}_{Z,t} - (\mathbf{H}_r^{-1})^\intercal\mathbf{F}_{Z,t}\mathbf{H}_c^{-1}\big\|_F^2 = O_P\big(p^{1-\delta_{r,k_r}}q^{1-\delta_{c,k_c}} + T^{-1}p^{1+2\delta_{r,1}-2\delta_{r,k_r}}q^{1-\delta_{c,1}} + p^{1+\delta_{r,1}-3\delta_{r,k_r}}q^{2-\delta_{c,1}}$$
$$+ T^{-1}q^{1+2\delta_{c,1}-2\delta_{c,k_c}}p^{1-\delta_{r,1}} + q^{1+\delta_{c,1}-3\delta_{c,k_c}}p^{2-\delta_{r,1}}\big),$$
$$(\widehat{C}_{t,ij} - C_{t,ij})^2 = O_P\big(p^{1-2\delta_{r,k_r}}q^{1-2\delta_{c,k_c}} + T^{-1}p^{1+2\delta_{r,1}-3\delta_{r,k_r}}q^{1-\delta_{c,1}-\delta_{c,k_c}} + p^{1+\delta_{r,1}-4\delta_{r,k_r}}q^{2-\delta_{c,1}-\delta_{c,k_c}}$$
$$+ T^{-1}q^{1+2\delta_{c,1}-3\delta_{c,k_c}}p^{1-\delta_{r,1}-\delta_{r,k_r}} + q^{1+\delta_{c,1}-4\delta_{c,k_c}}p^{2-\delta_{r,1}-\delta_{r,k_r}}\big).$$

Theorem 4 details the behaviors of the estimators of factor loading matrices, core factors, and thus common components, under different factor strengths. For a clear comparison with results in the literature, Corollary 1 (in the following) spells out the rates of convergence in Theorem 4 when all factors are strong. From the corollary, our factor loading estimators have comparable performance as the α-PCA estimators in Chen and Fan (2023). Also, our results align with, for instance, Theorem 4.1 of He et al. (2024) and Theorem 4 of Chen and Fan (2023).

Corollary 1 (Simplified Theorem 4 under pervasive factors)**.** *Let all assumptions in Theorem 4 hold, and further assume that $\delta_{r,i} = \delta_{c,j} = 1$ for all $i \in [k_r]$, $j \in [k_c]$. Define the renormalized row and column loading estimators, and core factor estimator as $\widehat{\mathbf{A}}_r := \sqrt{p}\,\widehat{\mathbf{Q}}_r$,*

$\widehat{\mathbf{A}}_c := \sqrt{q}\,\widehat{\mathbf{Q}}_c$, *and* $\widehat{\mathbf{F}}_t := \widehat{\mathbf{F}}_{Z,t}/\sqrt{pq}$, *respectively. Then for any* $t \in [T]$, $i \in [p]$, $j \in [q]$,

$$\frac{1}{p}\big\|\widehat{\mathbf{A}}_r - \mathbf{A}_r\mathbf{H}_r^\intercal\big\|_F^2 = O_P\Big(\frac{1}{Tq} + \frac{1}{p^2}\Big), \quad \frac{1}{q}\big\|\widehat{\mathbf{A}}_c - \mathbf{A}_c\mathbf{H}_c^\intercal\big\|_F^2 = O_P\Big(\frac{1}{Tp} + \frac{1}{q^2}\Big),$$
$$\big\|\widehat{\mathbf{F}}_t - (\mathbf{H}_r^{-1})^\intercal\mathbf{F}_t\mathbf{H}_c^{-1}\big\|_F^2, \quad (\widehat{C}_{t,ij} - C_{t,ij})^2 = O_P\Big(\frac{1}{Tq} + \frac{1}{Tp} + \frac{1}{p^2} + \frac{1}{q^2}\Big).$$

4 Sparsity Recovery by Regularized Estimation

As a useful line of analysis, we study the sparsity of main effects in this section. Throughout, the shift-varying function in Assumption (IC1) takes the form in (2.2) with $u, c = 0$, i.e., $g(\cdot) = \min\{\cdot\}$ so that the main effects are identified by

$$\min\{\boldsymbol{\alpha}_t^*\} = 0, \quad \min\{\boldsymbol{\beta}_t^*\} = 0. \tag{4.1}$$

For each cross-sectional unit, the main effects can be sparse in certain periods. Formally, consider the row effects $\{\boldsymbol{\alpha}_t^*\}_{t\in[T]}$, and for any $i \in [p]$, we define the sparse block as $\mathcal{S}_{\alpha,i} = \{t : \alpha_{t,i}^* = 0\}$, and the dense block $\mathcal{B}_{\alpha,i} = [T] \setminus \mathcal{S}_{\alpha,i}$. The sparse and dense blocks for the column main effects are defined similarly, denoted as $\mathcal{S}_{\beta,j}$ and $\mathcal{B}_{\beta,j}$ respectively. Note that both $\mathcal{S}_{\alpha,i}$ and $\mathcal{S}_{\beta,j}$ can be potentially empty. From a data generating point of view, the sparse blocks should be viewed as non-random sets of timestamps where the corresponding main effects vanish, and the dense blocks are the remaining periods.

For illustration, consider again the example of economic indices from Section 1, where each row corresponds to one country and hence the time series $\{\alpha_{t,i}^*\}$ represents the country-i's effect which could disappear when the country's economy goes down for instance. To strengthen the idea that the sparsity can be piecewise, i.e., zeros in the main effects can be consecutive in timestamps, we address this by incorporating a total variation loss, which we defer to Section 4.1. Note that there could be singular zeros which live in between non-zero main effects, so that our framework balances between interpretability and generality. To summarize, the benefit of such a sparsity framework is at least two-fold:

1. With piecewise zeros, the main effects can be understood more easily by practitioners; on the other hand, singular zeros potentially imply influential events and merit further investigation.

2. Allowing for general non-zeros retains adequate flexibility, compared to e.g. piecewise constants which is restricted and may not be realistic. Moreover, the dense blocks are able to feature the important patterns in the observed matrix time series, such as high volatility for financial return data.

4.1 Regularized estimation

We address in this subsection the estimation problem under sparsity, and our goal is to recover the sparse blocks for the row and column main effects. By (2.8) and analogous steps in Section 2.2, we first obtain the initial estimators for the main effects as

$$\widetilde{\boldsymbol{\alpha}}_t := q^{-1}\mathbf{X}_t\mathbf{1}_q - q^{-1}\mathbf{1}_p \min\{\mathbf{X}_t\mathbf{1}_q\}, \tag{4.2}$$

$$\widetilde{\boldsymbol{\beta}}_t := p^{-1}\mathbf{X}_t^\intercal\mathbf{1}_p - p^{-1}\mathbf{1}_q \min\{\mathbf{X}_t^\intercal\mathbf{1}_p\}. \tag{4.3}$$

Inspired by the Lasso (Tibshirani, 1996), we may employ an L_1 penalty to select the non-zero main effects. Nevertheless, such a regularization penalizes uniformly on each main effect value, potentially leading to over- or under-penalization which, in our scenario, fails block consistency (i.e., sparse blocks and dense blocks are estimated exactly as the true sets). To circumvent the restricted yet necessary irrepresentable condition in Lasso, Zou (2006) proposes to adaptively penalize the magnitude of estimators by reweighing the L_1 loss by the inverse of some well-behaved initial estimators. This adaptive Lasso method enlightens us to consider a loss function such that, for $i \in [p]$,

$$L^\circ(\alpha_{1,i}, \ldots, \alpha_{T,i}) := \frac{1}{2}\sum_{t=1}^{T}\left(\alpha_{t,i}^* - \alpha_{t,i}\right)^2 + \lambda_\alpha \sum_{t=1}^{T}\gamma_{\alpha,t,i}\left|\alpha_{t,i}\right|, \tag{4.4}$$

where $\gamma_{\alpha,t,i} = 1/\widetilde{\alpha}_{t,i}$ and λ_α is the tuning parameter. The estimators obtained by minimizing (4.4) are theoretically solid, but suffer from two flaws in practice. First, as a part of the latent MEFM representation, the row main effects cannot be directly observed, and hence all the $\alpha_{t,i}^*$'s in (4.4) are unavailable. Secondly, even if the true sparse block contains consecutive indices, the resulted main effect estimators from (4.4) do not favor piecewise sparsity, thus undermining the interpretation empirically.

To address the first concern above, we leverage the initial estimator $\widetilde{\alpha}_{t,i}$ which turns out to be an appropriate proxy to $\alpha_{t,i}^*$ under very general assumptions; see Assumption (R2) in Section 4.2. To promote smoothness in the estimator, we further include a total variation

loss in the objective function, which is akin to the fused Lasso method (Tibshirani et al., 2005; Rinaldo, 2009). Different from the traditional use of the total variation loss, we borrow the idea from adaptive Lasso again and reweigh the penalty by $1/\max\{\widetilde{\alpha}_{t,i}, \widetilde{\alpha}_{t-1,i}\}$, so that the smoothness is mainly encouraged on the sparse blocks. To this end, we introduce a new regularized estimator called the *Doubly Adaptive Fused Lasso (DAFL)* estimator, detailed as follows. For each $i \in [p]$, the DAFL estimator for the i-th row effect, $\{\widehat{\alpha}_{t,i}\}_{t\in[T]}$, is obtained by minimizing the penalized loss

$$L(\alpha_{1,i}, \ldots, \alpha_{T,i}) := \frac{1}{2}\sum_{t=1}^{T}\left(\widetilde{\alpha}_{t,i} - \alpha_{t,i}\right)^2 + \lambda_\alpha \sum_{t=2}^{T} u_{\alpha,t,i}\left|\alpha_{t,i} - \alpha_{t-1,i}\right| + \lambda_\alpha \sum_{t=1}^{T} \gamma_{\alpha,t,i}\left|\alpha_{t,i}\right|, \tag{4.5}$$

where $u_{\alpha,t,i} = 1/\max\{\widetilde{\alpha}_{t,i}, \widetilde{\alpha}_{t-1,i}\}$, and λ_α and $\gamma_{\alpha,t,i}$ are defined in (4.4). Note that $L(\cdot)$ depends on $\{\widetilde{\alpha}_{t,i}\}_{t\in[T]}$, which is not implied from our notation for the ease of presentation. The DAFL estimators for the column effects can be obtained similarly, denoted by $\{\widehat{\beta}_{t,j}\}_{t\in[T]}$ for $j \in [q]$. Then the sparse block estimators and their corresponding dense block estimators for the row and column main effects can be respectively defined as follows, for $i \in [p]$, $j \in [q]$,

$$\begin{aligned} &\widehat{\mathcal{S}}_{\alpha,i} := \{t : \widehat{\alpha}_{t,i} \le 0\}, \quad \widehat{\mathcal{S}}_{\beta,j} := \{t : \widehat{\beta}_{t,j} \le 0\}, \\ &\widehat{\mathcal{B}}_{\alpha,i} := [T] \setminus \widehat{\mathcal{S}}_{\alpha,i}, \quad \widehat{\mathcal{B}}_{\beta,j} := [T] \setminus \widehat{\mathcal{S}}_{\beta,j}. \end{aligned} \tag{4.6}$$

Note that we expect the DAFL estimators to be non-negative since the initial estimators are non-negative by how they are constructed. Hence, $\widehat{\mathcal{S}}_{\alpha,i}$ and $\widehat{\mathcal{S}}_{\beta,j}$ should technically correspond to the periods where $\widehat{\alpha}_{t,i}$ and $\widehat{\beta}_{t,j}$ are exactly zero, which might not hold empirically due to outliers or over-penalization. We thus define the sparse blocks as in (4.6) to address this; see more details in Remark 1.

Lastly, we construct the final estimators for the main effects by replacing all entries in $\widetilde{\boldsymbol{\alpha}}_t$ and $\widetilde{\boldsymbol{\beta}}_t$ by zero except for those according to the estimated dense blocks. That is, for each $t \in [T]$, $i \in [p]$, $j \in [q]$,

$$\widetilde{\alpha}^{\mathcal{B}}_{t,i} := \begin{cases} \widetilde{\alpha}_{t,i}, & t \in \widehat{\mathcal{B}}_{\alpha,i}; \\ 0, & \text{otherwise.} \end{cases}, \quad \widetilde{\beta}^{\mathcal{B}}_{t,j} := \begin{cases} \widetilde{\beta}_{t,j}, & t \in \widehat{\mathcal{B}}_{\beta,j}; \\ 0, & \text{otherwise.} \end{cases}$$

This unconventional step is to compensate for the use of e.g. $\widetilde{\alpha}_{t,i}$ in (4.5), which already induces an estimation error. Thus, even though the DAFL estimator $\{\widehat{\boldsymbol{\alpha}}_t\}_{t\in[T]}$ can con-

sistently recover the sparse and dense blocks, it inevitably inherits the error from $\widetilde{\alpha}_{t,i}$ and the rate of convergence deteriorates in general, cf. the discussion in Remark 1. Intuitively, $\{\widetilde{\boldsymbol{\alpha}}_t^{\mathcal{B}}\}_{t\in[T]}$ can be regarded as a thresholding estimator based on $\{\widetilde{\boldsymbol{\alpha}}_t\}_{t\in[T]}$, except that the thresholding is carried out by solving a constrained optimization problem with desirable features. Those final estimators are a compromise for the fact that main effects are latent, pinpointing again the difficulty of the sparsity problem in factor models.

4.2 Theoretical guarantee on sparsity recovery

Before we show the properties of the DAFL estimators, we require an additional rate assumptions as follows.

(R2) (Further rate assumptions). *We assume that,*

$$\lambda_\alpha^{-1}q^{-1}\log\Big(p\sum_{i=1}^{p}|\mathcal{S}_{\alpha,i}|\Big),\quad \lambda_\beta^{-1}p^{-1}\log\Big(q\sum_{j=1}^{q}|\mathcal{S}_{\beta,j}|\Big)=o(1),$$
$$\Big\{q^{-1}\log\Big(p\sum_{i=1}^{p}|\mathcal{B}_{\alpha,i}|\Big)+\lambda_\alpha\Big\}\Big(\min_{i\in[p]}\min_{t\in\mathcal{B}_{\alpha,i}}\{\alpha_{t,i}^{*2}\}\Big)^{-1}=o_P(1),$$
$$\Big\{p^{-1}\log\Big(q\sum_{j=1}^{q}|\mathcal{B}_{\beta,j}|\Big)+\lambda_\beta\Big\}\Big(\min_{j\in[q]}\min_{t\in\mathcal{B}_{\beta,j}}\{\beta_{t,j}^{*2}\}\Big)^{-1}=o_P(1).$$

Assumption (R2) is a set of rate assumptions required for block consistency to hold in general, restricting the sizes of the sparse and dense blocks, the tuning parameters, and the behavior of the non-zero main effects. Note that we presume

$$\min_{i\in[p]}\min_{t\in\mathcal{B}_{\alpha,i}}\{\alpha_{t,i}^{*}\},\ \min_{j\in[q]}\min_{t\in\mathcal{B}_{\beta,j}}\{\beta_{t,j}^{*}\}=O_P(1),$$

since otherwise block consistency can be trivially obtained by any constant thresholding, which is unrealistic. As long as the initial estimators of the main effects are consistent with rates according to Theorem 3 (ii), we can read the first line in Assumption (R2) as requiring $\log(T)/(q\lambda_\alpha+p\lambda_\beta)\to 0$. On the other hand, the second and third lines in (R2) restricts the tuning parameters to grow slower than the squared minimum non-zero main effects. Those constraints involving the sizes of the sparse and dense blocks are in parallel

to standard assumptions on the zero and non-zero coefficients in variable selection in linear regression, cf. Assumption (A4) in Huang et al. (2008). In what follows, we present the block consistency of the DAFL estimators, which further induces the results for the final estimators $\{\widetilde{\boldsymbol{\alpha}}_t^{\mathcal{B}}\}_{t\in[T]}$ and $\{\widetilde{\boldsymbol{\beta}}_t^{\mathcal{B}}\}_{t\in[T]}$.

Theorem 5 (Properties of DAFL estimators)**.** *Let all assumptions in Theorem 3 hold, with Assumption (IC1) (ii)–(iii) as in (4.1). Further given Assumption (R2), and consider* $\min\{p, q, T\} \to \infty$. *Then the DAFL estimators in Section 4.1 are consistent, and block consistency holds with*

$$\mathbb{P}(\widehat{\mathcal{S}}_{\alpha,1} = \mathcal{S}_{\alpha,1}, \ldots, \widehat{\mathcal{S}}_{\alpha,p} = \mathcal{S}_{\alpha,p}) \to 1, \quad \mathbb{P}(\widehat{\mathcal{S}}_{\beta,1} = \mathcal{S}_{\beta,1}, \ldots, \widehat{\mathcal{S}}_{\beta,q} = \mathcal{S}_{\beta,q}) \to 1.$$

Corollary 2. *Under the assumptions in Theorem 5, the final estimators for the main effects have the following properties.*

(i) (Block consistency). As $\min\{p, q, T\} \to \infty$, *it holds with probability tending 1 that*

$$\{t : \widetilde{\alpha}_{t,i}^{\mathcal{B}} = 0\} = \mathcal{S}_{\alpha,i} \quad \textit{for all } i \in [p],$$

$$\{t : \widetilde{\beta}_{t,j}^{\mathcal{B}} = 0\} = \mathcal{S}_{\beta,j} \quad \textit{for all } j \in [q].$$

(ii) (Uniform convergence). The final estimators in the dense blocks are consistent with

$$\max_{i\in[p]} \max_{t\in\mathcal{B}_{\alpha,i}} |\widetilde{\alpha}_{t,i}^{\mathcal{B}} - \alpha_{t,i}^*| = O_P\Big\{q^{-1/2}\log^{1/2}\Big(p\sum_{i=1}^{p}|\mathcal{B}_{\alpha,i}|\Big)\Big\},$$

$$\max_{j\in[q]} \max_{t\in\mathcal{B}_{\beta,j}} |\widetilde{\beta}_{t,j}^{\mathcal{B}} - \beta_{t,j}^*| = O_P\Big\{p^{-1/2}\log^{1/2}\Big(q\sum_{j=1}^{q}|\mathcal{B}_{\beta,j}|\Big)\Big\}.$$

Theorem 5 is key to the desired properties in the final estimators. Note that although the DAFL estimators are consistent, to derive a comparable rate of convergence as Theorem 3 (ii) requires stricter rate conditions than Assumption (R2). This is circumvented in the final estimators, as shown in Corollary 2 (ii). Intuitively, the final estimators for the main effects treat the DAFL estimation as a thresholding procedure. In terms of thresholding, by Assumption (R2), we require in the probability sense that the squared row main effects to asymptotically dominate the rate $q^{-1}\log\left(p\sum_{i=1}^{p}|\mathcal{S}_{\alpha,i}|\right)+q^{-1}\log\left(p\sum_{i=1}^{p}|\mathcal{B}_{\alpha,i}|\right)$ which is effectively of order $q^{-1}\log(pT)$; similar arguments hold for the column main effects. Finally, with the block consistency result from Theorem 5 and hence Corollary 2 (i), we present the behaviors of the main effect estimators in the dense blocks as in Corollary 2 (ii).

Remark 1. *Theorem 5 in Section 3.2 shows that the DAFL estimators are only consistent with arbitrary rates, compared to the rate from the initial estimators according to Theorem 3 (ii), unless more restrictive conditions hold. A more practical concern in directly using the DAFL estimators is related to Assumption (IC1). As explained below (4.6), the DAFL estimators might be negative in practice and would not fulfill Assumption (IC1) under $g(\cdot)=\min\{\cdot\}=0$, requiring that all main effects are at least zero. On the other hand, it is easy to see that the initial estimators from (4.2) and (4.3) satisfy the identification. This property is retained for the final estimators due to the way they are constructed.*

4.3 Practical implementation

We now discuss the practical optimization to compute our DAFL estimators. We only focus on estimating the row main effects by minimizing (4.5), as the arguments for the column main effects follow similarly. It turns out that our DAFL estimators can be obtained by equivalently solving a generalized lasso problem (Tibshirani and Taylor, 2011). More

specifically, for $i \in [p]$, if we define $\boldsymbol{\alpha}_{\cdot,i} := (\alpha_{1,i}, \ldots, \alpha_{T,i})^\intercal$, $\mathbf{D}_{\alpha,i} := \big(\mathbf{D}_{\alpha,i}^{(\mathrm{F})\intercal}, \mathbf{D}_{\alpha,i}^{(\mathrm{L})\intercal}\big)^\intercal$, where

$$\mathbf{D}_{\alpha,i}^{(\mathrm{L})} := \mathrm{diag}\big(\{1/\widetilde{\alpha}_{1,i}, \ldots, 1/\widetilde{\alpha}_{T,i}\}\big),$$

$$\mathbf{D}_{\alpha,i}^{(\mathrm{F})} := \begin{pmatrix} -(\widetilde{\alpha}_{1,i} \vee \widetilde{\alpha}_{2,i})^{-1} & (\widetilde{\alpha}_{1,i} \vee \widetilde{\alpha}_{2,i})^{-1} & 0 & \cdots & 0 \\ 0 & -(\widetilde{\alpha}_{2,i} \vee \widetilde{\alpha}_{3,i})^{-1} & (\widetilde{\alpha}_{2,i} \vee \widetilde{\alpha}_{3,i})^{-1} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -(\widetilde{\alpha}_{T-1,i} \vee \widetilde{\alpha}_{T,i})^{-1} & (\widetilde{\alpha}_{T-1,i} \vee \widetilde{\alpha}_{T,i})^{-1} \end{pmatrix},$$

then (4.5) can be rewritten as

$$L(\boldsymbol{\alpha}_{\cdot,i}) = \frac{1}{2}\big\|\widetilde{\boldsymbol{\alpha}}_{\cdot,i} - \boldsymbol{\alpha}_{\cdot,i}\big\|_2^2 + \lambda_\alpha \big\|\mathbf{D}_{\alpha,i}\boldsymbol{\alpha}_{\cdot,i}\big\|_1, \tag{4.7}$$

which has the form of Equation (2) in Tibshirani and Taylor (2011) and the solution path can be computed by algorithms therein in $O(T^3)$ times; see Algorithm 2 and the discussion in Section 8 in Tibshirani and Taylor (2011). It might be worth pointing out that although one can stack all rows and directly compute the solution path on $(\boldsymbol{\alpha}_{\cdot,1}^\intercal, \ldots, \boldsymbol{\alpha}_{\cdot,p}^\intercal)^\intercal$, this is not recommended since the computational complexity is of order T^3p^3, compared to T^3p by solving the above problem for each $i \in [p]$.

For such an ℓ_1 penalized regression problem as (4.7), it is often of interest to study its degrees of freedom which characterizes the effective number of parameters of a fitting procedure (Efron, 1986; Zou et al., 2007). In brief, for a data vector $\mathbf{y} \in \mathbb{R}^T$ whose elements are uncorrelated with homoscedastic mean μ and variance σ^2, the degrees of freedom of a function $g : \mathbb{R}^T \to \mathbb{R}^T$ is defined as

$$\mathrm{df}\,(g) := \frac{1}{\sigma^2}\sum_{t=1}^{T} \mathrm{Cov}\{g_t(\mathbf{y}), y_t\}.$$

As our DAFL estimator $\widehat{\boldsymbol{\alpha}}_{\cdot,i}$ is obtained by minimizing (4.7), we are interested in $\mathrm{df}\,(\widehat{\boldsymbol{\alpha}}_{\cdot,i})$ in terms of approximating the vector $\widetilde{\boldsymbol{\alpha}}_{\cdot,i}$. For a fixed tuning parameter λ_α, we may apply Theorem 1 in Tibshirani and Taylor (2011) and leverage the relation between the primal

and dual solutions to conclude

$$\mathrm{df}\,(\widehat{\boldsymbol{\alpha}}_{\cdot,i}) = \mathbb{E}\{\mathrm{nullity}(\mathbf{D}_{\mathcal{A}})\},$$

where $\mathbf{D}_{\mathcal{A}}$ denotes the matrix $\mathbf{D}_{\alpha,i}$ with rows restricted on the index set $\mathcal{A} = \{t : (\mathbf{D}_{\alpha,i}\widehat{\boldsymbol{\alpha}}_{\cdot,i})_t = 0\}$. Without the adaptive terms $u_{\alpha,t,i}$ and $\gamma_{\alpha,t,i}$ in (4.7), our problem boils down to the sparse fused lasso problem in Equation (47) in Tibshirani and Taylor (2011) and the degrees of freedom can be reduced to the expected number of non-zero fused groups in $\widehat{\boldsymbol{\alpha}}_{\cdot,i}$. Although we do not have such interpretation in our complicated scenario, we may readily use the realized $\mathrm{nullity}(\mathbf{D}_{\mathcal{A}})$ and hence compute an estimate of the degrees of freedom. This allows us to modify the Mallows's C_p statistic (Mallows, 1973) as

$$\widehat{C}_p(\lambda_\alpha) := \left\|\widetilde{\boldsymbol{\alpha}}_{\cdot,i} - \widehat{\boldsymbol{\alpha}}_{\cdot,i}\right\|_2^2 - T\widehat{\sigma}^2 + 2\widehat{\sigma}^2\mathrm{nullity}(\mathbf{D}_{\mathcal{A}}),$$

where $\widehat{\sigma}^2$ is the sample variance of $\widetilde{\boldsymbol{\alpha}}_{\cdot,i}$. Therefore, in terms of model selection, we may choose the tuning parameter λ_α to minimize $\widehat{C}_p(\lambda_\alpha)$, among a grid of candidates. Note that we use the same λ_α over $i \in [p]$ in (4.5), so that the theoretical guarantee holds uniformly. In practice, we can either minimize the aggregated $\widehat{C}_p(\lambda_\alpha)$ over all $i \in [p]$ or, more generally, in our numerical experiments, we select different λ_α's for each i.

5 Numerical Analysis

5.1 Simulation

In this subsection, we showcase the numerical performance of the proposed estimators using Monte Carlo experiments. We first evaluate the accuracy of the main effect and grand mean estimators under various linear and quantile identifications. We then examine the accuracy of the estimated common components and demonstrate their invariance to the identification imposed on the main effects. Finally, under the minimum-type identification, we investigate the performance of the DAFL and final estimators in terms of sparsity recovery and estimation accuracy.

Data generating process. We first generate grand mean series $\{\mu_t^\circ\}$, main effect series $\{\boldsymbol{\alpha}_t^\circ\}$ and $\{\boldsymbol{\beta}_t^\circ\}$. The grand mean and each entry of the two main effect processes independently follow AR(2) processes with coefficients (0.5,-0.2) and i.i.d. $\mathcal{N}(0,1)$ innovations.

Based on these, we construct different versions of grand mean and main effects according to different shift-varying functions in Assumption (IC1), as detailed in each experiment.

The noise and factor series $\mathbf{E}_t$ and $\mathbf{F}_t$ are formed as in Assumptions (E1), (E2), (E3) and (F1). Specifically, elements in $\mathbf{F}_t$ are jointly independent and each follows a standardized AR(2) process with coefficients (0.5, -0.3). Elements in $\mathbf{F}_{e,t}$ and $\boldsymbol{\epsilon}_t$ are similarly constructed, except that the AR coefficients are (-0.4, 0.4) and (0.6, 0.2), respectively. Furthermore, elements in the standard deviation matrix $\boldsymbol{\Sigma}_\epsilon$ are generated by i.i.d. $|\mathcal{N}(0,1)|$. The innovation processes in generating $\mathbf{F}_t$, $\mathbf{F}_{e,t}$ and $\boldsymbol{\epsilon}_t$ are i.i.d. $\mathcal{N}(0,1)$. To incorporate factor strengths, the row factor loading matrix is generated as $\mathbf{A}_r = \mathbf{M}_p\mathbf{U}_r\mathbf{B}_r$, where $\mathbf{U}_r \in \mathbb{R}^{p\times k_r}$ consists of i.i.d. $\mathcal{N}(0,1)$ elements, $\mathbf{M}_p$ is defined in Section 1 so that item (i) of Assumption (IC1) is satisfied, and $\mathbf{B}_r = \text{diag}(p^{-\zeta_{r,1}}, \ldots, p^{-\zeta_{r,k_r}})$. Note that $\zeta_{r,j} \in [0, 0.5]$, with pervasive factors represented by $\zeta_{r,j} = 0$ and weak factors otherwise. The column loading $\mathbf{A}_c$ is similarly generated. Each entry of $\mathbf{A}_{e,r}$ and $\mathbf{A}_{e,c}$ is i.i.d. standard normal and has probability of 0.95 being exact 0. Throughout the simulation, we fix $k_{e,r} = k_{e,c} = 2$ and $k_r = k_c = 2$. All numerical results are based on 500 Monte Carlo replications, unless specified otherwise.

To evaluate estimation accuracy, for any parameter series $\boldsymbol{\theta} = \{\boldsymbol{\theta}_t\}_{t\in[T]}$, where $\boldsymbol{\theta}_t$ can be a scalar, vector, or matrix, and its estimator $\widehat{\boldsymbol{\theta}} = \{\widehat{\boldsymbol{\theta}}_t\}_{t\in[T]}$, we define the mean squared errors as

$$\text{MSE}(\widehat{\boldsymbol{\theta}}) := \frac{\sum_{t=1}^{T} \|\boldsymbol{\theta}_t - \widehat{\boldsymbol{\theta}}_t\|_F^2}{T d_{\boldsymbol{\theta}}},$$

where $d_{\boldsymbol{\theta}}$ denotes the number of elements in $\boldsymbol{\theta}_t$. Moreover, for any given $\mathbf{Q}$ and $\widehat{\mathbf{Q}}$, we use the column space distance to measure their discrepancy:

$$\mathcal{D}(\mathbf{Q}, \widehat{\mathbf{Q}}) := \left\| \mathbf{Q}(\mathbf{Q}^\intercal\mathbf{Q})^{-1}\mathbf{Q}^\intercal - \widehat{\mathbf{Q}}(\widehat{\mathbf{Q}}^\intercal\widehat{\mathbf{Q}})^{-1}\widehat{\mathbf{Q}}^\intercal \right\|,$$

which is widely used in the literature such as Chen et al. (2022), among others. Unless otherwise stated, all reported performance measures are averaged over replications.

Accuracy of grand mean and main effects. Given shift-varying functions g_α and g_β and their solution functions $\widetilde{g}_\alpha$ and $\widetilde{g}_\beta$ (cf. Assumptions (IC1) and (IC2)), the identified parameters are constructed as

$$\boldsymbol{\alpha}_t^* := \boldsymbol{\alpha}_t^\circ - \widetilde{g}_\alpha(\boldsymbol{\alpha}_t^\circ)\mathbf{1}_p, \qquad \boldsymbol{\beta}_t^* := \boldsymbol{\beta}_t^\circ - \widetilde{g}_\beta(\boldsymbol{\beta}_t^\circ)\mathbf{1}_q, \qquad \mu_t^* := \mu_t^\circ + \widetilde{g}_\alpha(\boldsymbol{\alpha}_t^\circ) + \widetilde{g}_\beta(\boldsymbol{\beta}_t^\circ).$$

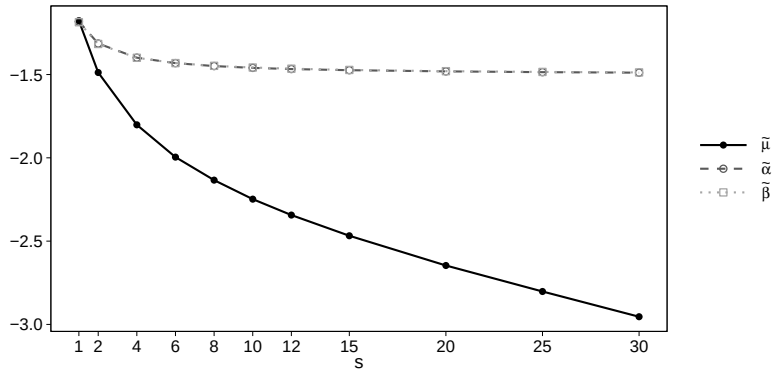

Figure 2: MSEs under linear identification as the number s varies (on the $\log_{10}$ scale).

It follows that $g_\alpha(\boldsymbol{\alpha}_t^*) = g_\beta(\boldsymbol{\beta}_t^*) = 0$, while

$$\mu_t^* \mathbf{1}_p \mathbf{1}_q^\intercal + \boldsymbol{\alpha}_t^* \mathbf{1}_q^\intercal + \mathbf{1}_p \boldsymbol{\beta}_t^{*\intercal} = \mu_t^\circ \mathbf{1}_p \mathbf{1}_q^\intercal + \boldsymbol{\alpha}_t^\circ \mathbf{1}_q^\intercal + \mathbf{1}_p \boldsymbol{\beta}_t^{\circ\intercal}.$$

Hence, within each realization, the observed matrix time series $\{\mathbf{X}_t\}$ is exactly the same under any identification, while the identification only changes the values of the grand mean and main effects. This allows all identifications to be compared using the same set of generated data. Throughout this experiment, all factors are pervasive such that $\zeta_{r,j} = \zeta_{c,j} = 0$. We first experiment the linear identifications described in $\mathcal{L}$. Given any positive integer d and $s \in [d]$, let $\mathbf{v}_{d,s} := (\mathbf{1}_s^\intercal, \mathbf{0}_{d-s}^\intercal)^\intercal \in \mathbb{R}^d$. We set

$$g_\alpha(\mathbf{a}) = \mathbf{v}_{p,s_p}^\intercal \mathbf{a} \quad \text{and} \quad g_\beta(\mathbf{b}) = \mathbf{v}_{q,s_q}^\intercal \mathbf{b}.$$

A smaller value of s_p or s_q corresponds to a more concentrated identification, with $s_p = s_q = 1$ giving a reference-unit identification.

To examine the effect of weight concentration, we fix $T = 100$ and $p = q = 30$ and impose $s_p = s_q = s$. Figure 2 shows how the accuracy of the grand mean and main effect estimators varies with s. The MSEs decrease as s increases, showing that less concentrated identifications lead to more accurate estimation. The improvement is much stronger for the grand mean than for the main effects, consistent with Theorem 3 (i).

We then examine the effects of varying p, q, and T under the reference-unit and equal-weight identifications, corresponding to $(s_p, s_q) = (1, 1)$ and (p, q), respectively. The results reported in Table 1 are largely unchanged as T increases. The equal-weight identification is substantially more accurate than the reference-unit identification, which is consistent with

Table 1: Accuracy under linear identification. Entries are MSE means multiplied by 10^2, with standard errors in parentheses.

(s_p, s_q)	p	q	MSE($\widetilde{\mu}$)		MSE($\widetilde{\boldsymbol{\alpha}}$)		MSE($\widetilde{\boldsymbol{\beta}}$)	
			$T = 100$	$T = 200$	$T = 100$	$T = 200$	$T = 100$	$T = 200$
$(1, 1)$	30	30	6.652(0.110)	6.464(0.108)	6.577(0.065)	6.430(0.060)	6.516(0.063)	6.413(0.057)
	80	80	2.526(0.037)	2.549(0.029)	2.499(0.022)	2.554(0.020)	2.505(0.020)	2.491(0.016)
	30	80	4.591(0.078)	4.656(0.071)	2.474(0.022)	2.426(0.015)	6.628(0.059)	6.712(0.062)
(p, q)	30	30	0.111(0.002)	0.112(0.001)	3.245(0.011)	3.237(0.010)	3.256(0.012)	3.249(0.010)
	80	80	0.016(0.000)	0.016(0.000)	1.250(0.002)	1.247(0.002)	1.251(0.002)	1.249(0.002)
	30	80	0.042(0.001)	0.042(0.000)	1.223(0.004)	1.223(0.003)	3.332(0.007)	3.339(0.007)

Figure 2. The grand mean becomes more accurate as either p or q increases. For the main effects, increasing q improves the row main effect estimator but not the column main effect estimator, while increasing p shows the reverse pattern. These patterns are also consistent with Theorem 3 (i).

We next experiment the quantile identifications described in $\mathcal{Q}$. Specifically, we set

$$g_\alpha(\mathbf{a}) = q_u(\mathbf{a}) \quad \text{and} \quad g_\beta(\mathbf{b}) = q_u(\mathbf{b}),$$

and vary the quantile level over $u \in \{0, 0.25, 0.5\}$. The results over different dimension combinations are reported in Table 2. The estimation accuracy improves as the identifying quantile moves toward the center. This is because tail quantiles are more sensitive to extreme cross-sectional noise, with $u = 0$ being directly determined by the minimum. The results change little as T increases, while the effects of the cross-sectional dimensions follow the same patterns as under linear identification.

Accuracy of common components. We next examine the estimation accuracy of $\widehat{\mathbf{Q}}_r$, $\widehat{\mathbf{Q}}_c$, and $\widehat{\mathbf{C}}_t$. Since these estimators are invariant to the identification imposed on the main effects, we fix the median identification throughout this experiment, namely, $g_\alpha(\mathbf{a}) = q_{0.5}(\mathbf{a})$ and $g_\beta(\mathbf{b}) = q_{0.5}(\mathbf{b})$. The grand mean and main effects are constructed accordingly as in the preceding experiment. We vary p, q, T, and the factor strength to examine their effects on the estimation accuracy, as reported in Table 3. In general, increasing T or the cross-sectional dimensions improves estimation accuracy. Weak factors substantially reduce the accuracy of the loading estimators, while their effect on the common component is relatively mild. corroborating with Theorem 4.

Table 2: Accuracy under quantile identification. Entries are MSE means multiplied by 10^2, with standard errors in parentheses.

p	q	u	MSE($\widetilde{\mu}$)		MSE($\widetilde{\boldsymbol{\alpha}}$)		MSE($\widetilde{\boldsymbol{\beta}}$)	
			$T = 100$	$T = 200$	$T = 100$	$T = 200$	$T = 100$	$T = 200$
30	30	0	6.160(0.047)	6.155(0.034)	6.180(0.030)	6.162(0.023)	6.192(0.030)	6.162(0.023)
		0.25	2.062(0.014)	2.069(0.011)	4.206(0.015)	4.202(0.013)	4.220(0.015)	4.223(0.014)
		0.5	1.609(0.011)	1.616(0.008)	3.996(0.014)	3.990(0.012)	4.008(0.014)	4.002(0.012)
80	80	0	2.362(0.015)	2.358(0.011)	2.400(0.009)	2.401(0.007)	2.402(0.009)	2.400(0.007)
		0.25	0.579(0.004)	0.574(0.003)	1.529(0.003)	1.523(0.003)	1.534(0.003)	1.530(0.003)
		0.5	0.458(0.003)	0.459(0.002)	1.471(0.003)	1.469(0.002)	1.474(0.003)	1.471(0.002)
30	80	0	4.245(0.030)	4.236(0.021)	2.362(0.011)	2.368(0.007)	6.309(0.024)	6.323(0.020)
		0.25	1.016(0.007)	1.023(0.005)	1.692(0.006)	1.691(0.004)	3.838(0.009)	3.845(0.008)
		0.5	0.805(0.006)	0.803(0.004)	1.595(0.005)	1.593(0.004)	3.719(0.008)	3.728(0.007)

Table 3: Loading-space and common component estimation accuracy, where $\zeta = \zeta_r = \zeta_c$. Entries are Monte Carlo means multiplied by 10^2, with standard errors in parentheses.

p	q	ζ	$\mathcal{D}(\mathbf{Q}_r, \widehat{\mathbf{Q}}_r)$		$\mathcal{D}(\mathbf{Q}_c, \widehat{\mathbf{Q}}_c)$		MSE($\widehat{\mathbf{C}}$)	
			$T = 100$	$T = 200$	$T = 100$	$T = 200$	$T = 100$	$T = 200$
30	30	0	2.016(0.022)	1.445(0.017)	1.976(0.022)	1.455(0.020)	0.637(0.004)	0.550(0.003)
		0.2	16.844(0.450)	13.491(0.315)	16.528(0.383)	13.529(0.336)	1.284(0.033)	1.014(0.027)
80	80	0	1.046(0.007)	0.728(0.004)	1.052(0.007)	0.732(0.004)	0.127(0.001)	0.096(0.000)
		0.2	10.064(0.158)	8.601(0.304)	10.020(0.135)	8.567(0.314)	0.235(0.004)	0.238(0.033)
30	80	0	1.179(0.012)	0.873(0.009)	1.751(0.015)	1.227(0.009)	0.288(0.001)	0.231(0.001)
		0.2	15.000(0.470)	11.906(0.270)	13.176(0.360)	9.313(0.143)	0.714(0.067)	0.434(0.010)

Sparsity recovery under minimum identification. We investigate the sparsity recovery performance of the DAFL estimators and whether exploiting the recovered sparsity improves the final estimation of the main effects. Throughout, we impose the minimum identification, namely, $g_\alpha(\mathbf{a}) = \min\{\mathbf{a}\}$ and $g_\beta(\mathbf{b}) = \min\{\mathbf{b}\}$. All factors are pervasive, the grand mean, common component, and noise processes are generated as in the beginning of this section, except that the row and column main effects are generated from sparse processes described below.

We describe the generating mechanism for the row main effects, with the column main effects constructed analogously. For each $i \in [p]$, denote the stay-in probability for sparse blocks and dense blocks given some sparsity level by $\pi^{\mathcal{S}}_{\alpha,i}$ and $\pi^{\mathcal{B}}_{\alpha,i}$ respectively, such that

$$\mathbb{P}(\alpha^\circ_{t+1,i} = 0 \mid \alpha^\circ_{t,i} = 0) = \pi^{\mathcal{S}}_{\alpha,i}, \quad \mathbb{P}(\alpha^\circ_{t+1,i} > 0 \mid \alpha^\circ_{t,i} > 0) = \pi^{\mathcal{B}}_{\alpha,i}, \tag{5.1}$$

Essentially, larger stay-in probabilities imply longer piecewise blocks, while a larger $\pi^{\mathcal{S}}_{\alpha,i}$ combined with a smaller $\pi^{\mathcal{B}}_{\alpha,i}$ implies a larger $|\mathcal{S}_{\alpha,i}|$, and vice versa. Define i.i.d. random variables $Z_{\alpha,t,i} \sim |\mathcal{N}(0,1)|$ for $t \in [T]$ and $i \in [p]$. Then given constant m_α, we generate $\alpha^\circ_{t,i}$ for $t \in [T]$ by the following steps:

Step 1. Define $p_{\alpha,i,*} := (1-\pi^{\mathcal{B}}_{\alpha,i})/(2-\pi^{\mathcal{S}}_{\alpha,i}-\pi^{\mathcal{B}}_{\alpha,i})$. Set $\alpha^\circ_{1,i} = 0$ with probability $p_{\alpha,i,*}$, and set $\alpha^\circ_{1,i} = m_\alpha + Z_{\alpha,1,i}$ otherwise.

Step 2. For $t = 2, \ldots, T$, conditional on the state at $t-1$, generate the state at t according to (5.1). Set $\alpha^\circ_{t,i} = 0$ if the process is in the sparse block, and set $\alpha^\circ_{t,i} = m_\alpha + Z_{\alpha,t,i}$ if it is in the dense block.

The following proposition presents the properties of the resulting process $\{\alpha^\circ_{t,i}\}_{t\in[T]}$.

Proposition 1. *For $\{\alpha^\circ_{t,i}\}_{t\in[T]}$ generated by Steps 1–2 above, it holds that: (i) $\{\mathbb{1}\{\alpha^\circ_{t,i} = 0\}\}_{t\in[T]}$ is a stationary Markov process, with stationary probability $\mathbb{P}(\alpha^\circ_{t,i} = 0) = p_{\alpha,i,*} := (1-\pi^{\mathcal{B}}_{\alpha,i})/(2-\pi^{\mathcal{S}}_{\alpha,i}-\pi^{\mathcal{B}}_{\alpha,i})$; (ii) with $\mathcal{S}^\circ_{\alpha,i} := \{t : \alpha^\circ_{t,i} = 0\}$, we have $\mathbb{E}(|\mathcal{S}^\circ_{\alpha,i}|) = Tp_{\alpha,i,*}$; and (iii) the expected length of a consecutive sparse interval is $L^{\mathcal{S}}_{\alpha,i} = (1-\pi^{\mathcal{S}}_{\alpha,i})^{-1}$.*

Proposition 1 shows that the two stay-in probabilities provide a convenient way to control two distinct features of sparsity. In particular, $p_{\alpha,i,*}$ determines the overall proportion of sparse observations, whereas $L^{\mathcal{S}}_{\alpha,i}$ determines their temporal persistence. In the experiments below, these parameters are taken to be homogeneous across units, and we write them as p_* and $L_{\mathcal{S}}$. For any admissible pair $(p_*, L_{\mathcal{S}})$, the corresponding stay-in probabilities are $\pi^{\mathcal{S}} = 1 - 1/L_{\mathcal{S}}$ and $\pi^{\mathcal{B}} = 1 - \frac{p_*}{(1-p_*)L_{\mathcal{S}}}$. After generating $\boldsymbol{\alpha}^\circ_t$ and $\boldsymbol{\beta}^\circ_t$, we impose the minimum identification through the same shift transformation as in the preceding experiments. Specifically, we set

$$\boldsymbol{\alpha}^*_t = \boldsymbol{\alpha}^\circ_t - \big(\min\{\boldsymbol{\alpha}^\circ_t\}\big)\mathbf{1}_p, \quad \boldsymbol{\beta}^*_t = \boldsymbol{\beta}^\circ_t - \big(\min\{\boldsymbol{\beta}^\circ_t\}\big)\mathbf{1}_q, \quad \mu^*_t = \mu^\circ_t + \min\{\boldsymbol{\alpha}^\circ_t\} + \min\{\boldsymbol{\beta}^\circ_t\}.$$

For every replication, we obtain the DAFL estimators $\widehat{\boldsymbol{\alpha}}_t$ and $\widehat{\boldsymbol{\beta}}_t$ using the tuning parameters selected by the modified C_p criterion in Section 4.3. We evaluate sparsity recovery using sensitivity and specificity. For the row main effects, define

$$\text{sensitivity}_{\boldsymbol{\alpha}} := \frac{\sum_{i=1}^p |\widehat{\mathcal{B}}_{\alpha,i} \cap \mathcal{B}_{\alpha,i}|}{\sum_{i=1}^p |\mathcal{B}_{\alpha,i}|}, \quad \text{specificity}_{\boldsymbol{\alpha}} := \frac{\sum_{i=1}^p |\widehat{\mathcal{S}}_{\alpha,i} \cap \mathcal{S}_{\alpha,i}|}{\sum_{i=1}^p |\mathcal{S}_{\alpha,i}|}.$$

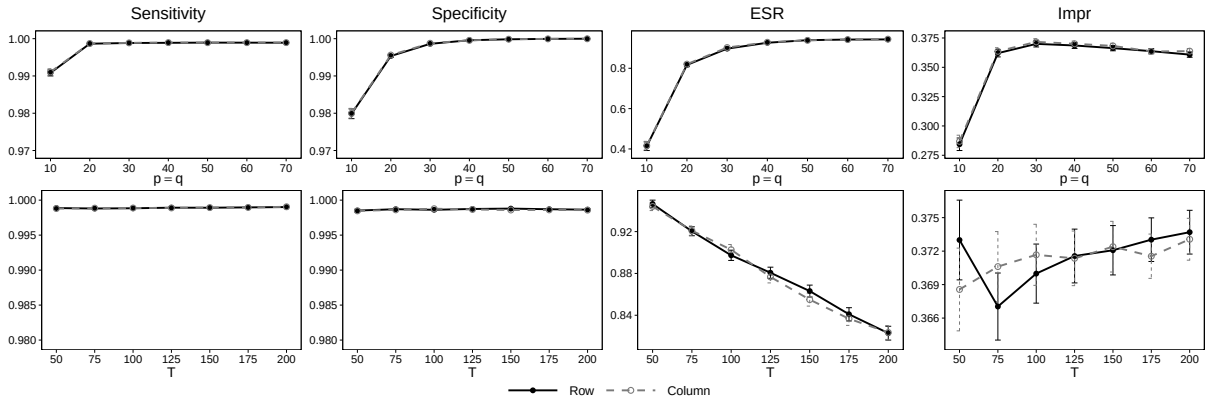

Figure 3: Recovery performance. The columns report sensitivity, specificity, ESR, and Impr, respectively. The first row varies $p = q$ with $T = 100$, while the second row varies T with $p = q = 30$. Curves show the row and column Monte Carlo means with 95% Monte Carlo error bars.

Also, define a unit-level exact sparsity recovery rate as $\mathrm{ESR}_{\boldsymbol{\alpha}} := p^{-1} \sum_{i=1}^{p} \mathbb{1}\{\widehat{\mathcal{S}}_{\alpha,i} = \mathcal{S}_{\alpha,i}\}$. Lastly, to assess whether exploiting the estimated sparsity improves main-effect estimation, we report the relative MSE improvement $\mathrm{Impr}_{\alpha} := 1 - \mathrm{MSE}(\widetilde{\boldsymbol{\alpha}}^{\mathcal{B}})/\mathrm{MSE}(\widetilde{\boldsymbol{\alpha}})$. Let the corresponding quantities $\mathrm{sensitivity}_{\boldsymbol{\beta}}$, $\mathrm{specificity}_{\boldsymbol{\beta}}$, $\mathrm{ESR}_{\boldsymbol{\beta}}$, and Impr_{β} be analogously defined.

We investigate sparsity recovery with respect to dimensions and series length, signal strength, and sparsity patterns. First, we vary the matrix dimensions and series length, while fixing a representative sparsity pattern with $p_* = 0.4$ and $L_{\mathcal{S}} = 4$, and setting $m_{\alpha} = m_{\beta} = 2$. Figure 3 presents the recovery performance. Except for $p = q = 10$, the sensitivity and specificity both exceed 0.99 and quickly approach one as $p = q$ increases. ESR shows a stronger dimensional effect and also approaches one as the cross-sectional dimensions increase. As T increases, the sensitivity and specificity remain nearly unchanged near one, while the ESR gradually decreases but remains above 0.8. This is natural because ESR requires the sparse block of a unit to be recovered exactly over its entire time trajectory. A longer series therefore makes exact recovery more demanding, even when the overall recovery accuracy remains nearly unchanged. Furthermore, Impr_{α} and Impr_{β} remain well above zero across all settings. Hence, exploiting the recovered sparsity indeed improves the estimation accuracy of the main effects relative to the initial estimators, showing that sparsity recovery also leads to estimation gains.

Second, we examine the effect of signal strength under $p = q = 30$ and $T = 100$, while fixing $p_* = 0.4$ and $L_{\mathcal{S}} = 4$. The first row of Figure 4 shows that both sensitivity and specificity rapidly approach one as $m = m_{\alpha} = m_{\beta}$ increases. ESR increases from near zero under the weakest signal to nearly one. Under weak signals, specificity is lower than sensitivity, indicating that true sparse observations are more often classified as dense. When m is

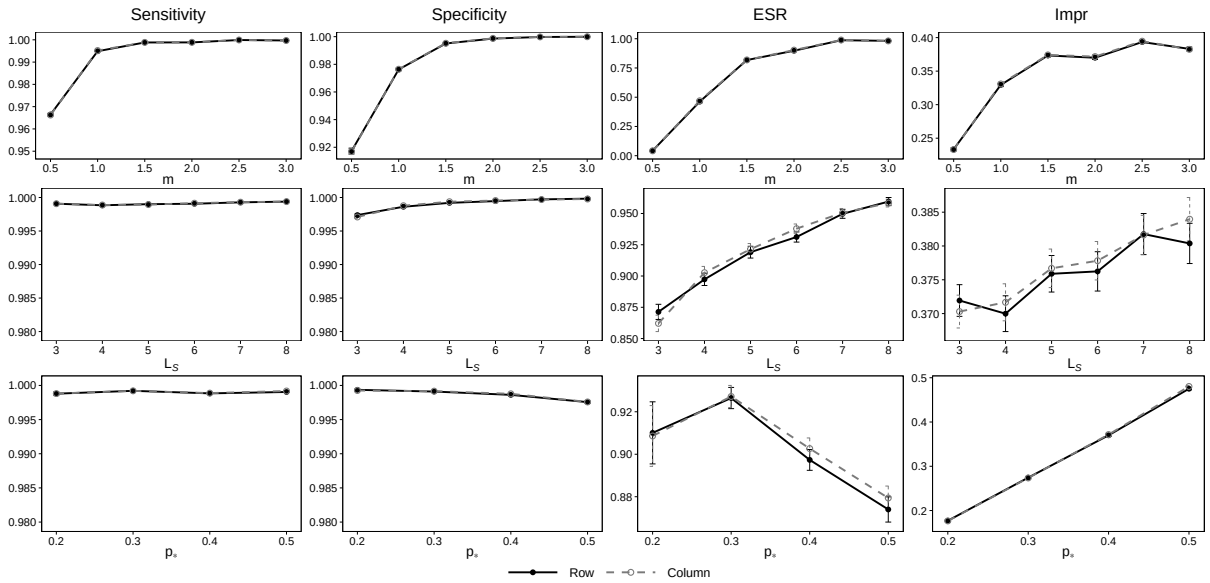

Figure 4: Recovery performance. The rows vary signal strength, sparse-block persistence, and sparsity level, respectively. The columns report sensitivity, specificity, ESR, and Impr, respectively. Curves show the row and column Monte Carlo means with 95% Monte Carlo error bars.

small, sparse and dense effects are less well separated relative to estimation noise, so some sparse observations receive insufficient adaptive penalization and remain positive. As m increases, sparse and dense effects become better separated, which is also consistent with Assumption (R2).

We next examine the effects of sparse-block persistence and sparsity level under the same $p = q = 30$, $T = 100$, and $m_\alpha = m_\beta = 2$ setting, as shown in the second and third rows of Figure 4. When $p_* = 0.4$ is fixed, sensitivity and specificity remain close to one as $L_{\mathcal{S}}$ varies, while ESR steadily approaches one as $L_{\mathcal{S}}$ increases. This pattern is consistent with the role of the fused term in DAFL, which exploits longer piecewise sparse blocks for sparsity recovery. When $L_{\mathcal{S}} = 4$ is fixed, increasing p_* leaves $\pi^{\mathcal{S}} = 0.75$ unchanged but decreases $\pi^{\mathcal{B}}$. Hence, the sparse blocks have the same expected length, while the dense blocks become shorter and sparse-dense transitions occur more frequently.

Across settings, Impr_α and Impr_β remain positive. They generally increase as the signal strength or sparse-block persistence increases, following the improvement in sparsity recovery. The increase is particularly clear as p_* increases, since a larger proportion of true zeros allows the final estimators to remove estimation error from more sparse entries. Overall, under the minimum identification, DAFL accurately recovers the sparse blocks across a wide range of settings, with stronger signals and longer sparse blocks further improving

recovery. Exploiting the recovered sparsity also consistently improves the final main effect estimators over their initial counterparts.

5.2 A real application on labor market

Consider the QCEW data introduced in Section 1, where the details on the dataset are included in the supplementary material. To showcase our sparsity recovery procedure, we now in turn consider the identification $\min\{\boldsymbol{\alpha}_t^*\} = 0$. The minimum is natural on the state categorization because its zero set identifies the lower boundary relevant for persistent downturn in state employment. This can be also of particular interest due to the fact that the raw minimum is highly unstable over time: among the 28 states, 25 attain the exact minimum at least once, and the minimum state changes in 41.5% of adjacent months. We can therefore apply our procedure for sparsity recovery in Section 4, with the tuning parameter selected by the modified C_p criterion in Section 4.3. DAFL exploits temporal persistence to recover a more stable zero set from the noisy lower-boundary estimates. Such the recovered zeros allow us to identify states with persistent lower-tail employment growth. For the sector main effect, we simply fix $\text{med}\{\boldsymbol{\beta}_t^*\} = 0$ and omit its interpretation because of space limitation.

Figure 5 compares the initial and final state-effect estimates. Panel A shows substantial month-to-month variation, whereas Panel B reveals persistent zero blocks associated with recognizable events, such as Michigan before the GFC, Nevada and Florida during the housing bust, West Virginia during the 2014–16 energy decline, and New York during the pandemic around 2020. We also report in Panel C the fraction of estimated zero sets. During the 2014–16 energy decline and the pandemic, such fractions are relatively narrow while much of the remaining cross section lies farther above the lower boundary. These large shocks widen cross-state differences and make the states with persistent low employment growth more distinct from the rest.

Next, we examine the estimated factor structure. To select the factor numbers (k_r, k_c), we experiment the estimators in Yu et al. (2022) and Section 4.6 of Lam and Cen (2025) (which can be viewed as Zhang et al. (2024) for matrix data) on $\widetilde{\mathbf{L}}_t$ in (2.6), with both resulting in $(\widehat{k}_r, \widehat{k}_c) = (1, 2)$. Figure 6 summarizes the estimated common component. Panel A shows a single state loading direction, with Oklahoma, Mississippi, West Virginia, and Texas toward one end and Washington, Oregon, California, and New York toward the other, indicating different patterns of industry-specific departures across states. In Panel B, one estimated industry-loading direction is dominated by mining, and its factor series moves sharply during the 2014–16 energy decline shown in Panel C. The second direction is dominated by arts and recreation, with a smaller contribution from accommodation and food services, and its factor shows significant movements during the pandemic, thus showing how the life

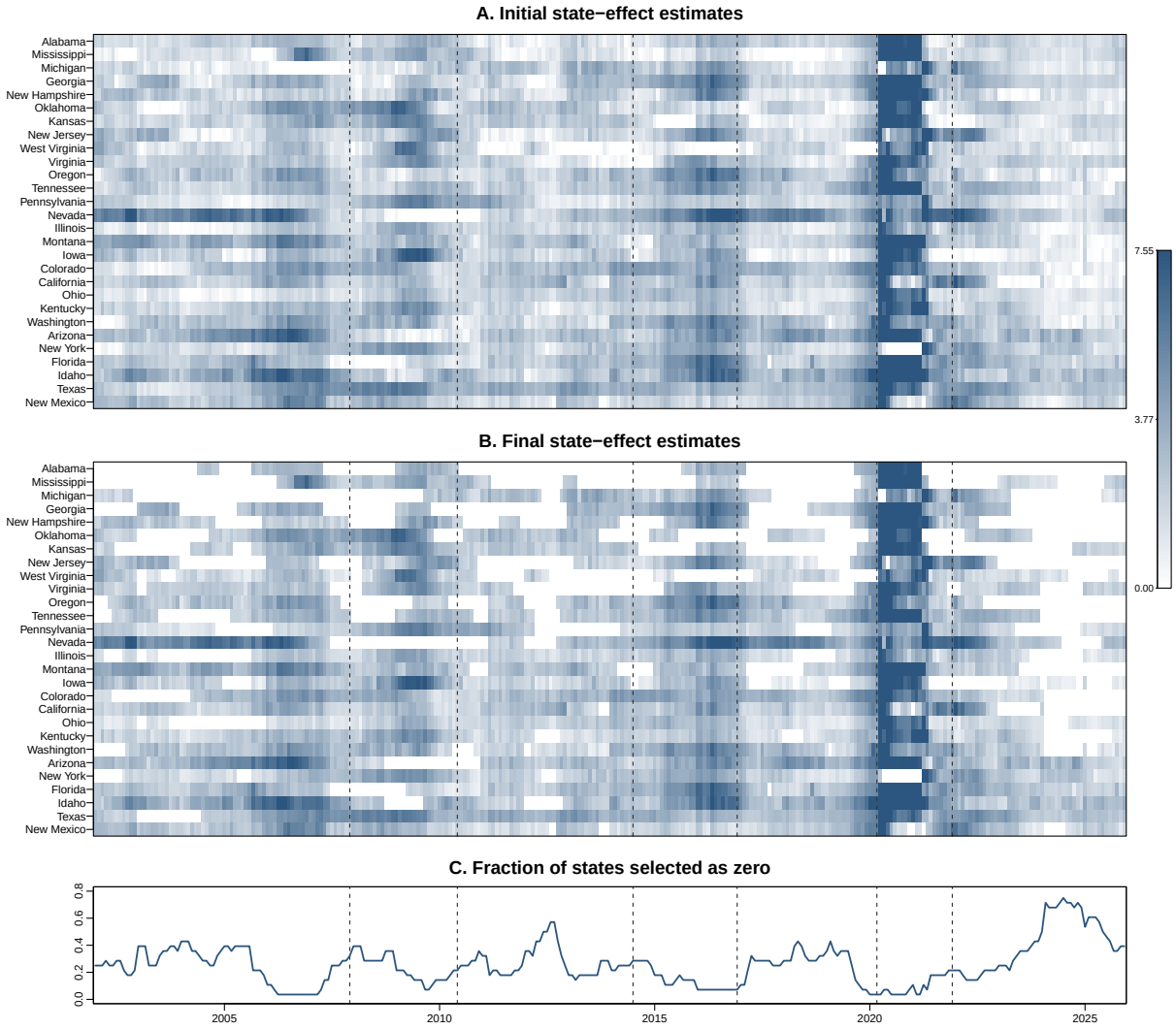

Figure 5: Initial and final state-effect estimates in the QCEW data. Panels A and B show the initial estimates and the final estimates after DAFL zero-set recovery, respectively, using a common color scale. Panel C reports the fraction of zero state effects over time. Dashed vertical lines enclose the periods representing the GFC (line 1–2), the 2014–16 energy decline (line 3–4), and the pandemic (line 5–6).

styles in U.S. were affected by the pandemic.

We also display in Panel D the magnitude of the estimated common components over time. From the diagram, the magnitude rises sharply during both episodes and reaches its sample maximum during the pandemic. The GFC is less prominent in the common component, compared to the peaks in the energy decline and the pandemic.

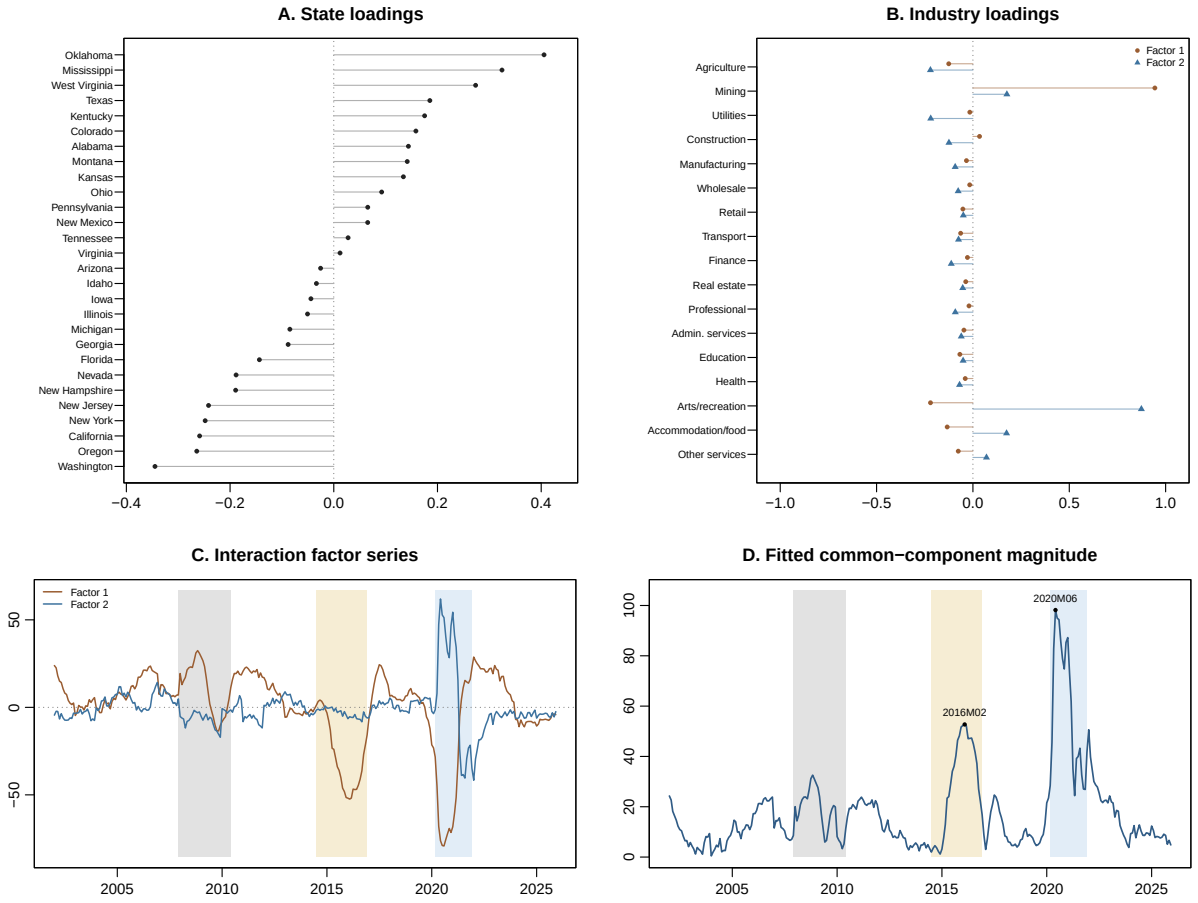

Figure 6: Estimated factor structure for the QCEW panel under $(\widehat{k}_r, \widehat{k}_c) = (1, 2)$. Panels A and B show the state and industry loadings, Panel C the factor series, and Panel D the Frobenius norm of the fitted common component. Shaded areas mark the GFC (gray), the 2014–16 energy decline (orange), and the pandemic period (blue).

Taken together, this QCEW application highlights a different role of identification from that in Section 1. The minimum identification makes the lower boundary of the state effects an object of interest, and DAFL turns its unstable monthly estimates into persistent slow employment growth that are more interpretable. The common component captures complementary state-industry interactions, most clearly during the energy decline and the pandemic.

5.3 A real application on macroeconomic indices

Motivated by Chen et al. (2020) and Yu et al. (2022), we study the quarterly macroeconomic data from Organization for Economic Co-operation and Development (OECD). At each quarter t, $\mathbf{X}_t \in \mathbb{R}^{8\times 10}$ contains 10 macroeconomic indicators for eight countries: the United States, the United Kingdom, Canada, France, Germany, Norway, Australia and New Zealand. The indicators consist of three price measures (total, energy and core CPI), two interest rates (3-month and long-term rates), three real-activity measures (GDP, industrial production and manufacturing production), and two trade measures (goods exports and imports). The data contains $T = 128$ quarters from 1988-Q3 to 2020-Q2.

For identification on the main effects, we use the median country and the real-activity group as our primary choice, such that

$$\text{Identification I}: \quad \text{med}\{\boldsymbol{\alpha}_t^*\} = 0, \quad (\beta^*_{t,\text{GDP}} + \beta^*_{t,\text{IP}} + \beta^*_{t,\text{Manuf}})/3 = 0.$$

Hence, the country effects are measured relative to a contemporaneous median country, while the indicator effects are measured relative to the average real-activity dynamics. Two alternative sets of identification are also considered: on the country side, we use the United States as a reference unit while retaining the real-activity group as above; and on the indicator side, we retain the median-country identification but use the three price indicators as the reference group. Formally,

$$\begin{aligned}
&\text{Identification II}: \quad \alpha^*_{t,\text{USA}} = 0, \quad (\beta^*_{t,\text{GDP}} + \beta^*_{t,\text{IP}} + \beta^*_{t,\text{Manuf}})/3 = 0; \\
&\text{Identification III}: \quad \text{med}\{\boldsymbol{\alpha}_t^*\} = 0, \quad (\beta^*_{t,\text{CPI}} + \beta^*_{t,\text{Energy}} + \beta^*_{t,\text{Core}})/3 = 0.
\end{aligned}$$

Figure 7 shows the estimated main effects under these identifications. Panels A and C report the estimated main effects under Identification I. In Panel A, around the global financial crisis (GFC) in 2008, the country main effect for the U.S. stepped below the median (of value zero) for several quarters, whereas that of Norway often moved in the opposite direction. In Panel C, indicator effects are measured relative to the average of GDP, industrial production and manufacturing production. Around GFC in 2008, it is clear to see that Goods imports, Goods exports, and all CPI indicators stood prominently above our real-activity benchmark, indicating that, for instance, imports and exports were less influenced by GFC. Also, the effects of Long-term rate and 3-month rate appeared more volatile in the early 1990s, while CPI energy stood out during the 2014–16 oil-price decline.

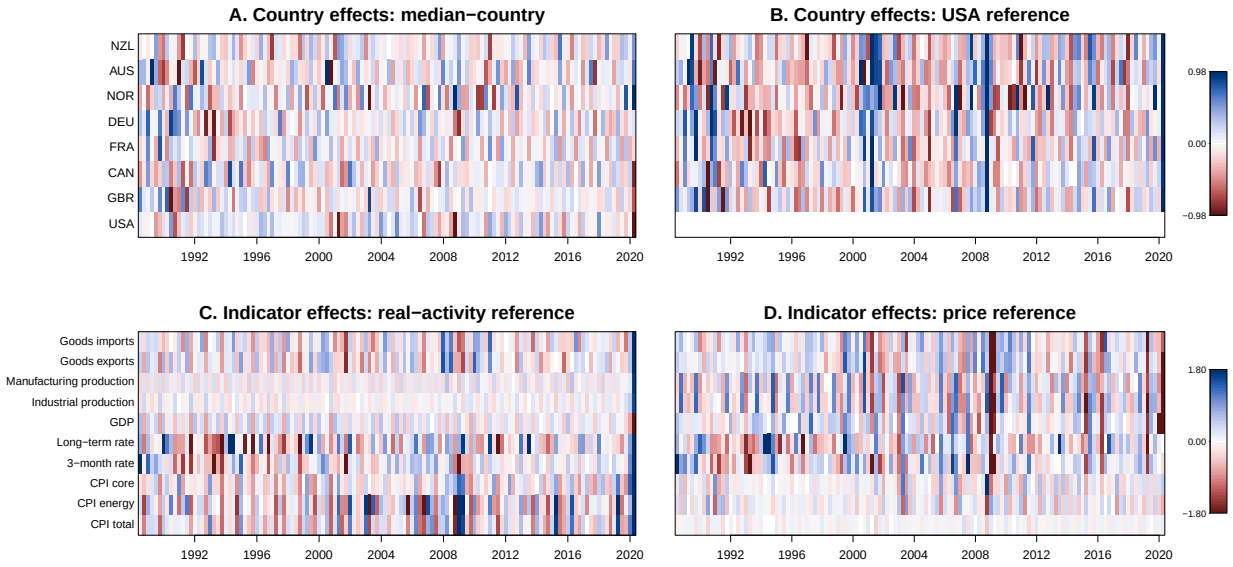

Figure 7: OECD main effects under different identifications. Panels A and B show country effects under the median-country and U.S. reference identification, holding the real-activity reference fixed. Panels C and D show indicator effects under the real-activity and price(-group) identification, holding the median-country (as in Panel A) fixed. Color scales are symmetrically clipped at the 98th percentile of absolute effect magnitudes for readability.

Panels B and D provide alternative perspectives under Identifications II and III. With the U.S. as the country reference in Panel B, all other countries' main effects appeared above the U.S. baseline, both during GFC around 2008 and the pandemic in 2020-Q2. With the price group as the indicator reference in Panel D, the influence by GFC around 2008 appeared in the opposite directions compared to that in Panel C, which is due to real-activity reference moving well below the price reference. Thus, the choice of identification could change the interpretations of main effects, while preserving the relative differences across countries and indicators.

We next examine the factor structure, which, markedly, is estimated the same regardless of our identification on the main effects. Similarly to Section 5.2, we apply the factor number estimators in Yu et al. (2022) and select $(\widehat{k}_r, \widehat{k}_c) = (1, 3)$. Figure 8 reports the estimated loadings and factor series. Panel A shows the estimated country loading matrix, where a geographical grouping can be seen from the signs of the elements. In Panel B, the first column can be regarded as CPI-related indices; the second is dominated by interest rates against trade; and the third represents industrial and manufacturing production with interest rates. Panel C displays when these factors become more active. For instance, The second factor is more variable in the early 1990s and around the GFC, while the first shows

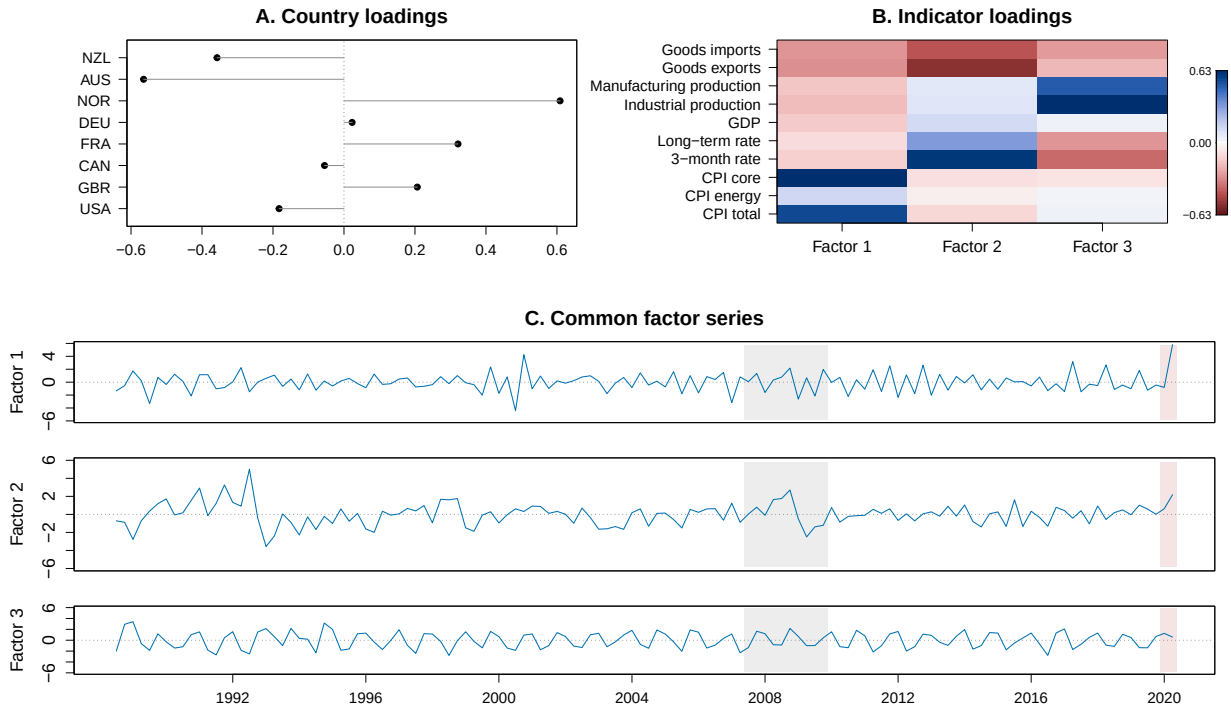

Figure 8: Estimated common factor structure for the OECD data. Panels A and B show the country and indicator loadings, and Panel C shows the three common factor series, with the GFC and pandemic periods shaded.

a large movement in 2020-Q2. Overall, the OECD application shows how general identification separates interpretable additive effects from a low-dimensional common component.

6 Concluding Remarks

In this paper, we address a critical gap in the literature of matrix factor models by developing a flexible framework for MEFM identification based on shift-varying functions, and by designing methodologies to recover sparse structures in the main effects. In particular, we reconcile the dual demands of interpretability for applied econometricians and structural parsimony inherent in matrix-valued time series. Our general identification theory allows practitioners to choose domain-appropriate constraints, from sum-to-zero to quantile-based or reference-unit anchors, while maintaining identifiability. Building on this, we propose a Doubly Adaptive Fused Lasso estimator that consistently recovers the sparse and dense blocks of the main effects, enabling a clear separation of periods with and without row- or

column-specific deviations.

Although we only consider sparsity in the main effects, it is possible to study approximately sparse loading matrices as in Assumption 2 in Freyaldenhoven (2022), or exactly sparse loadings as in Assumption 2 in Uematsu and Yamagata (2023a), among others. Their methods can be directly applied to the vectorization of the matrix time series $\widetilde{\mathbf{L}}_t$ ($t \in [T]$) defined in (2.6), at the cost of overlooking the matrix nature in each observed data matrix and inflating number of parameters. These observations demand further development of sparsity in matrix factor models. In fact, methodologies designed for the classical factor models are all applicable to $\widetilde{\mathbf{L}}_t$ ($t \in [T]$) in practice, and hence the approach proposed in this paper can be used as a powerful pre-processing of the observed data set. We hope that this article sheds light on sparsity structures in data analysis for matrix-valued or even higher-order tensor time series, and provides new insights in addressing dependency in complicated data structures.

References

Ando, T. and Bai, J. (2017). Clustering huge number of financial time series: A panel data approach with high-dimensional predictors and factor structures. *Journal of the American Statistical Association*, 112(519):1182–1198.

Ayvazyan, S. A. and Ulyanov, V. V. (2023). A multivariate clt for weighted sums with rate of convergence of order o(1/n). In Belomestny, D., Butucea, C., Mammen, E., Moulines, E., Reiß, M., and Ulyanov, V. V., editors, *Foundations of Modern Statistics*, pages 225–257, Cham. Springer International Publishing.

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

Barigozzi, M. and Hallin, M. (2024). The dynamic, the static, and the weak: Factor models and the analysis of high-dimensional time series. *arXiv preprint arXiv:2407.10653v3*.

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577 – 2604.

Cen, Z. and Lam, C. (2025). Tensor time series imputation through tensor factor modelling. *Journal of Econometrics*, 249:105974.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1281–1304.

Chen, E. Y. and Fan, J. (2023). Statistical inference for high-dimensional matrix-variate factor models. *Journal of the American Statistical Association*, 118(542):1038–1055.

Chen, E. Y., Tsay, R. S., and Chen, R. (2020). Constrained factor models for high-dimensional matrix-variate time series. *Journal of the American Statistical Association*, 115(530):775–793.

Chen, R., Yang, D., and Zhang, C.-H. (2022). Factor models for high-dimensional tensor time series. *Journal of the American Statistical Association*, 117(537):94–116.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81(394):461–470.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680.

Fan, J., Lou, Z., and Yu, M. (2024). Are latent factor regression and sparse regression adequate? *Journal of the American Statistical Association*, 119(546):1076–1088.

Fan, J., Masini, R. P., and Medeiros, M. C. (2023). Bridging factor and sparse models. *The Annals of Statistics*, 51(4):1692 – 1717.

Freyaldenhoven, S. (2022). Factor models with local factors — determining the number of relevant factors. *Journal of Econometrics*, 229(1):80–102.

He, Y., Kong, X., Trapani, L., and Yu, L. (2023). One-way or two-way factor model for matrix sequences? *Journal of Econometrics*, 235(2):1981–2004.

He, Y., Kong, X., Yu, L., Zhang, X., and Zhao, C. (2024). Matrix factor analysis: From least squares to iterative projection. *Journal of Business & Economic Statistics*, 42(1):322–334.

Hirzel, A. H., Hausser, J., Chessel, D., and Perrin, N. (2002). Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? *Ecology*, 83(7):2027–2036.

Hochreiter, S., Clevert, D.-A., and Obermayer, K. (2006). A new summarization method for affymetrix probe level data. *Bioinformatics*, 22(8):943–949.

Hu, J., Li, T., and Wang, X. (2025). Aggregated projection method: A new approach for group factor model. *Journal of the American Statistical Association*, 0(0):1–13.

Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *statistica sinica*, pages 1603–1618.

Lam, C. and Cen, Z. (2025). Matrix-valued factor model with time-varying main effects. *Journal of Econometrics*, 252:106105.

Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *The Annals of Statistics*, 40(2):694–726.

Lam, C., Yao, Q., and Bathia, N. (2011). Estimation of latent factors for high-dimensional time series. *Biometrika*, 98(4):901–918.

Mallows, C. L. (1973). Some comments on cp. *Technometrics*, 15(4):661–675.

McCrae, R. R. and John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, 60(2):175–215.

Mosley, L., Chan, T.-S. T., and Gibberd, A. (2024). The sparse dynamic factor model: a regularised quasi-maximum likelihood approach. *Statistics and Computing*, 34(68).

Pan, J. and Yao, Q. (2008). Modelling multiple time series via common factors. *Biometrika*, 95(2):365–379.

Rinaldo, A. (2009). Properties and refinements of the fused lasso. *The Annals of Statistics*, 37(5B):2922–2952.

Stock, J. H. and Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

Stock, J. H. and Watson, M. W. (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162.

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

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(1):91–108.

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

Uematsu, Y. and Yamagata, T. (2023a). Estimation of sparsity-induced weak factor models. *Journal of Business & Economic Statistics*, 41(1):213–227.

Uematsu, Y. and Yamagata, T. (2023b). Inference in sparsity-induced weak factor models. *Journal of Business & Economic Statistics*, 41(1):126–139.

Wang, D., Liu, X., and Chen, R. (2019). Factor models for matrix-valued high-dimensional time series. *Journal of Econometrics*, 208(1):231–248. Special Issue on Financial Engineering and Risk Management.

Wang, P. (2008). Large dimensional factor models with a multi-level factor structure: Identification, estimation, and inference.

Wei, J. and Zhang, Y. (2024). Can principal component analysis preserve the sparsity in factor loadings? *arXiv preprint arXiv:2305.05934v2*.

Yu, L., He, Y., Kong, X., and Zhang, X. (2022). Projected estimation for large-dimensional matrix factor models. *Journal of Econometrics*, 229(1):201–217.

Zhang, B., Pan, G., Yao, Q., and Zhou, W. (2024). Factor modeling for clustering high-dimensional time series. *Journal of the American Statistical Association*, 119(546):1252–1263.

Zhang, S., Shen, Y., Chen, I. A., and Lee, J. (2025). Sparse bayesian group factor model for feature interactions in multiple count tables data. *Journal of the American Statistical Association*, 120(550):723–736.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173 – 2192.

Supplementary materials for the paper "Main Effect Factor Models in High-Dimensional Matrix Time Series: Identification and Sparsity"

A Additional Experiments and Details

A.1 Additional simulation experiments

Accuracy of grand mean and main effects Table A.1 examines the accuracy of the grand mean and main effect estimators when the Gaussian noise innovations are replaced by standardized t_5 and t_3 innovations. We consider representative identifications from both $\mathcal{L}$ and $\mathcal{Q}$. Overall, the patterns under Gaussian noise are well preserved as the tails become heavier. The equal-weight identification remains the most accurate among the linear identifications, while the reference-unit identification has errors similar to those under the minimum identification.

The central and tail quantile identifications respond differently to heavier tails. The MSEs under the median identification remain nearly unchanged from Gaussian to t_3 noise, including their upper replication-level quantiles. In contrast, the minimum identification shows some deterioration under t_3 noise, most clearly for the grand mean. This reflects the greater sensitivity of an extreme quantile to large noise realizations. The results therefore also illustrate the relative robustness of using a central quantile as the identifying benchmark.

Accuracy of common components Table A.2 repeats the common component experiment using standardized t_3 noise innovations. The main qualitative patterns are preserved. Increasing T and the cross-sectional dimensions generally improves estimation accuracy, while weak factors lead to larger loading-space errors.

Compared with the Gaussian setting, the deterioration is mild under pervasive factors in most cases. The loading-space errors remain of similar magnitude, and the common component is also estimated accurately in most settings. The effect of heavy tails is more pronounced under weak factors. Both loading-space errors and common-component MSEs increase substantially relative to the Gaussian setting. Increasing T still improves the results in general.

Table A.1: Accuracy under heavy-tailed noise. Entries report MSE means multiplied by 10^2, with the 95th percentile of replication-level MSEs in brackets.

Noise	Identification	MSE($\widetilde{\mu}$)	MSE($\widetilde{\boldsymbol{\alpha}}$)	MSE($\widetilde{\boldsymbol{\beta}}$)
Gaussian	$g(\mathbf{a}) = a_1$	6.439[10.634]	6.454[8.899]	6.504[9.220]
	$g(\mathbf{a}) = \mathbf{1}_d^\intercal \mathbf{a}$	0.111[0.172]	3.246[3.728]	3.257[3.667]
	$g(\mathbf{a}) = \min\{\mathbf{a}\}$	6.149[8.145]	6.151[7.358]	6.185[7.410]
	$g(\mathbf{a}) = \text{med}\{\mathbf{a}\}$	1.621[2.043]	4.003[4.549]	4.008[4.522]
t_5	$g(\mathbf{a}) = a_1$	6.574[11.191]	6.477[9.494]	6.505[9.321]
	$g(\mathbf{a}) = \mathbf{1}_d^\intercal \mathbf{a}$	0.109[0.173]	3.262[3.710]	3.262[3.703]
	$g(\mathbf{a}) = \min\{\mathbf{a}\}$	6.219[8.238]	6.199[7.476]	6.190[7.420]
	$g(\mathbf{a}) = \text{med}\{\mathbf{a}\}$	1.630[2.075]	4.021[4.571]	4.018[4.614]
t_3	$g(\mathbf{a}) = a_1$	6.558[11.406]	6.361[9.164]	6.527[9.644]
	$g(\mathbf{a}) = \mathbf{1}_d^\intercal \mathbf{a}$	0.112[0.180]	3.217[3.791]	3.218[3.780]
	$g(\mathbf{a}) = \min\{\mathbf{a}\}$	6.971[8.435]	6.341[7.604]	6.336[7.784]
	$g(\mathbf{a}) = \text{med}\{\mathbf{a}\}$	1.552[1.980]	3.936[4.601]	3.946[4.630]

Table A.2: Loading-space and common component estimation accuracy under standardized t_3 noise innovations, where $\zeta = \zeta_r = \zeta_c$. Entries are Monte Carlo means multiplied by 10^2, with standard errors in parentheses.

p	q	ζ	$\mathcal{D}(\mathbf{Q}_r, \widehat{\mathbf{Q}}_r)$		$\mathcal{D}(\mathbf{Q}_c, \widehat{\mathbf{Q}}_c)$		MSE($\widehat{\mathbf{C}}$)	
			$T = 100$	$T = 200$	$T = 100$	$T = 200$	$T = 100$	$T = 200$
30	30	0	2.346(0.200)	1.652(0.042)	2.217(0.060)	1.610(0.029)	0.764(0.085)	0.575(0.008)
		0.2	26.713(1.052)	21.815(0.928)	26.205(1.049)	21.693(0.958)	5.540(1.021)	3.632(0.468)
80	80	0	1.277(0.196)	0.761(0.007)	1.284(0.195)	0.754(0.007)	1.653(1.521)	0.097(0.001)
		0.2	21.170(1.074)	15.785(0.934)	21.196(1.101)	15.648(0.927)	1.782(0.235)	1.466(0.329)
30	80	0	1.406(0.046)	1.290(0.152)	1.831(0.020)	1.379(0.076)	0.316(0.009)	0.417(0.106)
		0.2	24.381(1.005)	20.455(0.928)	21.308(0.981)	16.764(0.932)	2.898(0.635)	2.205(0.477)

A.2 Additional details on real data applications

QCEW. For our QCEW data studied in the main text, the industries cover agriculture, mining, utilities, construction, manufacturing, wholesale and retail trade, transportation, finance, real estate, professional and administrative services, education, health care, arts and recreation, accommodation and food services, and other services, all classified according to the North American Industry Classification System. The observation for each state–

industry pair is computed as the year-over-year log employment growth, i.e.,

$$X_{t,ij} = 100\,(\log N_{t,ij} - \log N_{t-12,ij}),$$

where $N_{t,ij}$ denotes monthly employment.

OECD. We process the OECD raw data as follows. Following the pre-processing adopted in Chen et al. (2020) and Yu et al. (2022), logarithmic and differencing transformations are applied: the CPI series are second log-differenced, the interest rates are first-differenced, the real-activity series are first log-differenced, and exports and imports are transformed into growth rates. Each transformed country–indicator series is further standardized over time to remove scale differences across heterogeneous macroeconomic variables.

A.3 Additional real data application on taxi trips

We illustrate the proposed method for MEFM using the publicly available New York City (NYC) Yellow Taxi Trip Records; see the details in https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Each raw trip record includes pick-up/drop-off timestamps, pick-up/drop-off location codes, trip distance, fare components, payment information, etc. The Taxi and Limousine Commission's map partitions NYC into 265 zones, and we focus on the 69 zones on Manhattan Island which accounts for the vast majority of trip activity. Let $\mathbf{X}_t \in \mathbb{R}^{69\times 24}$ denote the matrix for calendar day t. Its (i,j)-th entry counts all trips with drop-off zone i and pick-up time in the j-th hourly slot. We analyze the period from 1 January 2019 to 31 December 2022, which spans the dramatic mobility collapse during the 2020 pandemic lockdown and the subsequent recovery. To respect the distinct spatial/temporal rhythms of the city, we further split the series into weekday and weekend subsets, containing 1044 and 417 days, respectively, and estimate the model on each subset separately. In this application, we impose the minimum identification

$$g_\alpha(\mathbf{a}) = \min\{\mathbf{a}\} \quad \text{and} \quad g_\beta(\mathbf{b}) = \min\{\mathbf{b}\},$$

so that the location and hour main effects are nonnegative and, at each time point, their minimum is normalized to zero. Accordingly, a zero main effect indicates the absence of an additional location- or hour-specific additive effect relative to the estimated baseline, rather than the absence of taxi trips. The tuning parameters are chosen via the modified C_p criterion described in Section 4.3.

In Figure A.1, the location main effects for weekdays and weekends are demonstrated. Firstly, an abrupt shift from dark red to pale white in mid-March 2020 signals the pandemic lockdown, during which the location main effect estimators for almost all Manhattan zones collapse to zero for several months. Secondly, the rebound is asymmetric: weekday activity re-emerges more quickly than weekend activity in most zones, reflecting a faster return to work routines while leisure remains depressed. Thirdly, several locations—Liberty, Ellis, Governors, and Randalls Islands—stay faint almost throughout, yet the model still detects sporadic taxi traffic. On the other hand, zones such as Harlem, Hamilton Heights, Manhattan Valley, and Washington Heights preserve moderate intensities even during the pandemic, underscoring local demand. Finally, neither weekdays nor weekends regain their 2019 pre-pandemic intensity by late 2021–2022, demonstrating the impact of the pandemic on Manhattan mobility and, by extension, urban economic and lifestyle patterns. Notably, the fused penalty is what allows the model to delineate the prolonged "silent" pandemic interval, yet it remains sensitive enough to uncover the faint residual taxi activity that persisted throughout the lockdown.

To further demonstrate the spatial pattern, Figure A.2 offers a snapshot of how the estimated location main effects evolved across Manhattan. Six representative days are selected to demonstrate the before, during, and after the pandemic period for weekdays and weekends. Before the pandemic, both rows are covered by dark reds, but their centers of taxi rides differ: weekdays peak in business core in Midtown whereas weekends tilt toward leisure areas in the West Village, SoHo and the Lower East Side. However, during lockdown, the island converts almost uniformly pale, and only a few residential and hospital-adjacent blocks in Upper Manhattan retain faint activity during weekday, while the corresponding weekend map is blank, underscoring the sharper collapse of leisure travel. After a year of rehabilitation, demand partially returns, but the asymmetry remains: weekend drop-offs stay muted across several entertainment districts downtown, and weekday rebounds more strongly, especially in residential uptown zones.

Furthermore, the hour main effect is displayed in Figure A.3. We may observe that weekday mobility peaks in the late-afternoon commute (3pm–6pm) and tapers off by 1am, whereas weekends sustain a pronounced nightlife pulse until about 4am and begin later in the morning. In addition, the lockdown starts in mid-March 2020 and erases nearly all hourly traffic for several weeks. The weekend panel stays blank longer, demonstrating a sharper hit to entertainment travel. Moreover, recovery unfolds asymmetrically, where weekday evening activity re-emerges first (from mid-June 2020), with the signal band sliding from 5pm down to midnight over the next year. Weekend nightlife restarts only in the early autumn of 2020, and does not regain its pre-pandemic 4am endpoint by 2022, indicating a lasting contraction of late-night leisure demand.

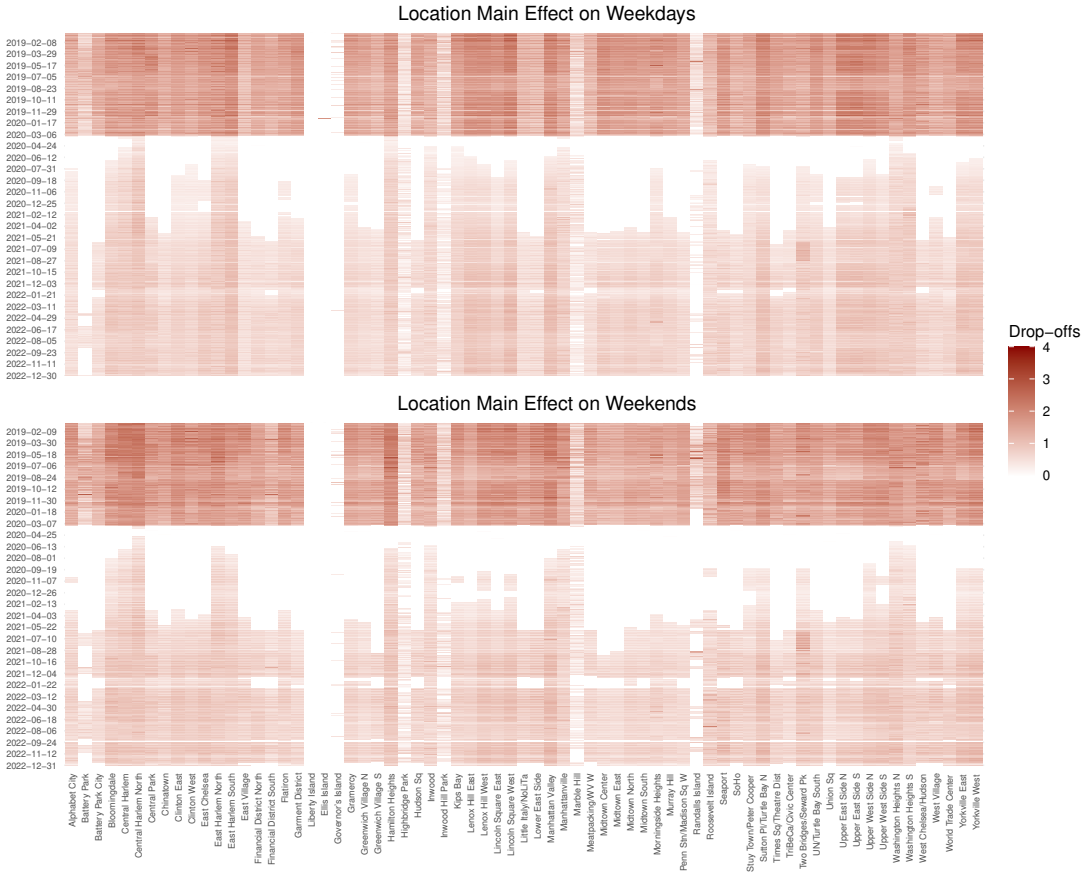

Figure A.1: Heatmap of the location main effect $\widetilde{\boldsymbol{\alpha}}_l^{\mathcal{P}}$.

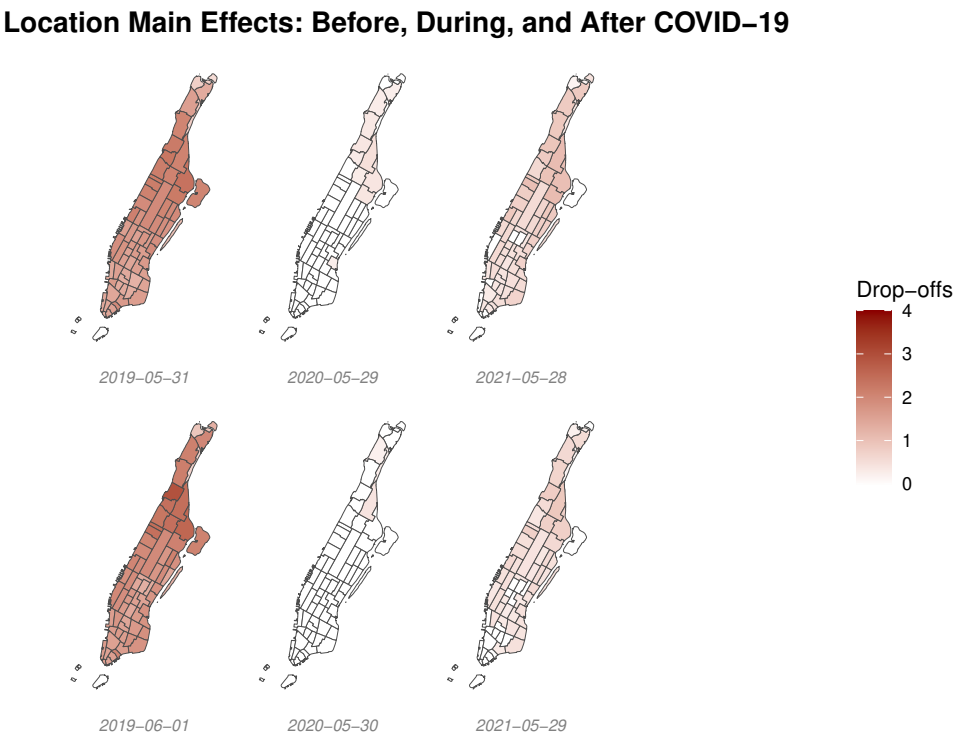

Figure A.2: Snapshot heatmaps of location main effect. The top rows are weekdays, and the bottom rows are weekends. The color scale matches Figure A.1.

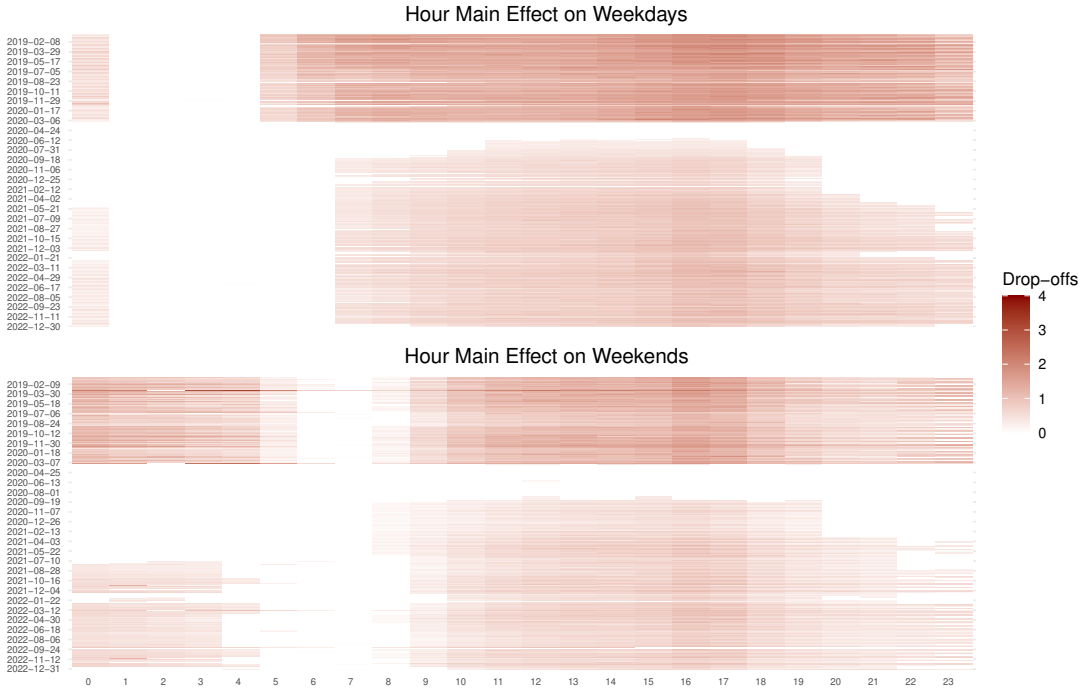

Figure A.3: Heatmap of the hour main effect $\widetilde{\boldsymbol{\beta}}_t^{\mathcal{B}}$.

B Technical Proofs

B.1 Proof of theorems

Proof of Theorem 1. First, consider part (i) in the statement. Suppose we have another set of parameters, $\left(\ddot{\mu}_t, \ddot{\boldsymbol{\alpha}}_t, \ddot{\boldsymbol{\beta}}_t, \ddot{\mathbf{C}}_t\right)$ for $t \in [T]$, also satisfying (1.1). For each $t \in [T]$, right-multiply $\mathbf{1}_q$ on both sides of (1.1), by Assumption (IC1) (i), it holds that

$$\begin{aligned}\mathbf{X}_t\mathbf{1}_q &= q\boldsymbol{\alpha}_t^* + \mathbf{1}_p(q\mu_t + \mathbf{1}_q^\intercal\boldsymbol{\beta}_t^*) + \mathbf{E}_t\mathbf{1}_q \\ &= q\ddot{\boldsymbol{\alpha}}_t + \mathbf{1}_p(q\ddot{\mu}_t + \mathbf{1}_q^\intercal\ddot{\boldsymbol{\beta}}_t) + \mathbf{E}_t\mathbf{1}_q.\end{aligned}$$

Note that both $(q\mu_t + \mathbf{1}_q^\intercal \boldsymbol{\beta}_t^*)$ and $(q\ddot{\mu}_t + \mathbf{1}_q^\intercal \ddot{\boldsymbol{\beta}}_t)$ are scalars, we obtain $\boldsymbol{\alpha}_t^* - \ddot{\boldsymbol{\alpha}}_t = c_t \mathbf{1}_p$ for some scalar c_t. This implies each row main effect is at least identified up to a shift. Then by Assumption (IC1) (ii), we have

$$0 = g_\alpha(\boldsymbol{\alpha}_t^*) = g_\alpha(\ddot{\boldsymbol{\alpha}}_t + c_t \mathbf{1}_p),$$

which immediately implies $c_t = 0$ by (IC1) (iii), and thus $\boldsymbol{\alpha}_t^* = \ddot{\boldsymbol{\alpha}}_t$. For the column main effects, all the above arguments follow analogously by left-multiplying $\mathbf{1}_p^\intercal$ on both sides of (1.1), and hence $\boldsymbol{\beta}_t^* = \ddot{\boldsymbol{\beta}}_t$. For each $t \in [T]$, by respectively left- and right-multiplying $\mathbf{1}_p^\intercal$ and $\mathbf{1}_q$ on both sides of (1.1), we obtain

$$\begin{aligned}
\mathbf{1}_p^\intercal \mathbf{X}_t \mathbf{1}_q &= pq\mu_t + q\mathbf{1}_p^\intercal \boldsymbol{\alpha}_t^* + p\mathbf{1}_q^\intercal \boldsymbol{\beta}_t^* + \mathbf{1}_p^\intercal \mathbf{E}_t \mathbf{1}_q \\
&= pq\ddot{\mu}_t + q\mathbf{1}_p^\intercal \ddot{\boldsymbol{\alpha}}_t + p\mathbf{1}_q^\intercal \ddot{\boldsymbol{\beta}}_t + \mathbf{1}_p^\intercal \mathbf{E}_t \mathbf{1}_q,
\end{aligned}$$

so that $\mu_t = \ddot{\mu}_t$ given all main effects are identified previously. Lastly, the remaining components $\mathbf{C}_t$ and $\ddot{\mathbf{C}}_t$ trivially coincide for all $t \in [T]$. This completes the "if" part in the statement of this theorem.

To show the "only if" part, it suffices to show that Assumption (IC1) (iii) is necessary to identify the row main effects. Suppose Assumption (IC1) (iii) does not hold. Namely, there exists some vector $\mathbf{a}$ with $g_\alpha(\mathbf{a}) = 0$ and non-zero constant d such that $g_\alpha(\mathbf{a} + d\mathbf{1}_p) = 0$. This implies both row main effects $\mathbf{a}$ and $\mathbf{a} + d\mathbf{1}_p$ fulfill Assumption (IC1) (ii), but $d \neq 0$

and hence those two vectors are not the same. The row main effect is thus unidentifiable in this case, indicating that Assumption (IC1) (iii) is necessary in identifying the row main effects. This completes the proof of part (i) in the theorem.

For part (ii), let $\mathbf{X}_t = \acute{\mathbf{A}}_r \mathbf{F}_t \acute{\mathbf{A}}_c^\intercal + \mathbf{E}_t =: \acute{\mathbf{C}}_t + \mathbf{E}_t$. For each $t \in [T]$, define $c_{\alpha,t}$ as the scalar such that

$$g_\alpha(q^{-1}\mathbf{M}_p \acute{\mathbf{C}}_t \mathbf{1}_q + c_{\alpha,t}\mathbf{1}_p) = 0,$$

with $c_{\alpha,t}$ guaranteed to exist according to the condition in the statement of Theorem 1 (ii). Define analogously $c_{\beta,t}$. Then we can construct

$$\begin{aligned}
\mathbf{A}_r &= \mathbf{M}_p \acute{\mathbf{A}}_r = (\mathbf{I}_p - p^{-1}\mathbf{1}_p\mathbf{1}_p^\intercal)\acute{\mathbf{A}}_r, \quad \boldsymbol{\alpha}_t^* = q^{-1}\mathbf{M}_p \acute{\mathbf{C}}_t \mathbf{1}_q + \mathbf{1}_p c_{\alpha,t}, \\
\mathbf{A}_c &= \mathbf{M}_q \acute{\mathbf{A}}_c = (\mathbf{I}_q - q^{-1}\mathbf{1}_q\mathbf{1}_q^\intercal)\acute{\mathbf{A}}_c, \quad \boldsymbol{\beta}_t^* = p^{-1}\mathbf{M}_q \acute{\mathbf{C}}_t^\intercal \mathbf{1}_p + \mathbf{1}_q c_{\beta,t}, \\
\mu_t &= (pq)^{-1}\mathbf{1}_p^\intercal \acute{\mathbf{C}}_t \mathbf{1}_q - c_{\alpha,t} - c_{\beta,t}.
\end{aligned}$$

By Remark 1 in Lam and Cen (2025), it holds immediately that

$$\begin{aligned}
\acute{\mathbf{C}}_t &= (pq)^{-1}(\mathbf{1}_p^\intercal \acute{\mathbf{C}}_t \mathbf{1}_q)\mathbf{1}_p\mathbf{1}_q^\intercal + q^{-1}(\mathbf{M}_p \acute{\mathbf{C}}_t \mathbf{1}_q)\mathbf{1}_q^\intercal + p^{-1}\mathbf{1}_p(\mathbf{M}_q \acute{\mathbf{C}}_t^\intercal \mathbf{1}_p)^\intercal + (\mathbf{M}_p \acute{\mathbf{A}}_r)\mathbf{F}_t(\mathbf{M}_q \acute{\mathbf{A}}_c)^\intercal \\
&= \mu_t \mathbf{1}_p\mathbf{1}_q^\intercal + \boldsymbol{\alpha}_t^*\mathbf{1}_q^\intercal + \mathbf{1}_p\boldsymbol{\beta}_t^{*\intercal} + \mathbf{A}_r\mathbf{F}_t\mathbf{A}_c^\intercal,
\end{aligned}$$

as in (1.1). By the definition of $c_{\alpha,t}$ and $c_{\beta,t}$, it is easy to see $g_\alpha(\boldsymbol{\alpha}_t^*) = 0$ and $g_\beta(\boldsymbol{\beta}_t^*) = 0$, thus satisfying Assumption (IC1) (ii). Assumption (IC1) (i) is indeed satisfied by $\mathbf{M}_a\mathbf{1}_a = \mathbf{0}$

and hence

$$\mathbf{1}_p^\intercal(\mathbf{M}_p\acute{\mathbf{A}}_r) = \mathbf{0}, \quad \mathbf{1}_q^\intercal(\mathbf{M}_q\acute{\mathbf{A}}_c) = \mathbf{0}.$$

For completeness, Assumption (IC1) (iii) holds automatically. This completes the proof of the theorem. □

Proof of Theorem 2. First, consider the estimator of row main effects. Fix any $t \in [T]$. To begin with, by Lemma B.1 (i), we have

$$\mathbb{E}\big[\|q^{-1}\mathbf{E}_t\mathbf{1}_q\|^2\big] = q^{-2}\sum_{i=1}^{p}\mathrm{Var}\Big(\sum_{j=1}^{q}E_{t,ij}\Big) = O(pq^{-1}). \tag{B.1}$$

As discussed in Section 2.2, using Assumptions (IC1) and (IC2), it holds that

$$\begin{aligned}\boldsymbol{\alpha}_t^* &= q^{-1}\mathbf{X}_t\mathbf{1}_q - \mathbf{1}_p(\mu_t + q^{-1}\mathbf{1}_q^\intercal\boldsymbol{\beta}_t^*) - q^{-1}\mathbf{E}_t\mathbf{1}_q \\ &= q^{-1}\mathbf{X}_t\mathbf{1}_q - \mathbf{1}_p\cdot\widetilde{g}_\alpha(q^{-1}\mathbf{X}_t\mathbf{1}_q - q^{-1}\mathbf{E}_t\mathbf{1}_q) - q^{-1}\mathbf{E}_t\mathbf{1}_q.\end{aligned}$$

Together with (2.3), we have

$$\widetilde{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}_t^* = \mathbf{1}_p\cdot\big\{\widetilde{g}_\alpha(q^{-1}\mathbf{X}_t\mathbf{1}_q - q^{-1}\mathbf{E}_t\mathbf{1}_q) - \widetilde{g}_\alpha(q^{-1}\mathbf{X}_t\mathbf{1}_q)\big\} + q^{-1}\mathbf{E}_t\mathbf{1}_q. \tag{B.2}$$

This gives

$$\frac{1}{p}\|\widetilde{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}_t^*\|^2 = \frac{1}{p}\big\|\mathbf{1}_p\cdot\big\{\widetilde{g}_\alpha(q^{-1}\mathbf{X}_t\mathbf{1}_q - q^{-1}\mathbf{E}_t\mathbf{1}_q) - \widetilde{g}_\alpha(q^{-1}\mathbf{X}_t\mathbf{1}_q)\big\} + q^{-1}\mathbf{E}_t\mathbf{1}_q\big\|^2$$

$$\leq 2\left\{\widetilde{g}_\alpha(q^{-1}\mathbf{X}_t\mathbf{1}_q - q^{-1}\mathbf{E}_t\mathbf{1}_q) - \widetilde{g}_\alpha(q^{-1}\mathbf{X}_t\mathbf{1}_q)\right\}^2 + \frac{2}{p}\left\|q^{-1}\mathbf{E}_t\mathbf{1}_q\right\|^2$$
$$\leq 2L_\alpha^2\left\|q^{-1}\mathbf{E}_t\mathbf{1}_q\right\|_*^2 + \frac{2}{p}\left\|q^{-1}\mathbf{E}_t\mathbf{1}_q\right\|^2 = O_P\left(\frac{L_\alpha^2}{q^2}\left\|\mathbf{E}_t\mathbf{1}_q\right\|_*^2 + \frac{1}{q}\right).$$

where the second inequality used Assumption (S1) and the notations therein, and the last equality used (B.1). Analogous arguments follow for the estimator of the column main effects.

Next, consider the grand mean estimator at $t \in [T]$. For this, note that by (B.2) and Assumption (S1) again, we have

$$\begin{aligned}\frac{1}{p^2}\left\{\mathbf{1}_p^\intercal(\widetilde{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}_t^*)\right\}^2 &= \frac{1}{p^2}\left[p\left\{\widetilde{g}_\alpha(q^{-1}\mathbf{X}_t\mathbf{1}_q - q^{-1}\mathbf{E}_t\mathbf{1}_q) - \widetilde{g}_\alpha(q^{-1}\mathbf{X}_t\mathbf{1}_q)\right\} + q^{-1}\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{1}_q\right]^2 \\ &\leq 2\left\{\widetilde{g}_\alpha(q^{-1}\mathbf{X}_t\mathbf{1}_q - q^{-1}\mathbf{E}_t\mathbf{1}_q) - \widetilde{g}_\alpha(q^{-1}\mathbf{X}_t\mathbf{1}_q)\right\}^2 + \frac{2}{p^2q^2}(\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{1}_q)^2 \\ &\leq 2L_\alpha^2\left\|q^{-1}\mathbf{E}_t\mathbf{1}_q\right\|_*^2 + \frac{2}{p^2q^2}(\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{1}_q)^2. \end{aligned} \tag{B.3}$$

Similarly, it holds that

$$\frac{1}{q^2}\left\{\mathbf{1}_q^\intercal(\widetilde{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^*)\right\}^2 \leq 2L_\beta^2\left\|p^{-1}\mathbf{E}_t^\intercal\mathbf{1}_p\right\|_*^2 + \frac{2}{p^2q^2}(\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{1}_q)^2. \tag{B.4}$$

Recalling $\mu_t = (pq)^{-1}\mathbf{1}_p^\intercal\mathbf{X}_t\mathbf{1}_q - p^{-1}\mathbf{1}_p^\intercal\boldsymbol{\alpha}_t^* - q^{-1}\mathbf{1}_q^\intercal\boldsymbol{\beta}_t^* - (pq)^{-1}\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{1}_q$ and the construction in

(2.5), we have

$$
\begin{aligned}
(\widetilde{\mu}_t - \mu_t)^2 &= \left\{p^{-1}\mathbf{1}_p^\intercal(\widetilde{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}_t^*) + q^{-1}\mathbf{1}_q^\intercal(\widetilde{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^*) - (pq)^{-1}\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{1}_q\right\}^2 \\
&\lesssim \frac{1}{p^2}\left\{\mathbf{1}_p^\intercal(\widetilde{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}_t^*)\right\}^2 + \frac{1}{q^2}\left\{\mathbf{1}_q^\intercal(\widetilde{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^*)\right\}^2 + \frac{1}{p^2q^2}(\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{1}_q)^2 \\
&\lesssim L_\alpha^2\left\|q^{-1}\mathbf{E}_t\mathbf{1}_q\right\|_*^2 + L_\beta^2\left\|p^{-1}\mathbf{E}_t^\intercal\mathbf{1}_p\right\|_*^2 + \frac{1}{p^2q^2}(\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{1}_q)^2 \\
&= O_P\left(\frac{L_\alpha^2}{q^2}\left\|\mathbf{E}_t\mathbf{1}_q\right\|_*^2 + \frac{L_\beta^2}{p^2}\left\|\mathbf{E}_t^\intercal\mathbf{1}_p\right\|_*^2 + \frac{1}{pq}\right).
\end{aligned}
$$

where the third line used (B.3) and (B.4), and the last equality used the fact that

$$
\left((pq)^{-1}\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{1}_q\right)^2 = O_P(p^{-1}q^{-1})
$$

from the proof of Theorem 2 in Lam and Cen (2025). This completes the proof for the grand mean estimator and hence finishes this proof. □

Proof of Theorem 3. Before any part, first consider a series of sub-Gaussian random variables $\{X_i\}_{i=1}^p$ with zero mean and variance proxy σ^2. For any $\lambda \geq 0$, by the Jensen's inequality, we have

$$
\begin{aligned}
\exp\{\lambda\mathbb{E}(\max_{i\in[p]}\{X_i\})\} &\leq \mathbb{E}(\exp\{\lambda\max_{i\in[p]}\{X_i\}\}) \leq \mathbb{E}(\max_{i\in[p]}\{\exp\{\lambda X_i\}\}) \\
&\leq \sum_{i=1}^p \mathbb{E}(e^{\lambda X_i}) \leq pe^{\lambda^2\sigma^2/2},
\end{aligned}
$$

implying that $\mathbb{E}(\max_{i\in[p]}\{X_i\}) \leq \log(p)/\lambda + \lambda\sigma^2/2$. Choosing $\lambda = \sqrt{2\log(p)}/\sigma$ yields $\mathbb{E}(\max_{i\in[p]}\{X_i\}) \leq \sqrt{2\sigma^2\log(p)}$. Therefore, we can conclude that $\mathbb{E}(\max_{i\in[p]}|X_i|) \leq \sqrt{2\sigma^2\log(2p)}$. With the above argument, note that for each $i \in [p]$, $X_i = \sum_{j=1}^{q} E_{t,ij}$ is a sub-Gaussian random variables with variance proxy $\sigma^2 \asymp q$, according to Assumptions (E1), (E2), and (E3). Subsequently, we have $\mathbb{E}(\max_{i\in[p]}\{|\sum_{j=1}^{q} E_{t,ij}|\}) = O(\sqrt{q\log(p)})$, which in turn yields

$$\|\mathbf{E}_t\mathbf{1}_q\|_\infty = \max_{i\in[p]}\left(\left|\sum_{j=1}^{q} E_{t,ij}\right|\right) = O_P\left(\sqrt{q\log(p)}\right). \tag{B.5}$$

Now consider part (i). In this case, note that Assumption (S1) can be satisfied when $\|\cdot\|_*$ is set as the Euclidean norm $\|\cdot\|$. Then for any shift-varying function in the set $\mathcal{L}$, by (2.7), we have

$$|\widetilde{g}_\alpha(\mathbf{x}_1) - \widetilde{g}_\alpha(\mathbf{x}_2)| = \left|(\mathbf{v}^\intercal\mathbf{1}_p)^{-1}\mathbf{v}^\intercal(\mathbf{x}_1 - \mathbf{x}_2)\right| \leq \frac{\|\mathbf{v}\|}{|\mathbf{v}^\intercal\mathbf{1}_p|}\cdot\|\mathbf{x}_1 - \mathbf{x}_2\|,$$

thus indicating the Lipschitz constant as $L_\alpha = \|\mathbf{v}\|/|\mathbf{v}^\intercal\mathbf{1}_p|$. Also, by (B.1), $\|\mathbf{E}_t\mathbf{1}_q\|^2 = O_P(pq)$. Then we may apply Theorem 2 and obtain

$$\frac{1}{p}\|\widetilde{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}_t^*\|^2 = O_P\left(\frac{L_\alpha^2}{q^2}\|\mathbf{E}_t\mathbf{1}_q\|^2 + \frac{1}{q}\right) = O_P\left(\frac{p\|\mathbf{v}\|^2}{q(\mathbf{v}^\intercal\mathbf{1}_p)^2} + \frac{1}{q}\right). \tag{B.6}$$

On the other hand, Assumption (S1) can also be satisfied with $\|\cdot\|_*$ set as the L_∞ norm

$\|\cdot\|_\infty$, in which case we have $L_\alpha = \|\mathbf{v}\|_1/|\mathbf{v}^\intercal \mathbf{1}_p|$. Together with (B.5) and Theorem 2, we obtain

$$\frac{1}{p}\|\widetilde{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}_t^*\|^2 = O_P\Big(\frac{L_\alpha^2}{q^2}\,\|\mathbf{E}_t\mathbf{1}_q\|_\infty^2 + \frac{1}{q}\Big) = O_P\Big(\frac{\log(p)\|\mathbf{v}\|_1^2}{q(\mathbf{v}^\intercal\mathbf{1}_p)^2} + \frac{1}{q}\Big). \tag{B.7}$$

Combining (B.6) and (B.7), we have the desired for the row main effect estimators. The results for the column main effect estimators and the grand mean estimators follow similarly. Now when $\mathbf{v} = \mathbf{1}$, it is direct that the Lipschitz constant (for the row main effect estimators) has value $L_\alpha = \|\mathbf{v}\|/|\mathbf{v}^\intercal\mathbf{1}_p| = p^{-1/2}$, and similarly $L_\beta = q^{-1/2}$. It remains to show that the grand mean estimator in this case would not carry these two rates of $1/p$ and $1/q$. To see this, note that by (2.7),

$$\mathbf{1}_p^\intercal\widetilde{\boldsymbol{\alpha}}_t = q^{-1}\mathbf{1}_p^\intercal\mathbf{X}_t\mathbf{1}_q - \frac{p}{\mathbf{v}^\intercal\mathbf{1}_p}(q^{-1}\mathbf{v}^\intercal\mathbf{X}_t\mathbf{1}_q - c),$$

which has value c if $\mathbf{v} = \mathbf{1}_p$. Also, recall that Assumption (IC1) requires $\mathbf{1}_p^\intercal\boldsymbol{\alpha}_t^* = c$ in such case. Then from the decomposition of $(\widetilde{\mu}_t - \mu_t)^2$ in the proof of Theorem 2, it is direct that $(\widetilde{\mu}_t - \mu_t)^2 = O_P\{1/(pq)\}$. This completes the proof for part (i) of the theorem.

For part (ii), recall that

$$|\widetilde{g}_\alpha(\mathbf{x}_1) - \widetilde{g}_\alpha(\mathbf{x}_2)| = |q_u(\mathbf{x}_1) - q_u(\mathbf{x}_2)| \le \|\mathbf{x}_1 - \mathbf{x}_2\|_\infty \le \|\mathbf{x}_1 - \mathbf{x}_2\|.$$

Thus in Assumption (S1), the Lipschitz constant $L_\alpha = 1$ when $\|\cdot\|_*$ is set as either the Euclidean norm or the L_∞ norm. Since $\|\mathbf{E}_t\mathbf{1}_q\|^2$ has order larger than that of $\|\mathbf{E}_t\mathbf{1}_q\|_\infty^2$, we use the L_∞ norm and apply Theorem 2. This gives

$$\frac{1}{p}\|\widetilde{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}_t^*\|^2 = O_P\Big(\frac{L_\alpha^2}{q^2}\|\mathbf{E}_t\mathbf{1}_q\|_\infty^2 + \frac{1}{q}\Big) = O_P\Big(\frac{\log(p)}{q} + \frac{1}{q}\Big) = O_P\Big(\frac{\log(p)}{q}\Big),$$

as desired, which concludes the proof of this theorem. □

Proof of Theorem 4. First, we show the consistency of the factor loading estimators. It suffices to consider the row loading estimator, as the arguments for the column loading estimator are analogous. Recall the notation $\mathbf{M}_a = \mathbf{I}_a - a^{-1}\mathbf{1}_a\mathbf{1}_a^\intercal$, then from (2.6), we have

$$\begin{aligned}
\widetilde{\mathbf{L}}_t &= \mathbf{X}_t - (pq)^{-1}\mathbf{1}_p^\intercal\mathbf{X}_t\mathbf{1}_q\mathbf{1}_p\mathbf{1}_q^\intercal + p^{-1}\mathbf{1}_p^\intercal\widetilde{\boldsymbol{\alpha}}_t\mathbf{1}_p\mathbf{1}_q^\intercal + q^{-1}\mathbf{1}_q^\intercal\widetilde{\boldsymbol{\beta}}_t\mathbf{1}_p\mathbf{1}_q^\intercal - \widetilde{\boldsymbol{\alpha}}_t\mathbf{1}_q^\intercal - \mathbf{1}_p\widetilde{\boldsymbol{\beta}}_t^\intercal \\
&= \mathbf{M}_p\mathbf{X}_t(\mathbf{1}_q\mathbf{1}_q^\intercal/q) + \mathbf{X}_t\mathbf{M}_q - \mathbf{M}_p\widetilde{\boldsymbol{\alpha}}_t\mathbf{1}_q^\intercal - \mathbf{1}_p\widetilde{\boldsymbol{\beta}}_t^\intercal\mathbf{M}_q \\
&= \mathbf{M}_p\mathbf{X}_t(\mathbf{1}_q\mathbf{1}_q^\intercal/q) + \mathbf{X}_t\mathbf{M}_q - \mathbf{M}_p\mathbf{X}_t(\mathbf{1}_q\mathbf{1}_q^\intercal/q) - (\mathbf{1}_p\mathbf{1}_p^\intercal/p)\mathbf{X}_t\mathbf{M}_q = \mathbf{M}_p\mathbf{X}_t\mathbf{M}_q,
\end{aligned}$$

where the second last equality used the definitions of $\widetilde{\boldsymbol{\alpha}}_t$ and $\widetilde{\boldsymbol{\beta}}_t$ in (2.3) and (2.4) respectively, together with the key that $\mathbf{M}_a\mathbf{1}_a = \mathbf{0}$. Hence,

$$\begin{aligned}
\widetilde{\mathbf{L}}_t\widetilde{\mathbf{L}}_t^\intercal &= \mathbf{M}_p\mathbf{X}_t\mathbf{M}_q\mathbf{M}_q^\intercal\mathbf{X}_t^\intercal\mathbf{M}_p^\intercal = \mathbf{M}_p\mathbf{X}_t\mathbf{M}_q\mathbf{X}_t^\intercal\mathbf{M}_p \\
&= \mathbf{M}_p(\mathbf{C}_t + \mathbf{E}_t)\mathbf{M}_q(\mathbf{C}_t^\intercal + \mathbf{E}_t^\intercal)\mathbf{M}_p = \mathbf{C}_t\mathbf{C}_t^\intercal + \mathbf{C}_t\mathbf{E}_t^\intercal\mathbf{M}_p + \mathbf{M}_p\mathbf{E}_t\mathbf{C}_t^\intercal + \mathbf{M}_p\mathbf{E}_t\mathbf{M}_q\mathbf{E}_t^\intercal\mathbf{M}_p
\end{aligned}$$

$$= \mathbf{C}_t\mathbf{C}_t^\intercal + \mathbf{C}_t\mathbf{E}_t^\intercal + \mathbf{E}_t\mathbf{C}_t^\intercal + \mathbf{E}_t\mathbf{E}_t^\intercal + (pq)^{-1}\mathbf{E}_t\mathbf{1}_q\mathbf{1}_q^\intercal\mathbf{E}_t^\intercal\mathbf{1}_p\mathbf{1}_p^\intercal + (pq)^{-1}\mathbf{1}_p\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{1}_q\mathbf{1}_q^\intercal\mathbf{E}_t^\intercal$$
$$- p^{-1}\mathbf{C}_t\mathbf{E}_t^\intercal\mathbf{1}_p\mathbf{1}_p^\intercal - p^{-1}\mathbf{1}_p\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{C}_t^\intercal - p^{-1}\mathbf{E}_t\mathbf{E}_t^\intercal\mathbf{1}_p\mathbf{1}_p^\intercal - p^{-1}\mathbf{1}_p\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{E}_t^\intercal - q^{-1}\mathbf{E}_t\mathbf{1}_q\mathbf{1}_q^\intercal\mathbf{E}_t^\intercal$$
$$+ p^{-2}\mathbf{1}_p\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{E}_t^\intercal\mathbf{1}_p\mathbf{1}_p^\intercal - (pq)^{-1}p^{-1}\mathbf{1}_p\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{1}_q\mathbf{1}_q^\intercal\mathbf{E}_t^\intercal\mathbf{1}_p\mathbf{1}_p^\intercal. \tag{B.8}$$

To ease notations, define

$$\mathbf{R}_{r,t} := \mathbf{C}_t\mathbf{E}_t^\intercal + \mathbf{E}_t\mathbf{C}_t^\intercal + \mathbf{E}_t\mathbf{E}_t^\intercal + (pq)^{-1}\mathbf{E}_t\mathbf{1}_q\mathbf{1}_q^\intercal\mathbf{E}_t^\intercal\mathbf{1}_p\mathbf{1}_p^\intercal + (pq)^{-1}\mathbf{1}_p\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{1}_q\mathbf{1}_q^\intercal\mathbf{E}_t^\intercal$$
$$- p^{-1}\mathbf{C}_t\mathbf{E}_t^\intercal\mathbf{1}_p\mathbf{1}_p^\intercal - p^{-1}\mathbf{1}_p\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{C}_t^\intercal - p^{-1}\mathbf{E}_t\mathbf{E}_t^\intercal\mathbf{1}_p\mathbf{1}_p^\intercal - p^{-1}\mathbf{1}_p\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{E}_t^\intercal - q^{-1}\mathbf{E}_t\mathbf{1}_q\mathbf{1}_q^\intercal\mathbf{E}_t^\intercal$$
$$+ p^{-2}\mathbf{1}_p\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{E}_t^\intercal\mathbf{1}_p\mathbf{1}_p^\intercal - (pq)^{-1}p^{-1}\mathbf{1}_p\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{1}_q\mathbf{1}_q^\intercal\mathbf{E}_t^\intercal\mathbf{1}_p\mathbf{1}_p^\intercal,$$

so that from (B.8), we can write $\widetilde{\mathbf{L}}_t\widetilde{\mathbf{L}}_t^\intercal = \mathbf{C}_t\mathbf{C}_t^\intercal + \mathbf{R}_{r,t}$. In particular, from Lemma 2 of Lam and Cen (2025), we have

$$\left\| \sum_{t=1}^{T} \mathbf{R}_{r,t} \right\|_F^2 = O_P(Tp^2q + T^2pq^2). \tag{B.9}$$

Next, we show

$$\left\|\widehat{\mathbf{Q}}_r - \mathbf{Q}_r\mathbf{H}_r^\intercal\right\|_F^2 = O_P\Big(T^{-1}p^{2(1-\delta_{r,k_r})}q^{1-2\delta_{c,1}} + p^{1-2\delta_{r,k_r}}q^{2(1-\delta_{c,1})}\Big), \tag{B.10}$$

where $\mathbf{H}_r$ satisfies that $\left\|\mathbf{H}_r\mathbf{Q}_r^\intercal\mathbf{Q}_r\mathbf{H}_r^\intercal - \mathbf{I}_{k_r}\right\|_F^2 = o_P(1)$, which, since $\mathbf{Q}_r^\intercal\mathbf{Q}_r \to \mathbf{\Sigma}_{A,r}$ and $\mathbf{H}_r = O_P(1)$, implies $\|\mathbf{H}_r\mathbf{\Sigma}_{A,r}\mathbf{H}_r^\intercal - \mathbf{I}_{k_r}\|_F^2 = o_P(1)$ and the asymptotic invertibility of

$\mathbf{H}_r$. This would follow similar arguments in the proof of Theorem 2 in Lam and Cen (2025), which we delineate as follows for completeness. By construction, it holds that $\widehat{\mathbf{Q}}_r\widehat{\mathbf{D}}_r = T^{-1}\sum_{t=1}^{T}\widetilde{\mathbf{L}}_t\widetilde{\mathbf{L}}_t^{\intercal}\widehat{\mathbf{Q}}_r$, and hence for any $j \in [p]$:

$$\widehat{\mathbf{Q}}_{r,j\cdot} - \mathbf{H}_r\mathbf{Q}_{r,j\cdot} = \frac{1}{T}\widehat{\mathbf{D}}_r^{-1}\sum_{i=1}^{p}\widehat{\mathbf{Q}}_{r,i\cdot}\sum_{t=1}^{T}(\mathbf{R}_{r,t})_{ij}. \tag{B.11}$$

Hence by sub-multiplicativity of the Frobenius norm, we have

$$\begin{aligned}\big\|\widehat{\mathbf{Q}}_r - \mathbf{Q}_r\mathbf{H}_r^{\intercal}\big\|_F^2 &= \sum_{j=1}^{p}\big\|\widehat{\mathbf{Q}}_{r,j\cdot} - \mathbf{H}_r\mathbf{Q}_{r,j\cdot}\big\|^2 = \sum_{j=1}^{p}\bigg\|\frac{1}{T}\widehat{\mathbf{D}}_r^{-1}\widehat{\mathbf{Q}}_r^{\intercal}\Big(\sum_{t=1}^{T}\mathbf{R}_{r,t}\Big)_{\cdot j}\bigg\|^2 \\ &\le \frac{1}{T^2}\big\|\widehat{\mathbf{D}}_r^{-1}\big\|_F^2\cdot\big\|\widehat{\mathbf{Q}}_r\big\|_F^2\cdot\bigg\|\sum_{t=1}^{T}\mathbf{R}_{r,t}\bigg\|_F^2.\end{aligned}$$

Combining the above with (B.9) and that $\|\widehat{\mathbf{D}}_r^{-1}\|_F = O_P\big(p^{-\delta_{r,k_r}}q^{-\delta_{c,1}}\big)$ from Lemma 3 of Lam and Cen (2025), we may conclude (B.10). We further improve this rate of convergence in the following. To this end, note that for any $j \in [p]$, (B.11) gives

$$\widehat{\mathbf{Q}}_{r,j\cdot} - \mathbf{H}_r\mathbf{Q}_{r,j\cdot} = \frac{1}{T}\widehat{\mathbf{D}}_r^{-1}\sum_{i=1}^{p}(\widehat{\mathbf{Q}}_{r,i\cdot} - \mathbf{H}_r\mathbf{Q}_{r,i\cdot})\sum_{t=1}^{T}(\mathbf{R}_{r,t})_{ij} + \frac{1}{T}\widehat{\mathbf{D}}_r^{-1}\sum_{i=1}^{p}\mathbf{H}_r\mathbf{Q}_{r,i\cdot}\sum_{t=1}^{T}(\mathbf{R}_{r,t})_{ij}.$$

Thus similar to the above, we can write

$$\begin{aligned}&\big\|\widehat{\mathbf{Q}}_r - \mathbf{Q}_r\mathbf{H}_r^{\intercal}\big\|_F^2 = \sum_{j=1}^{p}\big\|\widehat{\mathbf{Q}}_{r,j\cdot} - \mathbf{H}_r\mathbf{Q}_{r,j\cdot}\big\|^2 \\ &\le \sum_{j=1}^{p}\bigg\|\frac{1}{T}\widehat{\mathbf{D}}_r^{-1}(\widehat{\mathbf{Q}}_r - \mathbf{Q}_r\mathbf{H}_r^{\intercal})^{\intercal}\Big(\sum_{t=1}^{T}\mathbf{R}_{r,t}\Big)_{\cdot j}\bigg\|^2 + \sum_{j=1}^{p}\bigg\|\frac{1}{T}\widehat{\mathbf{D}}_r^{-1}\mathbf{H}_r\mathbf{Q}_r^{\intercal}\Big(\sum_{t=1}^{T}\mathbf{R}_{r,t}\Big)_{\cdot j}\bigg\|^2\end{aligned}$$

$$\leq \frac{1}{T^2}\|\widehat{\mathbf{D}}_r^{-1}\|_F^2 \cdot \|\widehat{\mathbf{Q}}_r - \mathbf{Q}_r\mathbf{H}_r^\intercal\|_F^2 \cdot \Big\|\sum_{t=1}^{T}\mathbf{R}_{r,t}\Big\|_F^2 + \frac{1}{T^2}\|\widehat{\mathbf{D}}_r^{-1}\|_F^2 \cdot \|\mathbf{H}_r\|_F^2 \cdot \Big\|\sum_{t=1}^{T}\mathbf{Q}_r^\intercal\mathbf{R}_{r,t}\Big\|_F^2$$
$$= O_P\Big(T^{-1}p^{2(1-\delta_{r,k_r})}q^{1-2\delta_{c,1}} + p^{1-3\delta_{r,k_r}}q^{2-2\delta_{c,1}}\Big),$$

where the last equality used the rate on $\|\widehat{\mathbf{D}}_r^{-1}\|_F$, (B.9), (B.10), and the arguments in the proof of Theorem 5 of Lam and Cen (2025). This concludes the proof for the factor loading estimators. Finally, the proofs for the rates of the error of estimated factor series and individual common components are similar to that of Theorem 3 in Lam and Cen (2025), which completes the proof of this theorem. □

Proof of Theorem 5. It is sufficient to show the results for the row main effects, as the proof for the column main effects follows similarly. We first show the block consistency or equivalently, sign consistency of the DAFL estimators according to (4.5) in estimating the row main effects. Define $\widetilde{\boldsymbol{\alpha}}_{\cdot,i} = (\widetilde{\alpha}_{1,i}, \ldots, \widetilde{\alpha}_{T,i})^\intercal$ and $\widehat{\boldsymbol{\alpha}}_{\cdot,i} = (\widehat{\alpha}_{1,i}, \ldots, \widehat{\alpha}_{T,i})^\intercal$, for the ease of notation. By the KKT condition, $\widehat{\boldsymbol{\alpha}}_{\cdot,i}$ is a solution minimizing the loss function in (4.5) if

and only if there exists a subgradient

$$\mathbf{h} = \partial\Big(\sum_{t=2}^{T} u_{\alpha,t,i}\big|\widehat{\alpha}_{t,i} - \widehat{\alpha}_{t-1,i}\big|\Big) = \big\{\mathbf{h} \in \mathbb{R}^T \text{ such that}$$

$$h_1 : \begin{cases} h_1 = -u_{\alpha,2,i}\,\mathrm{sign}(\widehat{\alpha}_{2,i} - \widehat{\alpha}_{1,i}), & \widehat{\alpha}_{1,i} \neq \widehat{\alpha}_{2,i}; \\ |h_1| \leq u_{\alpha,2,i}, & \text{otherwise.} \end{cases};$$

$$h_T : \begin{cases} h_T = u_{\alpha,T,i}\,\mathrm{sign}(\widehat{\alpha}_{T,i} - \widehat{\alpha}_{T-1,i}), & \widehat{\alpha}_{T,i} \neq \widehat{\alpha}_{T-1,i}; \\ |h_T| \leq u_{\alpha,T,i}, & \text{otherwise.} \end{cases};$$

for $t = 2, \ldots, T-1$,

$$h_t : \left\{ \begin{array}{ll} h_t = u_{\alpha,t,i}\,\mathrm{sign}(\widehat{\alpha}_{t,i} - \widehat{\alpha}_{t-1,i}) - u_{\alpha,t+1,i}\,\mathrm{sign}(\widehat{\alpha}_{t+1,i} - \widehat{\alpha}_{t,i}), & \widehat{\alpha}_{t-1,i} \neq \widehat{\alpha}_{t,i} \neq \widehat{\alpha}_{t+1,i}; \\ \big|h_t + u_{\alpha,t+1,i}\,\mathrm{sign}(\widehat{\alpha}_{t+1,i} - \widehat{\alpha}_{t,i})\big| \leq u_{\alpha,t,i}, & \widehat{\alpha}_{t-1,i} = \widehat{\alpha}_{t,i} \neq \widehat{\alpha}_{t+1,i}; \\ \big|h_t - u_{\alpha,t,i}\,\mathrm{sign}(\widehat{\alpha}_{t,i} - \widehat{\alpha}_{t-1,i})\big| \leq u_{\alpha,t+1,i}, & \widehat{\alpha}_{t-1,i} \neq \widehat{\alpha}_{t,i} = \widehat{\alpha}_{t+1,i}; \\ \big|h_t\big| \leq u_{\alpha,t,i} + u_{\alpha,t+1,i}, & \text{otherwise.} \end{array} \right\}, \tag{B.12}$$

and also a subgradient

$$\mathbf{g} = \partial\Big(\sum_{t=1}^{T} \gamma_{\alpha,t,i}\big|\widehat{\alpha}_{t,i}\big|\Big) = \left\{ \mathbf{g} \in \mathbb{R}^T : \left\{ \begin{array}{ll} g_t = \gamma_{\alpha,t,i}\,\mathrm{sign}(\widehat{\alpha}_{t,i}), & \widehat{\alpha}_{t,i} \neq 0; \\ |g_t| \leq \gamma_{\alpha,t,i}, & \text{otherwise.} \end{array} \right\}, \right. \tag{B.13}$$

such that differentiating $L(\widehat{\boldsymbol{\alpha}}_{\cdot,i}) \equiv L(\widehat{\alpha}_{1,i}, \ldots, \widehat{\alpha}_{T,i})$ with respect to $\widehat{\boldsymbol{\alpha}}_{\cdot,i}$, we have

$$\widehat{\boldsymbol{\alpha}}_{\cdot,i} - \widetilde{\boldsymbol{\alpha}}_{\cdot,i} = -\partial\Big(\lambda_\alpha \sum_{t=2}^{T} u_{\alpha,t,i}|\widehat{\alpha}_{t,i} - \widehat{\alpha}_{t-1,i}| + \lambda_\alpha \sum_{t=1}^{T} \gamma_{\alpha,t,i}|\widehat{\alpha}_{t,i}|\Big) = -\lambda_\alpha \mathbf{h} - \lambda_\alpha \mathbf{g}. \quad \text{(B.14)}$$

Without loss of generality, we can always partition each sparse block into $\mathcal{S}_{\alpha,i} = \overline{\mathcal{S}}_{\alpha,i} \cup \mathcal{S}^{\circ}_{\alpha,i}$, where $\overline{\mathcal{S}}_{\alpha,i}$ is the periods when the main effect is piecewise zero and $\mathcal{S}^{\circ}_{\alpha,i}$ is the periods when the main effect is distinct zero. Formally, $\overline{\mathcal{S}}_{\alpha,i}$ is the largest set such that for all $t \in \overline{\mathcal{S}}_{\alpha,i}$: (i) $\alpha^*_{t,i} = 0$; (ii) $\{t-1, t+1\} \cap \overline{\mathcal{S}}_{\alpha,i} \neq \emptyset$.

Due to the intricacy in (B.12), we define the interior of $\overline{\mathcal{S}}_{\alpha,i}$ as

$$\mathcal{S}^*_{\alpha,i} = \big\{t \in \overline{\mathcal{S}}_{\alpha,i} : \text{ either } t \in \{1, T\} \text{ or } \{t-1, t+1\} \subseteq \overline{\mathcal{S}}_{\alpha,i}\big\},$$

that is, both (or only one neighbor when $t = 1, T$) neighboring timestamps are still in the sparse block.

To have sign consistency, we first show consistency in recovering the sparse block. (B.14) implies that uniformly over $i \in [p]$ and $t \in \mathcal{S}_{\alpha,i}$,

$$\widetilde{\alpha}_{t,i} = \lambda_\alpha h_t + \lambda_\alpha g_t, \quad \text{(B.15)}$$

where $|g_t| \leq \gamma_{\alpha,t,i}$ according to (B.13). Firstly consider $t \in \overline{\mathcal{S}}_{\alpha,i}$, we discuss this in three cases to ease the discussion, where Cases (a) and (c) correspond to consistency in recovering

the interior of the sparse block, Case (b) corresponds to the margin, $\overline{\mathcal{S}}_{\alpha,i} \setminus \mathcal{S}^*_{\alpha,i}$, of the sparse block:

(a) $t \in \mathcal{S}^*_{\alpha,i}$ and $t \notin \{1, T\}$, i.e., both the previous and the next timestamps are also in the sparse block;

(b) $t \in \overline{\mathcal{S}}_{\alpha,i} \setminus \mathcal{S}^*_{\alpha,i}$, recall that it means either the previous or the next timestamp is in the sparse block (but not both), i.e., $\{t-1, t\} \subseteq \mathcal{S}_{\alpha,i}$ or $\{t, t+1\} \subseteq \mathcal{S}_{\alpha,i}$; the definition also implies $t \neq 1$ and $t \neq T$, which facilitates us to better discuss the subgradient $\mathbf{h}$;

(c) $t \in \mathcal{S}^*_{\alpha,i}$ and $t \in \{1, T\}$.

Consider Case (a) first, then the subgradient $\mathbf{h}$ in (B.12) can only take values $\left|h_t\right| \leq u_{\alpha,t,i} + u_{\alpha,t+1,i}$. Hence for (B.15), we need to show that (uniformly over $i \in [p]$ and $t \in \overline{\mathcal{S}}_{\alpha,i}$),

$$|\widetilde{\alpha}_{t,i}| = \left|\lambda_\alpha h_t + \lambda_\alpha g_t\right| \leq \lambda_\alpha |h_t| + \lambda_\alpha |g_t| \leq \lambda_\alpha (u_{\alpha,t,i} + u_{\alpha,t+1,i}) + \lambda_\alpha \gamma_{\alpha,t,i}.$$

The right hand side above is lower bounded by $\lambda_\alpha / \widetilde{\alpha}_{t,i}$, according to the definition of $u_{\alpha,t,i}$ and $\gamma_{\alpha,t,i}$. Hence it suffices to show

$$\max_{i \in [p]} \max_{t \in \overline{\mathcal{S}}_{\alpha,i} : t-1 \in \overline{\mathcal{S}}_{\alpha,i}} \left\{\widetilde{\alpha}^2_{t,i}\right\} = \max_{i \in [p]} \|(\widetilde{\boldsymbol{\alpha}}_{\cdot,i})_{\mathcal{S}_{\alpha,i}}\|^2_\infty = o_P(\lambda_\alpha),$$

where the second equality indeed holds true from Assumption (R2) and the first result of Lemma B.2.

For Case (b), the subgradient value for h_t satisfies

$$\begin{aligned} \text{either} \quad & \left|h_t + u_{\alpha,t+1,i}\right| \le u_{\alpha,t,i} \quad \text{with} \quad 0 = \widehat{\alpha}_{t-1,i} = \widehat{\alpha}_{t,i} \neq \widehat{\alpha}_{t+1,i} \\ \text{or} \quad & \left|h_t + u_{\alpha,t,i}\right| \le u_{\alpha,t+1,i} \quad \text{with} \quad \widehat{\alpha}_{t-1,i} \neq \widehat{\alpha}_{t,i} = \widehat{\alpha}_{t+1,i} = 0. \end{aligned}$$

We only consider the first scenario above as the second follows a similar proof. Notice that $\widetilde{\alpha}_{t,i} \ge 0$, so with (B.15), it suffices to show that (uniformly over $i \in [p]$ and $t \in \overline{\mathcal{S}}_{\alpha,i}$ with $t+1 \in \mathcal{B}_{\alpha,i}$),

$$\begin{cases} -\lambda_\alpha u_{\alpha,t+1,i} - \lambda_\alpha u_{\alpha,t,i} - \lambda_\alpha \gamma_{\alpha,t,i} \le 0, \\ \widetilde{\alpha}_{t,i} \le -\lambda_\alpha u_{\alpha,t+1,i} + \lambda_\alpha u_{\alpha,t,i} + \lambda_\alpha \gamma_{\alpha,t,i}, \end{cases} \tag{B.16}$$

where the first inequality is immediate by noticing all $u_{\alpha,t+1,i}$, $u_{\alpha,t,i}$, and $\gamma_{\alpha,t,i}$ are non-negative. For the second inequality, since $\widetilde{\alpha}_{t,i}$ is dominated by $\lambda_\alpha u_{\alpha,t,i} + \lambda_\alpha \gamma_{\alpha,t,i}$ for $t \in \overline{\mathcal{S}}_{\alpha,i}$ using the argument in Case (a), it remains to show $u_{\alpha,t+1,i}$ is stochastically dominated by $\gamma_{\alpha,t,i}$, for which, noting that $\overline{\mathcal{S}}_{\alpha,i} \subseteq \mathcal{S}_{\alpha,i}$, it suffices to show

$$\max_{i\in[p]} \max_{t\in\mathcal{S}_{\alpha,i}: t+1\in\mathcal{B}_{\alpha,i}} \{\widetilde{\alpha}_{t,i}\} = o_P\big(\min_{i\in[p]} \min_{t\in\mathcal{S}_{\alpha,i}: t+1\in\mathcal{B}_{\alpha,i}} \{\widetilde{\alpha}_{t+1,i}\}\big). \tag{B.17}$$

In detail, we have

$$\begin{cases} \max_{i\in[p]} \max_{t\in\mathcal{S}_{\alpha,i}: t+1\in\mathcal{B}_{\alpha,i}} \{\widetilde{\alpha}_{t,i}\} \le \max_{i\in[p]} \|(\widetilde{\boldsymbol{\alpha}}_{\cdot,i})_{\mathcal{S}_{\alpha,i}}\|_\infty, \\ \min_{i\in[p]} \min_{t\in\mathcal{S}_{\alpha,i}: t+1\in\mathcal{B}_{\alpha,i}} \{\widetilde{\alpha}_{t+1,i}\} \\ \ge \min_{i\in[p]} \min_{t\in\mathcal{S}_{\alpha,i}: t+1\in\mathcal{B}_{\alpha,i}} \{\alpha^*_{t+1,i}\} - \max_{i\in[p]} \max_{t\in\mathcal{S}_{\alpha,i}: t+1\in\mathcal{B}_{\alpha,i}} |\widetilde{\alpha}_{t+1,i} - \alpha^*_{t+1,i}|, \end{cases}$$

which, together with Assumption (R2) and Lemma B.2, shows (B.17).

Lastly, consider Case (c). If $t = 1$, then we have $|h_1| \leq u_{\alpha,2,i}$; otherwise if $t = T$, then $|h_T| \leq u_{\alpha,T,i}$. Suppose $t = 1$, with (B.15), it is required that

$$|\widetilde{\alpha}_{1,i}| = \big|\lambda_\alpha h_1 + \lambda_\alpha g_1\big| \leq \lambda_\alpha |h_1| + \lambda_\alpha |g_1| \leq \lambda_\alpha u_{\alpha,2,i} + \lambda_\alpha \gamma_{\alpha,1,i}.$$

Similar to Case (a), it is sufficient to show $\max_{i\in[p]}\{\widetilde{\alpha}_{1,i}^2\} \leq \max_{i\in[p]} \|(\widetilde{\boldsymbol{\alpha}}_{\cdot,i})_{\mathcal{S}_{\alpha,i}}\|_\infty^2 = o_P(\lambda_\alpha)$, which is true by exactly the same argument in Case (a). The argument for the scenario $t = T$ is similar and hence omitted here. This completes the proof of consistency in recovering the sparse blocks.

It remains to consider $t \in \mathcal{S}_{\alpha,i}^\circ$ for consistency in the sparse block. Using (B.12), we require uniformly over $i \in [p]$ and $t \in \mathcal{S}_{\alpha,i}^\circ$ that

$$\big|\widetilde{\alpha}_{t,i} + u_{\alpha,t,i} + u_{\alpha,t+1,i}\big| = \widetilde{\alpha}_{t,i} + \lambda_\alpha u_{\alpha,t,i} + \lambda_\alpha u_{\alpha,t+1,i} \leq \lambda_\alpha \gamma_{\alpha,t,i},$$

which holds true if we can show

$$\begin{cases} \max_{i\in[p]} \|(\widetilde{\boldsymbol{\alpha}}_{\cdot,i})_{\mathcal{S}_{\alpha,i}^\circ}\|_\infty^2 \leq \max_{i\in[p]} \|(\widetilde{\boldsymbol{\alpha}}_{\cdot,i})_{\mathcal{S}_{\alpha,i}}\|_\infty^2 = o_P(\lambda_\alpha), \\ \max_{i\in[p]} \max_{t\in\mathcal{S}_{\alpha,i}^\circ : t+1\in\mathcal{B}_{\alpha,i}} \{\widetilde{\alpha}_{t,i}\} = o_P\big(\min_{i\in[p]} \min_{t\in\mathcal{S}_{\alpha,i}^\circ : t+1\in\mathcal{B}_{\alpha,i}} \{\widetilde{\alpha}_{t+1,i}\}\big). \end{cases}$$

The first equality is direct from Lemma B.2 and Assumption (R2), while the second equality from (B.17).

For sign consistency in $\mathcal{B}_{\alpha,i}$, from (B.14), we require with probability approaching 1 that

$$\mathbf{0} < (\widehat{\boldsymbol{\alpha}}_{\cdot,i})_{\mathcal{B}_{\alpha,i}} = (\widetilde{\boldsymbol{\alpha}}_{\cdot,i})_{\mathcal{B}_{\alpha,i}} - \lambda_\alpha(\mathbf{h})_{\mathcal{B}_{\alpha,i}} - \lambda_\alpha(\mathbf{g})_{\mathcal{B}_{\alpha,i}},$$

which holds true if both $\lambda_\alpha(\mathbf{h})_{\mathcal{B}_{\alpha,i}}$ and $\lambda_\alpha(\mathbf{g})_{\mathcal{B}_{\alpha,i}}$ are stochastically dominated by $(\widetilde{\boldsymbol{\alpha}}_{\cdot,i})_{\mathcal{B}_{\alpha,i}}$. By (B.12) and (B.13), it suffices to show λ_α is stochastically dominated by $\min_{i\in[p]} \min_{t\in\mathcal{B}_{\alpha,i}}\{\widetilde{\alpha}_{t,i}^2\}$. This is direct by combining Assumption (R2), Lemma B.2, and

$$\min_{i\in[p]} \min_{t\in\mathcal{B}_{\alpha,i}} \{\widetilde{\alpha}_{t,i}\} \geq \min_{i\in[p]} \min_{t\in\mathcal{B}_{\alpha,i}} \{\alpha_{t,i}^*\} - \max_{i\in[p]} \max_{t\in\mathcal{B}_{\alpha,i}} |\widetilde{\alpha}_{t,i} - \alpha_{t,i}^*|. \tag{B.18}$$

This ends the proof of block consistency. It remains to show the consistency for the DAFL estimators.

To this end, note that for any $i \in [p]$, $t \in \mathcal{B}_{\alpha,i}$,

$$\widehat{\alpha}_{t,i} - \alpha_{t,i}^* = (\widehat{\alpha}_{t,i} - \widetilde{\alpha}_{t,i}) + (\widetilde{\alpha}_{t,i} - \alpha_{t,i}^*) =: \mathcal{I}_1 + \mathcal{I}_2.$$

For $\mathcal{I}_1$, combining the KKT condition (B.14) and the subgradients from (B.12) and (B.13), we have

$$\begin{aligned} |\mathcal{I}_1| &\leq \lambda_\alpha |u_{\alpha,t,i} + u_{\alpha,t+1,i} + \gamma_{\alpha,t,i}| \leq \frac{\lambda_\alpha}{\min_{i\in[p]} \min_{t\in\mathcal{B}_{\alpha,i}}\{\widetilde{\alpha}_{t,i}\}} \\ &\leq \frac{\lambda_\alpha}{\min_{i\in[p]} \min_{t\in\mathcal{B}_{\alpha,i}}\{\alpha_{t,i}^*\} - \max_{i\in[p]} \max_{t\in\mathcal{B}_{\alpha,i}} |\widetilde{\alpha}_{t,i} - \alpha_{t,i}^*|}, \end{aligned} \tag{B.19}$$

where the last inequality used (B.18). Using the fact that $\min_{i\in[p]} \min_{t\in\mathcal{B}_{\alpha,i}}\{\alpha^*_{t,i}\} = O_P(1)$ and Assumption (R2), we have

$$\lambda_\alpha = o_P\Big(\min_{i\in[p]} \min_{t\in\mathcal{B}_{\alpha,i}}\{\alpha^{*2}_{t,i}\}\Big) = o_P\Big(\min_{i\in[p]} \min_{t\in\mathcal{B}_{\alpha,i}}\{\alpha^{*}_{t,i}\}\Big),$$

so that in (B.19), $|\mathcal{I}_1| = o_P(1)$. On the other hand, consider $\mathcal{I}_2$. For any $t \in [T]$ and $i \in \mathcal{B}_{\alpha,i}$, from the proof of Theorem 2, we can decompose

$$\begin{aligned}\widetilde{\alpha}_{t,i} - \alpha^*_{t,i} &= q^{-1}\mathbf{E}^\intercal_{t,i\cdot}\mathbf{1}_q - q^{-1}\min\{\mathbf{E}_t\mathbf{1}_q\} \\ &= q^{-1}(\mathbf{A}_{e,r})^\intercal_{i\cdot}\mathbf{F}_{e,t}\mathbf{A}^\intercal_{e,c}\mathbf{1}_q + q^{-1}\sum_{j=1}^{q}\Sigma_{\varepsilon,ij}\varepsilon_{t,ij} - q^{-1}\min\{\mathbf{E}_t\mathbf{1}_q\}.\end{aligned} \tag{B.20}$$

Since $\|\mathbf{A}_{e,r}\|_1, \|\mathbf{A}_{e,c}\|_1 = O(1)$ from Assumption (E1), we have $q^{-1}(\mathbf{A}_{e,r})^\intercal_{i\cdot}\mathbf{F}_{e,t}\mathbf{A}^\intercal_{e,c}\mathbf{1}_q = O_P(q^{-1})$. Moreover, $\mathbb{E}(q^{-1}(\mathbf{\Sigma}_\epsilon \circ \boldsymbol{\epsilon}_t)\mathbf{1}_q) = \mathbf{0}$ and $\mathrm{Var}(q^{-1}\sum_{j=1}^q \Sigma_{\epsilon,ij}\epsilon_{t,ij}) = q^{-2}\sum_{j=1}^q \Sigma^2_{\epsilon,ij} = O(q^{-1})$, implying that $|q^{-1}((\mathbf{\Sigma}_\epsilon \circ \boldsymbol{\epsilon}_t)\mathbf{1}_q)_i| = O_P(q^{-1/2})$. We also have $q^{-1}\min\{\mathbf{E}_t\mathbf{1}_q\} = O_P(q^{-1/2}\sqrt{\log(p)})$ from the proof of Theorem 3. Finally, let $\gamma^2_{\alpha,i} := \lim_{q\to\infty} q^{-1}\sum_{i=1}^q \Sigma^2_{\epsilon,ij}$, then using Theorem 1 in Ayvazyan and Ulyanov (2023), we have $q^{-1/2}\sum_{j=1}^q \Sigma_{\epsilon,ij}\epsilon_{t,ij} \xrightarrow{\mathcal{D}} \mathcal{N}(0,\gamma^2_{\alpha,i})$ and hence $|q^{-1}\sum_{j=1}^q \Sigma_{\varepsilon,ij}\varepsilon_{t,ij}| = O_P(q^{-1/2})$. We may now conclude from (B.20) that $|\mathcal{I}_2| = O_P(q^{-1/2}\sqrt{\log(p)})$. Hence finally,

$$|\widehat{\alpha}_{t,i} - \alpha^*_{t,i}| = |\mathcal{I}_1 + \mathcal{I}_2| = o_P(1),$$

which completes the proof of the theorem. □

B.2 Auxiliary results and proofs

Proof of Corollary 1. The results are direct from Theorem 4. □

Proof of Corollary 2. Result 1 is direct from Theorem 5, while result 2 is immediate by Lemma B.2, given the way we construct the final estimators. □

Proof of Proposition 1. For ease of notation, consider a time series $\{x_t\}_{t=1}^T$ with stay-in probabilities $\pi^{\mathcal{S}}$ and $\pi^{\mathcal{B}}$ and an initial probability $p_1 = \mathbb{P}(x_1 = 0)$. Let $p_t := \mathbb{P}(x_t = 0) = \mathbb{E}(\mathbb{1}\{x_t = 0\})$. Then for any $t = 2, \ldots, T$, we have

$$p_{t+1} = p_t\pi^{\mathcal{S}} + (1 - p_t)(1 - \pi^{\mathcal{B}}) = (1 - \pi^{\mathcal{B}}) + (\pi^{\mathcal{S}} + \pi^{\mathcal{B}} - 1)p_t,$$

which can be solved recursively as

$$p_t = p_* + (p_1 - p_*)(\pi^{\mathcal{S}} + \pi^{\mathcal{B}} - 1)^{t-1},$$

where $p_* = (1 - \pi^{\mathcal{B}})/(2 - \pi^{\mathcal{S}} - \pi^{\mathcal{B}})$. We hence choose p_1 to be the same as p_*, and $p_t = p_*$ is satisfied. Furthermore, we shall compute the expected number of zeros over the whole series, showcased by

$$\mathbb{E}(\#\{t : x_t = 0\}) = \sum_{t=1}^{T} p_t = Tp_*.$$

Consequently, we can also compute the expected length of each sparse sub-block $\{t_\ell + 1, \ldots, t_\ell + m_\ell\}$. Every run of consecutive zeros is a geometric string, i.e., for $k = 1, 2, \ldots,$

$\mathbb{P}(\text{block length} = k) = (\pi^{\mathcal{S}})^{k-1}(1-\pi^{\mathcal{S}})$. Hence the unconditional expected length of a sparse sub-block is

$$\mathbb{E}(\text{length of a sparse sub-block}) = \sum_{k=1}^{\infty} k(1-\pi^{\mathcal{S}})(\pi^{\mathcal{S}})^{k-1} = \frac{1}{1-\pi^{\mathcal{S}}}.$$

This completes the proof of Proposition 1. □

Our setting of factor structure is similar to that in Cen and Lam (2025) with tensor order $K = 2$ therein. For convenience, we state their Proposition 1 (in their supplement) as the following lemma.

Lemma B.1. *Let Assumptions (F1), (E1) and (E2) hold. Then*

(i) (Weak correlation of noise $\mathbf{E}_t$ across different rows, columns and times). We only require Assumptions (E1) and (E2) in this part. There exists some positive constant $C < \infty$ so that for any $t \in [T], i, j \in [p], h \in [q]$,

$$\sum_{k=1}^{p}\sum_{l=1}^{q}\left|\mathbb{E}[E_{t,ih}E_{t,kl}]\right| \le C,$$
$$\sum_{l=1}^{q}\sum_{s=1}^{T}\left|\operatorname{cov}(E_{t,ih}E_{t,jh}, E_{s,il}E_{s,jl})\right| \le C.$$

(ii) (Weak dependence between factor $\mathbf{F}_t$ and noise $\mathbf{E}_t$). There exists some positive constant $C < \infty$ so that for any $j \in [p], i \in [q]$, and any deterministic vectors $\mathbf{u} \in \mathbb{R}^{k_r}$

and $\mathbf{v} \in \mathbb{R}^{k_c}$ *with constant magnitudes,*

$$\mathbb{E}\left(\frac{1}{(qT)^{1/2}}\sum_{h=1}^{q}\sum_{t=1}^{T}E_{t,jh}\mathbf{u}^{\intercal}\mathbf{F}_t\mathbf{v}\right)^2 \leq C, \quad \mathbb{E}\left(\frac{1}{(pT)^{1/2}}\sum_{h=1}^{p}\sum_{t=1}^{T}E_{t,hi}\mathbf{v}^{\intercal}\mathbf{F}_t^{\intercal}\mathbf{u}\right)^2 \leq C.$$

(iii) (Further results on factor $\mathbf{F}_t$*). We only require Assumption (F1) in this part. For any* $t \in [T]$*, all elements in* $\mathbf{F}_t$ *are independent of each other, with mean* 0 *and unit variance. Moreover,*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{F}_t\mathbf{F}_t^{\intercal} \xrightarrow{p} \mathbf{\Sigma}_r := k_c\mathbf{I}_{k_r}, \quad \frac{1}{T}\sum_{t=1}^{T}\mathbf{F}_t^{\intercal}\mathbf{F}_t \xrightarrow{p} \mathbf{\Sigma}_c := k_r\mathbf{I}_{k_c},$$

with the number of factors k_r *and* k_c *fixed as* $\min\{T, p, q\} \to \infty$*.*

Lemma B.2. *Let Assumptions (IC1) (with parts (ii)–(iii) as in (4.1)), (IC2), (E1), (E2), and (E3) hold. For each* $i \in [p]$*, define the notation* $\boldsymbol{\alpha}_{\cdot,i}^* = (\alpha_{1,i}^*, \ldots, \alpha_{T,i}^*)^{\intercal}$ *and* $\widetilde{\boldsymbol{\alpha}}_{\cdot,i} = (\widetilde{\alpha}_{1,i}, \ldots, \widetilde{\alpha}_{T,i})^{\intercal}$*, then we have*

$$\max_{i\in[p]}\left\|(\widetilde{\boldsymbol{\alpha}}_{\cdot,i})_{\mathcal{S}_{\alpha,i}}\right\|_{\infty} = O_P\Big\{q^{-1/2}\log^{1/2}\Big(p\sum_{i=1}^{p}|\mathcal{S}_{\alpha,i}|\Big)\Big\},$$

where $(\widetilde{\boldsymbol{\alpha}}_{\cdot,i})_{\mathcal{S}_{\alpha,i}}$ *denotes the vector of* $\widetilde{\boldsymbol{\alpha}}_{\cdot,i}$ *with indices restricted on* $\mathcal{S}_{\alpha,i}$*, i.e., the vector consisting of* $\{\widetilde{\alpha}_{t,i}\}_{t\in\mathcal{S}_{\alpha,i}}$*. Let* $(\widetilde{\boldsymbol{\alpha}}_{\cdot,i} - \boldsymbol{\alpha}_{\cdot,i}^*)_{\mathcal{B}_{\alpha,i}}$ *be similarly defined by restricting indices on*

the set $\mathcal{B}_{\alpha,i}$. Then we have

$$\max_{i\in[p]} \big\|(\widetilde{\boldsymbol{\alpha}}_{\cdot,i} - \boldsymbol{\alpha}^*_{\cdot,i})_{\mathcal{B}_{\alpha,i}}\big\|_\infty = O_P\Big\{q^{-1/2}\log^{1/2}\Big(p\sum_{i=1}^{p}|\mathcal{B}_{\alpha,i}|\Big)\Big\}.$$

Proof of Lemma B.2. To show the first result, from (1.1) and Assumption (IC1), we have for any $i \in [p]$,

$$(\mathbf{X}_t\mathbf{1}_q)_i = \big(\mathbf{1}_p(q\mu_t + \mathbf{1}_q^\intercal\boldsymbol{\beta}_t^*) + q\boldsymbol{\alpha}_t^* + \mathbf{E}_t\mathbf{1}_q\big)_i = q\mu_t + \mathbf{1}_q^\intercal\boldsymbol{\beta}_t^* + q\alpha_{t,i}^* + \mathbf{1}_q^\intercal\mathbf{E}_{t,i\cdot},$$

so that using (4.2), it holds for any $t \in [T]$,

$$\begin{aligned}
&(\widetilde{\boldsymbol{\alpha}}_{\cdot,i})_t \equiv \widetilde{\alpha}_{t,i} = q^{-1}(\mathbf{X}_t\mathbf{1}_q)_i - q^{-1}\min\{\mathbf{X}_t\mathbf{1}_q\} \\
&= \mu_t + q^{-1}\mathbf{1}_q^\intercal\boldsymbol{\beta}_t^* + \alpha_{t,i}^* + q^{-1}\mathbf{1}_q^\intercal\mathbf{E}_{t,i\cdot} - \min_{j\in[p]}\big\{\mu_t + q^{-1}\mathbf{1}_q^\intercal\boldsymbol{\beta}_t^* + \alpha_{t,j}^* + q^{-1}\mathbf{1}_q^\intercal\mathbf{E}_{t,j\cdot}\big\} \qquad \text{(B.21)} \\
&= \alpha_{t,i}^* + q^{-1}\mathbf{1}_q^\intercal\mathbf{E}_{t,i\cdot} - \min_{j\in[p]}\big\{\alpha_{t,j}^* + q^{-1}\mathbf{1}_q^\intercal\mathbf{E}_{t,j\cdot}\big\}.
\end{aligned}$$

Then using the min-zero identification condition in the main effects, and with (B.21), we have

$$\begin{aligned}
&\big\|(\widetilde{\alpha}_{1,i} - \alpha_{1,i}^*, \ldots, \widetilde{\alpha}_{T,i} - \alpha_{T,i}^*)_{\mathcal{S}_{\alpha,i}}^\intercal\big\|_\infty = \max_{t\in\mathcal{S}_{\alpha,i}}\big|q^{-1}\mathbf{1}_q^\intercal\mathbf{E}_{t,i\cdot} - \min_{j\in[p]}\big\{\alpha_{t,j}^* + q^{-1}\mathbf{1}_q^\intercal\mathbf{E}_{t,j\cdot}\big\}\big| \\
&\le \max_{t\in\mathcal{S}_{\alpha,i}}\big|q^{-1}\mathbf{1}_q^\intercal\mathbf{E}_{t,i\cdot}\big| + \max_{t\in\mathcal{S}_{\alpha,i}}\max_{j\in[p]}\big|q^{-1}\mathbf{1}_q^\intercal\mathbf{E}_{t,j\cdot}\big| + 0 \le 2\max_{t\in\mathcal{S}_{\alpha,i}}\max_{j\in[p]}\big|q^{-1}\mathbf{1}_q^\intercal\mathbf{E}_{t,j\cdot}\big|.
\end{aligned} \qquad \text{(B.22)}$$

For any $j \in [p]$, $t \in [T]$, by Assumption (E1) and (E2),

$$E_{t,jh} = \mathbf{A}_{e,r,j\cdot}^{\intercal}\Big(\sum_{w\geq 0} a_{e,w}\mathbf{X}_{e,t-w}\Big)\mathbf{A}_{e,c,h\cdot} + \Sigma_{\epsilon,jh}\Big(\sum_{g\geq 0} a_{\epsilon,g}X_{\epsilon,t-g,jh}\Big),$$

so that we have

$$\mathbf{1}_q^{\intercal}\mathbf{E}_{t,j\cdot} = \sum_{h=1}^{q} E_{t,jh} = \mathbf{A}_{e,r,j\cdot}^{\intercal}\Big(\sum_{w\geq 0} a_{e,w}\mathbf{X}_{e,t-w}\Big)\sum_{h=1}^{q}\mathbf{A}_{e,c,h\cdot} + \sum_{h=1}^{q}\Sigma_{\epsilon,jh}\Big(\sum_{g\geq 0} a_{\epsilon,g}X_{\epsilon,t-g,jh}\Big).$$

By the sparsity of $\mathbf{A}_{e,c}$ from Assumption (E1), and Assumption (E2) and (E3), we conclude that the first term above is a zero-mean sub-Gaussian random variable with variance proxy of constant order. Similarly, the second term above is also zero-mean sub-Gaussian except that the variance proxy is of order q, and independent of the first term. Thus, $q^{-1}\mathbf{1}_q^{\intercal}\mathbf{E}_{t,j\cdot} \sim \text{subG}(C/q)$ for some arbitrary constant $C > 0$, and hence for any $\varepsilon > 0$ we have

$$\mathbb{P}\Big(\max_{i\in[p]}\max_{t\in\mathcal{S}_{\alpha,i}}\max_{j\in[p]}\big|q^{-1}\mathbf{1}_q^{\intercal}\mathbf{E}_{t,j\cdot}\big| > \varepsilon\Big) \leq 2\exp\Big\{\log\Big(p\sum_{i=1}^{p}|\mathcal{S}_{\alpha,i}|\Big) - q\varepsilon^2/2C\Big\},$$

implying that

$$\max_{i\in[p]}\max_{t\in\mathcal{S}_{\alpha,i}}\max_{j\in[p]}\big|q^{-1}\mathbf{1}_q^{\intercal}\mathbf{E}_{t,j\cdot}\big| = O_P\big\{q^{-1/2}\log^{1/2}\big(p\sum_{i=1}^{p}|\mathcal{S}_{\alpha,i}|\big)\big\}.$$

Together with the equation (B.22) and the fact that $\alpha_{t,i}^{*} = 0$ for any $i \in [p]$, $t \in \mathcal{S}_{\alpha,i}$, we conclude the first result of Lemma B.2. The remaining result of the lemma can be shown

by repeating all previous arguments, except that $\alpha^*_{t,i}$ is non-zero in (B.21). This completes the proof of Lemma B.2. □

Lemma B.3. *Under Assumptions (IC1) and (IC2), both loading estimators* $\widehat{\mathbf{Q}}_r$ *and* $\widehat{\mathbf{Q}}_c$ *satisfy Assumption (IC1) (i) such that* $\mathbf{A}_r^\intercal \mathbf{1}_p = \mathbf{0}$ *and* $\mathbf{A}_c^\intercal \mathbf{1}_q = \mathbf{0}$.

Proof of Lemma B.3. From the proof of Theorem 4, we have

$$\widetilde{\mathbf{L}}_t \widetilde{\mathbf{L}}_t^\intercal = \mathbf{M}_p \mathbf{X}_t \mathbf{M}_q \mathbf{X}_t^\intercal \mathbf{M}_p.$$

Together with $\mathbf{M}_a \mathbf{1}_a = \mathbf{0}$ for any positive integer a, it is direct that the any eigenvector of $\widetilde{\mathbf{L}}_t \widetilde{\mathbf{L}}_t^\intercal$ is orthogonal to $\mathbf{1}_p$. This shows $\mathbf{A}_r^\intercal \mathbf{1}_p = \mathbf{0}$. Similar arguments hold for $\widetilde{\mathbf{L}}_t^\intercal \widetilde{\mathbf{L}}_t$ and thus $\mathbf{A}_c^\intercal \mathbf{1}_q = \mathbf{0}$. This completes the proof. □