Sparsity of the Main Effect Matrix Factor Model

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

School of Mathematics, University of Bristol

Department of Statistics, London School of Economics and Political Science

Abstract

We introduce sparsity detection and estimation in main effect matrix factor models for matrix-valued time series. A carefully chosen set of identification conditions for the common component and the potentially nonstationary main effects is proposed to strengthen the interpretations of sparse main effects, while estimators of all model components are presented. Sparse estimation of the latent main effects is proposed using a doubly adaptive fused lasso estimation to allow for sparse sub-block detection, with theoretical guarantees and rates of convergence spelt out for the final estimators. Sparse block consistency for the main effects is also proved as a result. A realized Mallow's C_p is developed for tuning parameter selection, with practical implementation described. Simulation experiments are performed under a variety of settings, showing our proposed estimators work well. A set of NYC taxi traffic data is analyzed, clearly showing the effects of Covid-19 lockdown, with prolonged sparse main effects detected.

Key words and phrases: Generalized Lasso; Sparse main effects; Tucker decomposition factor model; Uniform consistency; Weak factors.

[*] Zetai Cen is Senior Research Associate, School of Mathematics, University of Bristol. Email: zetai.cen@bristol.ac.uk

[†] Kaixin Liu is PhD student, Department of Statistics, London School of Economics. Email: K.Liu31@lse.ac.uk

[‡] Clifford Lam is Professor, 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.

1.1 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 $\mathbf{x}_t^s \in \mathbb{R}^{p_s}$ $(t = 1, \dots, T,\ s = 1, \dots, S)$ 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

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

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 $\mathbf{X}_t$, $t \in \{1, \ldots, T\} =: [T]$, 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 (2024), which generalizes the Tucker-decomposition factor model to a Main Effect Factor Model (MEFM), significantly improving model interpretability. In particular, given a matrix observation $\mathbf{X}_t \in \mathbb{R}^{p \times q}$ for $t = 1, \ldots, T$, MEFM decomposes each $\mathbf{X}_t$ as

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

where $\mathbf{A}_r \mathbf{F}_t \mathbf{A}_c^\intercal$ is akin to the common component in traditional matrix factor models with $\mathbf{F}_t \in \mathbb{R}^{k_r \times k_c}$, and μ_t, $\boldsymbol{\alpha}_t^*$, $\boldsymbol{\beta}_t^*$ are the time-varying grand mean, row and column effects respectively. 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 (2024) 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 $\boldsymbol{\alpha}_t^*$ (resp. $\boldsymbol{\beta}_t^*$) towards the observed $\mathbf{X}_t$ is the same. Hence row or column structures are solely featured by the main effects, and the interaction between rows and columns by the common component.

1.2 Main Contributions

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.

However, the identification condition (1.2) gives rise to complications since all entries in the main effects are jointly constrained by (1.2). 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. Therefore, it is crucial to develop a new, reasonable identification that allows us to consider sparsity in the main effects. Our method takes advantage of the data format and we develop different sets of identification conditions, significantly enhancing the interpretability of our model.

1.3 Notations and Organizations

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 $b = O(a)$; and $a \vee b := \max\{a, b\}$. A random variable X is sub-Gaussian with variance proxy σ^2, denoted as $X \sim \text{subG}(\sigma^2)$, if $\mathbb{E}[\exp\{s(X - \mathbb{E}[X])\}] \leq \exp(s^2\lambda^2/2)$ for all $s \in \mathbb{R}$. A random variable X is sub-exponential with parameter λ, denoted as $X \sim \text{subE}(\lambda)$, 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 $\mathbf{M}_a = \mathbf{I}_a - a^{-1}\mathbf{1}_a\mathbf{1}_a^\intercal$. The i-th largest eigenvalue of a matrix $\mathbf{A}$ is denoted by $\lambda_i(\mathbf{A})$; $\mathbf{A}^\intercal$ denotes the transpose of $\mathbf{A}$; and $\text{diag}(\{\mathbf{A}_1, \ldots, \mathbf{A}_n\})$ represents the block diagonal matrix with matrices $\{\mathbf{A}_1, \ldots, \mathbf{A}_n\}$ on the diagonal. For a given set, we denote by $|\cdot|$ its cardinality. We use $\|\cdot\|$ to denote the spectral norm of a matrix or the L_2 norm of a vector, and $\|\cdot\|_F$ to denote the Frobenius norm of a matrix. We use $\|\cdot\|_{\max}$ to denote the maximum absolute value of the elements in a vector or a matrix. The notations $\|\cdot\|_1$ and $\|\cdot\|_\infty$ denote the L_1 and L_∞-norm of a matrix respectively, defined by $\|\mathbf{A}\|_1 := \max_j \sum_i |(\mathbf{A})_{ij}|$ and $\|\mathbf{A}\|_\infty := \max_i \sum_j |(\mathbf{A})_{ij}|$. Without loss of generality, we always assume the eigenvalues of a matrix are arranged by descending orders, and so are their corresponding eigenvectors.

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 discusses the implementation of the algorithm and hyperparameter tuning based on a modified C_p statistic. Lastly, numerical results are shown in Section 5, where a real dataset on taxi traffic is also analyzed. Conclusion of the paper is in Section 6. All theoretical proofs and additional information are relegated to the supplement.

2 Model and Estimation

2.1 A matrix factor model with sparse time-varying main effects

Based on (1.1), we propose to study sparsity in MEFM under a new, simple identification condition which naturally allows for sparse main effects. Consider mean-zero matrix-valued observations $\mathbf{X}_t \in \mathbb{R}^{p\times q}$ for $t \in [T]$, each admits an MEFM representation such that

$$\mathbf{X}_t = \mu_t \mathbf{1}_p \mathbf{1}_q^\intercal + \boldsymbol{\alpha}_t^* \mathbf{1}_q^\intercal + \mathbf{1}_p \boldsymbol{\beta}_t^{*\intercal} + \mathbf{C}_t + \mathbf{E}_t, \tag{2.1}$$

where μ_t is a scalar coined as the *base effect*, $\boldsymbol{\alpha}_t^* \in \mathbb{R}^p$ and $\boldsymbol{\beta}_t^* \in \mathbb{R}^q$ are respectively the row and column main effects with potentially many zero entries (will be further explained later), $\mathbf{C}_t = \mathbf{A}_r \mathbf{F}_t \mathbf{A}_c^\intercal$ 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. More importantly, we impose the following for identification:

(IC1) (Identification) *For any* $t \in [T]$, *we assume* $\mathbf{A}_r^\intercal \mathbf{1}_p = \mathbf{0}$, $\mathbf{A}_c^\intercal \mathbf{1}_q = \mathbf{0}$, $\min\{\boldsymbol{\alpha}_t^*\} = 0$ *and* $\min\{\boldsymbol{\beta}_t^*\} = 0$.

Unlike (1.2), (IC1) identifies each main effect on one or some indices where the main effects vanish, and effectively handles the example of economic indicators in Section 1. The interpretation of main effects is similar to those in Lam and Cen (2024) in the sense that, 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. Such an interpretation holds similarly for the column effects. 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. Due to (IC1), the main effects should be read sign-less, i.e., as a magnitude. Another subtlety lies in μ_t which is called grand mean by Lam and Cen (2024) where $\mu_t = (pq)^{-1}\mathbf{1}_p^\intercal(\mathbf{X}_t - \mathbf{E}_t)\mathbf{1}_q$, but is analogous to a baseline level (and hence termed as the base effect) using our (IC1) since $\mu_t + p^{-1}\mathbf{1}_p^\intercal \boldsymbol{\alpha}_t^* + q^{-1}\mathbf{1}_q^\intercal \boldsymbol{\beta}_t^* = (pq)^{-1}\mathbf{1}_p^\intercal(\mathbf{X}_t - \mathbf{E}_t)\mathbf{1}_q$ with all main effect entries non-negative.

With (1.2), Lam and Cen (2024) first identifies μ_t, so that the row and column main effects can be subsequently identified. This is impossible under our identification in Condition (IC1), so we take another route by identifying the main effects first. This again indicates that (IC1), in spite of its simple form, is a non-trivial extension of the identification (1.2). In fact, (IC1) is a special case among a large class of valid identification conditions to MEFM; see Remark 2 in Section 3.2 for more details. Besides being identifiable with (IC1), it is unsurprising that model (2.1) is more general than the Tucker-decomposition matrix factor model, just as in Lam and Cen (2024). We present these results in the following theorem and the proof is included in the supplement.

Theorem 1 *Under Assumption (IC1), it holds for* (2.1) *that: (i) each* μ_t*,* $\boldsymbol{\alpha}_t^*$*,* $\boldsymbol{\beta}_t^*$ *and* $\mathbf{C}_t$ *can be identified; (ii) if* $\mathbf{X}_t$ *follows the Tucker-decomposition matrix factor model* $\mathbf{X}_t = \acute{\mathbf{A}}_r \mathbf{F}_t \acute{\mathbf{A}}_c^\intercal + \mathbf{E}_t$ *with full rank loadings* $\acute{\mathbf{A}}_r$ *and* $\acute{\mathbf{A}}_c$*, then* $\mathbf{X}_t$ *also follows* (2.1) *with the resulting parameters satisfying (IC1).*

To facilitate interpretation of the main effects, we consider that, for each cross-sectional unit, its main effects are 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 period.

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 2.2. 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.

2.2 Regularized estimation

We discuss the estimation of model parameters in (2.1) in this subsection. First, we may utilize Condition (IC1) to subsequently estimate the factor structure in $\mathbf{X}_t$ for each $t \in [T]$. By left-multiplying $\mathbf{1}_p^\intercal$ and right-multiplying $\mathbf{1}_q$ on $\mathbf{X}_t$, we obtain

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

according to (2.1). We can also right-multiply $\mathbf{1}_q$ or left-multiply $\mathbf{1}_p^\intercal$ on $\mathbf{X}_t$ to respectively obtain

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

Thus, together with Condition (IC1), we may obtain the initial estimators for the main effects and hence the base effect as

$$\widetilde{\boldsymbol{\alpha}}_t := q^{-1}\mathbf{X}_t \mathbf{1}_q - q^{-1}\mathbf{1}_p \min\{\mathbf{X}_t \mathbf{1}_q\}, \tag{2.2}$$

$$\widetilde{\boldsymbol{\beta}}_t := p^{-1}\mathbf{X}_t^\intercal \mathbf{1}_p - p^{-1}\mathbf{1}_q \min\{\mathbf{X}_t^\intercal \mathbf{1}_p\}, \tag{2.3}$$

$$\widetilde{\mu}_t := (pq)^{-1}\mathbf{1}_p^\intercal \mathbf{X}_t \mathbf{1}_q - p^{-1}\mathbf{1}_p^\intercal \widetilde{\boldsymbol{\alpha}}_t - q^{-1}\mathbf{1}_q^\intercal \widetilde{\boldsymbol{\beta}}_t. \tag{2.4}$$

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; Cen and Lam, 2025), 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

$$\widetilde{\mathbf{L}}_t := \mathbf{X}_t - \widetilde{\mu}_t \mathbf{1}_p \mathbf{1}_q^\intercal - \widetilde{\boldsymbol{\alpha}}_t \mathbf{1}_q^\intercal - \mathbf{1}_p \widetilde{\boldsymbol{\beta}}_t^\intercal. \tag{2.5}$$

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

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

Next, our goal is to recover the sparse blocks for the row and column main effects. 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 $i \in [p]$,

$$L^{\circ}(\alpha_{1,i}, \ldots, \alpha_{T,i}) := \frac{1}{2} \sum_{t=1}^{T} \left(\alpha_{t,i}^{*} - \alpha_{t,i}\right)^2 + \lambda_\alpha \sum_{t=1}^{T} \gamma_{\alpha,t,i} \left|\alpha_{t,i}\right|, \tag{2.6}$$

where $\gamma_{\alpha,t,i} = 1/\widetilde{\alpha}_{t,i}$ and λ_α is the tuning parameter. The estimators obtained by minimizing (2.6) 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 (2.6) are unavailable. Secondly, even if the true sparse block contains consecutive indices, the resulted main effect estimators from (2.6) 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 3.1. 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

$$L(\alpha_{1,i}, \ldots, \alpha_{T,i}) := \frac{1}{2} \sum_{t=1}^{T} \left(\widetilde{\alpha}_{t,i} - \alpha_{t,i}\right)^2 + \lambda_\alpha \sum_{t=2}^{T} u_{\alpha,t,i} \left|\alpha_{t,i} - \alpha_{t-1,i}\right| + \lambda_\alpha \sum_{t=1}^{T} \gamma_{\alpha,t,i} \left|\alpha_{t,i}\right|, \tag{2.7}$$

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 (2.6). 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 $i \in [p]$, $j \in [q]$,

$$\begin{aligned} &\widehat{\mathcal{S}}_{\alpha,i} := \{t : \widehat{\alpha}_{t,i} \le 0\}, \quad \widehat{\mathcal{S}}_{\beta,j} := \{t : \widehat{\beta}_{t,j} \le 0\}, \\ &\widehat{\mathcal{B}}_{\alpha,i} := [T] \setminus \widehat{\mathcal{S}}_{\alpha,i}, \quad \widehat{\mathcal{B}}_{\beta,j} := [T] \setminus \widehat{\mathcal{S}}_{\beta,j}. \end{aligned} \tag{2.8}$$

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 (2.8) 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 (2.7), which already induces an estimation error. Thus, even though the DAFL estimator $\{\widehat{\boldsymbol{\alpha}}_t\}_{t\in[T]}$ can consistently 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, $\{\widetilde{\boldsymbol{\alpha}}^{\mathcal{B}}_t\}_{t\in[T]}$ 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.

Remark 1 *Theorem 4 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 2.1, unless more restrictive conditions hold. A more practical concern in directly using the DAFL estimators is related to Condition (IC1). As explained below* (2.8)*, the DAFL estimators might be negative in practice and would not fulfill Condition (IC1) requiring that all main effects are at least zero. On the other hand, it is easy to see that the initial estimators from* (2.2) *and* (2.3) *satisfy (IC1). This property is retained for the final estimators due to the way they are constructed.*

Moreover, we could also update the base effect estimator as in (2.4)*, with $\widetilde{\boldsymbol{\alpha}}_t$ and $\widetilde{\boldsymbol{\beta}}_t$ replaced by $\widetilde{\boldsymbol{\alpha}}^{\mathcal{B}}_t$ and $\widetilde{\boldsymbol{\beta}}^{\mathcal{B}}_t$, respectively. However, this is unnecessary since each $\widetilde{\mu}_t$ is already a consistent estimator for the base effect at time t, while block consistency relies on further assumptions. This line of analysis is therefore not pursued in this paper.*

3 Assumptions and Theoretical Results

3.1 Assumptions

On top of the identification condition (IC1), we present below a set of assumptions to characterize the model in (2.1). The explanations to each assumption is 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* $(\mathbf{Z}_c)_{jj} \asymp p^{\delta_{c,j}}$ *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 (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,*

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

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 is 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 (2.1). 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 a 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) and hence this issue is circumvented. Lastly, Assumption (R1) spells out the rates required on the dimensions and factor strengths, which directly hold if all factors are pervasive.

3.2 Theoretical results

We discuss the main theoretical results in this subsection. We first present in Theorem 2 the consistency results for the initial estimators of the base effect and main effects. To show the consistency of the factor loading matrix estimators, we define the following square matrices

$$\begin{aligned}
\mathbf{H}_r &:= T^{-1}\widehat{\mathbf{D}}_r^{-1}\widehat{\mathbf{Q}}_r^\intercal\mathbf{Q}_r\sum_{t=1}^T(\mathbf{F}_{Z,t}\mathbf{Q}_c^\intercal\mathbf{Q}_c\mathbf{F}_{Z,t}^\intercal),\\
\mathbf{H}_c &:= T^{-1}\widehat{\mathbf{D}}_c^{-1}\widehat{\mathbf{Q}}_c^\intercal\mathbf{Q}_c\sum_{t=1}^T(\mathbf{F}_{Z,t}^\intercal\mathbf{Q}_r^\intercal\mathbf{Q}_r\mathbf{F}_{Z,t}),
\end{aligned} \tag{3.3}$$

where $\widehat{\mathbf{D}}_r := \widehat{\mathbf{Q}}_r^\intercal(T^{-1}\sum_{t=1}^T\widehat{\mathbf{L}}_t\widehat{\mathbf{L}}_t^\intercal)\widehat{\mathbf{Q}}_r$ is the $k_r \times k_r$ diagonal matrix of eigenvalues of $T^{-1}\sum_{t=1}^T\widehat{\mathbf{L}}_t\widehat{\mathbf{L}}_t^\intercal$. Similarly, $\widehat{\mathbf{D}}_c := \widehat{\mathbf{Q}}_c^\intercal(T^{-1}\sum_{t=1}^T\widehat{\mathbf{L}}_t^\intercal\widehat{\mathbf{L}}_t)\widehat{\mathbf{Q}}_c$ is the $k_c \times k_c$ diagonal matrix of eigenvalues of $T^{-1}\sum_{t=1}^T\widehat{\mathbf{L}}_t^\intercal\widehat{\mathbf{L}}_t$. The matrices $\mathbf{H}_r$ and $\mathbf{H}_c$ are shown to be asymptotically invertible in the proof of Theorem 2.

Theorem 2 *Under Assumptions (IC1), (F1), (L1), (E1), (E2), (E3) and (R1), we have the following.*

1. *The initial estimators for the main effects and the base effect are consistent such that*

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

2. *With the matrices $\mathbf{H}_r$ and $\mathbf{H}_c$ defined in* (3.3)*, both the row and column factor loading matrix estimators are consistent such that*

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

3. *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*

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

From Theorem 2, the rates of convergence for the main effects are worse off by logarithmic factors compared to the results in Theorem 2 in Lam and Cen (2024), which is understandable since the extrema of the noise series are inevitably involved due to Condition (IC1). Those rates are comparable to the results in Lam and Cen (2024) when p and q are of similar polynomial orders, which is satisfied for most data sets for factor modeling. As mentioned in Section 2.1, the base effect is identified on top of the main effects, so it is expected that the base effect estimator absorbs the asymptotic rates from the estimators of the row and column main effects.

On the other hand, as a non-trivial result, our identification condition (IC1) allows us to consider the common component separately from the base effect and main effects. Hence in Theorem 2, the rates for the loadings, core factors, and common component estimators are exactly the same as those in Lam and Cen (2024). We may also prove asymptotic normality for the factor loading matrix estimators and estimate the covariance matrices as in Lam and Cen (2024) in a trivial sense. For a detailed discussion, we refer to Remark 3. Furthermore, to provide a better reading experience, we directly show the consistency results related to the common components when all factors are strong, summarized in the corollary below.

Corollary 3 *(Simplified Theorem 2.2 and Theorem 2.3 under pervasive factors). Let all assumptions in Theorem 2 hold, and further assume that $\delta_{r,i} = \delta_{c,j} = 1$ for any $i \in [k_r]$, $j \in [k_c]$. Define the renormalized row and column loading estimators and core factor estimator as*

$$\widehat{\mathbf{A}}_r := \sqrt{p}\,\widehat{\mathbf{Q}}_r, \quad \widehat{\mathbf{A}}_r := \sqrt{q}\,\widehat{\mathbf{Q}}_c, \quad \widehat{\mathbf{F}}_t := \widehat{\mathbf{F}}_{Z,t}/\sqrt{pq}.$$

Then we have the following for any $t \in [T]$, $i \in [p]$, $j \in [q]$:

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

From Corollary 3, our renormalized loading estimators have the same performance as the α-PCA estimators considered by Chen and Fan (2023) in their Theorem 1. It is worth pointing out that this rate of $1/Tq + 1/p$ (resp. $1/Tp + 1/q$) for the row (resp. column) renormalized loading estimator can be improved to $1/Tq + 1/p^2$ (resp. $1/Tp + 1/q^2$), which will need a rate from asymptotic normality of the factor loading matrix estimators; see Lemma 5 in Cen and Lam (2025). Hence, together with the results for the core factor estimator in Corollary 3, our results align with Theorem 4.1 in He et al. (2024). Lastly, our results on the common component estimator are consistent to Theorem 4 in Chen and Fan (2023).

Before we show the properties of the DAFL estimators, we require some 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 $\min_{i\in[p]} \min_{t\in\mathcal{B}_{\alpha,i}} \{\alpha_{t,i}^*\}$, $\min_{j\in[q]} \min_{t\in\mathcal{B}_{\beta,j}} \{\beta_{t,j}^*\} = O_P(1)$, 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 2, we can read the first line in (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 4 *Let assumptions in Theorem 2 hold. Further given Assumption (R2), then the DAFL estimators are consistent. As* $\min\{p, q, T\} \to \infty$*, we also have block consistency such that*

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

Corollary 5 *Under the assumptions in Theorem 4, the final estimators for the main effects have the following properties.*

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

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

2. *Uniform convergence. The final estimators in the dense blocks are convergent such that*

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

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

Theorem 4 is a key result to derive the desired properties in the final estimators. Note that although the DAFL estimators are consistent, to derive a comparable rate of convergence as Theorem 2.1 requires stricter rate conditions than Assumption (R2). This is circumvented in the final estimators, as shown in Corollary 5.2. 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 $q^{-1}\log\big(p\sum_{i=1}^{p}|\mathcal{S}_{\alpha,i}|\big) + q^{-1}\log\big(p\sum_{i=1}^{p}|\mathcal{B}_{\alpha,i}|\big)$ 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 4 and hence Corollary 5.1, we present the behaviors of the main effect estimators in the dense blocks as in Corollary 5.2.

Remark 2 *In Condition (IC1), the row and column main effects are identified such that* $\min_{i\in[p]}\{\alpha^*_{t,i}\} = 0$ *and* $\min_{j\in[q]}\{\beta^*_{t,j}\} = 0$*. This is not the only possibility besides the identification used by Lam and Cen (2024). From the proof of Theorem 1, actually any quantile of the main effects can be used as an identification condition. To be precise, consider the row main effects as an example, which can then be identified by assuming that for some pre-specified constants* $u \in [0,1]$ *and* v*, for any* $t \in [T]$*,*

$$u\text{-th quantile of } \{\alpha^*_{t,1},\ldots,\alpha^*_{t,p}\} = v.$$

Similarly, the column main effects can also be identified using such a quantile condition. These constitute a class of identification conditions for the MEFM framework, and could be of independent interests depending upon the data sets in practice or purpose of the analysis. Examples include the median of $\{\alpha^*_{t,j}\}_{j\in[p]} = 0$ *as a tail-robust framework to allow for outliers in the main effects, under which the main effect estimators can have the same rates of convergence as in Theorem 2 in Lam and Cen (2024) and hence asymptotic normality therein.*

For the current set of (IC1) we are using, asymptotic normality does not follow unfortunately. However, we stress that our (IC1) is the most natural for handling sparsity in the main effects, which is the reason why we chose that for further investigations in this paper.

Remark 3 *In this remark, we elucidate the details in doing inferences on the factor loading matrix estimators* $\widehat{\mathbf{Q}}_r$ *and* $\widehat{\mathbf{Q}}_c$*. As noted in the explanations below Theorem 2, it is by no mere coincidences that*

Theorem 2.3 coincides with Theorems 2 and 3 in Lam and Cen (2024). In detail, the key step lies in the proof of Theorem 2 where we manage to simplify and write $\widetilde{\mathbf{L}}_t = \mathbf{M}_p \mathbf{X}_t \mathbf{M}_q$, so that the estimators for the loadings and core factor can be considered separately from the base effect and main effects. Therefore, the asymptotic normality for each row of the factor loading matrix estimators can be constructed as detailed in Section 4.4 in Lam and Cen (2024), where a consistent HAC-estimator for the covariance matrix is also proposed.

In what follows, we discuss the generality of Assumption (M1) used in Lam and Cen (2024) for the above asymptotic normality to hold. In particular, their Assumption (M1) requires the vector processes $\textbf{vec}(\mathbf{F}_t)$ and $\textbf{vec}(\mathbf{E}_t)$ to be α-mixing, which is to facilitate proofs using central limit theorem for time series without losing too much generality as in Chen and Fan (2023). In fact, given the general linear processes in Assumption (F1) and the additional approximately sparse factor structure in (E1), such an α-mixing condition directly holds when the linear processes have Gaussian innovations. Other than this, as discussed in Section 15.3 in Davidson (2021): "[...] allowing more general distributions for the innovations yields surprising results. Contrary to what might be supposed, having the θ_j tend to zero even at an exponential rate is not sufficient by itself for strong mixing [...]", where θ_j is the coefficient in the linear process, (M1) can be complicated to verify. We refer interested readers to Theorem 15.9 in Davidson (2021) for a fairly general result which requires certain non-trivial smoothness conditions on the innovations' density functions and decays on the coefficients for a univariate linear process to be α-mixing.

4 Practical Implementation

In this section, we discuss the practical optimization to compute our DAFL estimators. We only focus on estimating the row main effects by minimizing (2.7), 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 $\boldsymbol{\alpha}_{\cdot,i} := (\alpha_{1,i}, \dots, \alpha_{T,i})^{\intercal}$, $\mathbf{D}_{\alpha,i} := \big(\mathbf{D}_{\alpha,i}^{(\mathrm{F})\intercal}, \mathbf{D}_{\alpha,i}^{(\mathrm{L})\intercal}\big)^{\intercal}$, where

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

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

then (2.7) can be rewritten as

$$L(\boldsymbol{\alpha}_{\cdot,i}) = \frac{1}{2}\big\|\widetilde{\boldsymbol{\alpha}}_{\cdot,i} - \boldsymbol{\alpha}_{\cdot,i}\big\|_2^2 + \lambda_\alpha \big\|\mathbf{D}\boldsymbol{\alpha}_{\cdot,i}\big\|_1, \tag{4.1}$$

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 $(\boldsymbol{\alpha}_{\cdot,1}^{\mathsf{T}}, \ldots, \boldsymbol{\alpha}_{\cdot,p}^{\mathsf{T}})^{\mathsf{T}}$, 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 (4.1), 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 (4.1), 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 $\mathbf{D}$ with rows restricted on the index set $\mathcal{A} = \{t : (\mathbf{D}\widehat{\boldsymbol{\alpha}}_{\cdot,i})_t = 0\}$. Without the adaptive terms $u_{\alpha,t,i}$ and $\gamma_{\alpha,t,i}$ in (4.1), 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 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

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

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 (2.7), 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 Results

5.1 Simulation

In this subsection, we showcase the numerical performance of the proposed estimators using Monte Carlo experiments. First, we experiment different settings to evaluate consistency results as described in Theorem 2, before further showing the results in Theorem 4 under various sparsity and signal settings.

For the data generating process, we use general linear processes for the noise and factor series $\mathbf{E}_t$ and $\mathbf{F}_t$ as pinned down by 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 (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 $k_{e,r} = k_{e,c} = 2$.

For any $t \in [T]$, the base effect is formed as $\mu_t = v_{\mu,t}$ with each $v_{\mu,t}$ following i.i.d. $\mathcal{N}(2,1)$. We next depict the generating mechanism for the row main effect, and all other main effects are constructed in the similar manners. Consider any $i \in [p]$, we first construct $\boldsymbol{\alpha}^\circ_{\cdot,i}$, before making it fulfilling Assumptions (R2) and (IC1). In detail, 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

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

where $0 \le \pi^{\mathcal{S}}_{\alpha,i} + \pi^{\mathcal{B}}_{\alpha,i} < 2$. Essentially, larger stay-in probabilities imply more occurrence of piecewise blocks, while larger $\pi^{\mathcal{S}}_{\alpha,i}$ combined with smaller $\pi^{\mathcal{B}}_{\alpha,i}$ imply larger $|\mathcal{S}_{\alpha,i}|$ and vice versa. Then given some constants m_α and σ_α, we generate $\alpha^\circ_{t,i}$ for $t \in [T]$ by the following steps:

Step 1. Let $\alpha^\circ_{1,i} \sim |\mathcal{N}(m_\alpha, \sigma^2_\alpha)|$ with probability $(1-\pi^{\mathcal{B}}_{\alpha,i})/(2-\pi^{\mathcal{S}}_{\alpha,i}-\pi^{\mathcal{B}}_{\alpha,i})$, or $\alpha^\circ_{1,i} = 0$ otherwise;

Step 2. For $t = 2, \ldots, T$, generate $\alpha^\circ_{t,i}$ as i.i.d. $|\mathcal{N}(m_\alpha, \sigma^2_\alpha)|$ if $\alpha^\circ_{t,i}$ is in the dense block according to (5.1).

In the proposition below, we present the sparsity properties of the resulted $\boldsymbol{\alpha}^\circ_{\cdot,i}$ from the above steps.

Proposition 1 *For $\{\alpha^\circ_{t,i}\}_{t\in[T]}$ generated by Steps 1–2 above, it holds that:*

(1) The process $\mathbb{1}\{\alpha^{\circ}_{t,i} = 0\}$*,* $t \in [T]$*, is a stationary Markov process with state space* $\{0, 1\}$ *and correspondingly stationary distribution* $(p_{\alpha,i,*}, 1 - p_{\alpha,i,*})$ *where* $p_{\alpha,i,*} = (1 - \pi^{\mathcal{B}}_{\alpha,i})/(2 - \pi^{\mathcal{S}}_{\alpha,i} - \pi^{\mathcal{B}}_{\alpha,i})$*;*

(2) Let $\mathcal{S}^{\circ}_{\alpha,i} = \{t : \alpha^{\circ}_{t,i} = 0\}$ *be the sparse block induced by* $\{\alpha^{\circ}_{t,i}\}_{t\in[T]}$*, then* $\mathbb{E}(|\mathcal{S}^{\circ}_{\alpha,i}|) = Tp_{\alpha,i,*}$*;*

(3) Let $\{t_\ell + 1, \ldots, t_\ell + m_\ell\}$ *be any subset of* $\mathcal{S}^{\circ}_{\alpha,i}$ *such that* $t_\ell, t_\ell + m_\ell + 1 \notin \mathcal{S}^{\circ}_{\alpha,i}$*, then* $\mathbb{E}(m_\ell) = (1 - \pi^{\mathcal{S}}_{\alpha,i})^{-1}$*.*

The proof of the proposition is relegated to the supplement. From (2) and (3) above, we may adjust the overall sparsity in the main effects and the length of each sparse intervals based on the interplay between $\pi^{\mathcal{S}}_{\alpha,i}$ and $\pi^{\mathcal{B}}_{\alpha,i}$. Next, as discussed below Corollary 5, the elements in the dense block are identifiable up to the threshold of order $q^{-1/2}\log^{1/2}(pT)$. For simplicity, all $\alpha^{\circ}_{t,i}$ less than $q^{-1/2}\log^{1/2}(pT)$ are set as zero. Finally, let $\boldsymbol{\alpha}^*_t$ be the same as $\boldsymbol{\alpha}^{\circ}_t$ except that the minimum of $\boldsymbol{\alpha}^{\circ}_t$ is replaced by zero if not already so, so that Condition (IC1) is fulfilled. Analogously, the column main effects can be formed given m_β, σ_β, $\pi^{\mathcal{S}}_{\beta,j}$ and $\pi^{\mathcal{B}}_{\beta,j}$. Every experiment in this subsection is repeated 500 times unless otherwise stated.

5.1.1 Accuracy of estimators

To evaluate the accuracy of our estimators on the factor structure, given any series of parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_t\}_{t\in[T]}$, 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

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

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. If not specified, for simplicity and ease of notation, let $\pi^{\mathcal{S}} := \pi^{\mathcal{S}}_{\alpha,i} = \pi^{\mathcal{S}}_{\beta,j}$, $\pi^{\mathcal{B}} := \pi^{\mathcal{B}}_{\alpha,i} = \pi^{\mathcal{B}}_{\beta,j}$, and $p_* := p_{\alpha,i,*} = p_{\beta,j,*}$. We consider the following settings:

(Ia) **Baseline, weak sparsity.** $T = 100$, $p = q = 40$, $k_r = 1$, $k_c = 2$. All factors are pervasive with $\zeta_{r,j} = \zeta_{c,j} = 0$. Set also $\sigma_\alpha = \sigma_\beta = 1$, and the signal strength $m_\alpha = m_\beta = 1$. The sparseness is weak with $\pi^{\mathcal{S}} = 0.4$, $\pi^{\mathcal{B}} = 0.8$.

(Ib) **One weak factor.** Same as (Ia), but one factor is weak with $\zeta_{r,1} = \zeta_{c,1} = 0.2$.

(Ic) **Stronger sparsity.** Same as (Ib), except that the sparseness is stronger with $\pi^{\mathcal{B}} = 0.4$.

(Id) **Longer series.** Same as (Ic) but $T = 200$.

(Ie) **Larger dimensions.** Same as (Id) but $p = q = 80$.

(IIa–e) **Temporal independence.** Same as (Ia) to (Ie) respectively, except that elements in $\mathbf{F}_t$, $\mathbf{F}_{e,t}$ and $\boldsymbol{\epsilon}_t$ are white noise rather than AR(2).

The six panels in Figure 1 summarize the accuracy of the estimators across different settings. From the transitions from (Ia) to (Ie), introducing weak factors does not affect the accuracy of our estimators on the base and main effects, corroborating the fact that the base and main effect estimations are detached from the estimation of the common component, due to our model identification. On the other hand, the factor loading space errors—hence the common component errors—are inflated by weak factors, aligning to the findings in Lam and Cen (2024).

The detrimental effects of stronger sparsity with respect to initial estimators are shown by the bump from (Ib) to (Ic), since the definitions in (2.2) and (2.3) suffer from larger bias and variance while incorporating more zeros, which leads to the motivation of our DAFL estimator. We would expect a better performance of $\widehat{\boldsymbol{\alpha}}_t^{\mathcal{B}}$ and $\widehat{\boldsymbol{\beta}}_t^{\mathcal{B}}$, demonstrated in Section 5.1.2. As dimensions increase in setting (Ie), the errors of initial estimators drop significantly, as described in Theorem 2. Furthermore, the increase in dimensions and series length benefit both the estimation of the factor loading space and common component. Note also that imposing stronger temporal dependence slightly inflates the variability of every estimator, demonstrated by comparing settings (Ia)–(Ie) and (IIa)–(IIe).

5.1.2 Performance of the final estimators

We consider the following settings to show the behavior of the final estimators in terms of block recovery and improvement compared to the initial estimators.

(IIIa) **Baseline with stronger signal.** Same as (Ia), but with $m_\alpha = m_\beta = 2$. Note that by Proposition 1, we have $p_* = 0.25$ and $\mathbb{E}(m_l) = 5/3$.

(IIIb) **Longer sparse blocks.** Same as (IIIa), but with $\pi^{\mathcal{B}} = 0.4$, so that $p_* = 0.5$ and $\mathbb{E}(m_l) = 5/3$.

(IIIc) **Longer sparse sub-blocks.** Same as (IIIa), but with $\pi^{\mathcal{S}} = 0.8$, $\pi^{\mathcal{B}} = 0.8$, so that $p_* = 0.5$ and $\mathbb{E}(m_l) = 5$.

(IIId) **Larger variation.** Same as (IIIb), but with $\sigma_\alpha = \sigma_\beta = 2$.

(IIIe) **Weaker signal.** Same as (IIIb), but with $m_\alpha = m_\beta = 1$.

(IIIf) **Larger dimensions.** Same as (IIIb), but with $p = q = 80$.

(IIIg) **Longer series.** Same as (IIIf), but with $T = 200$.

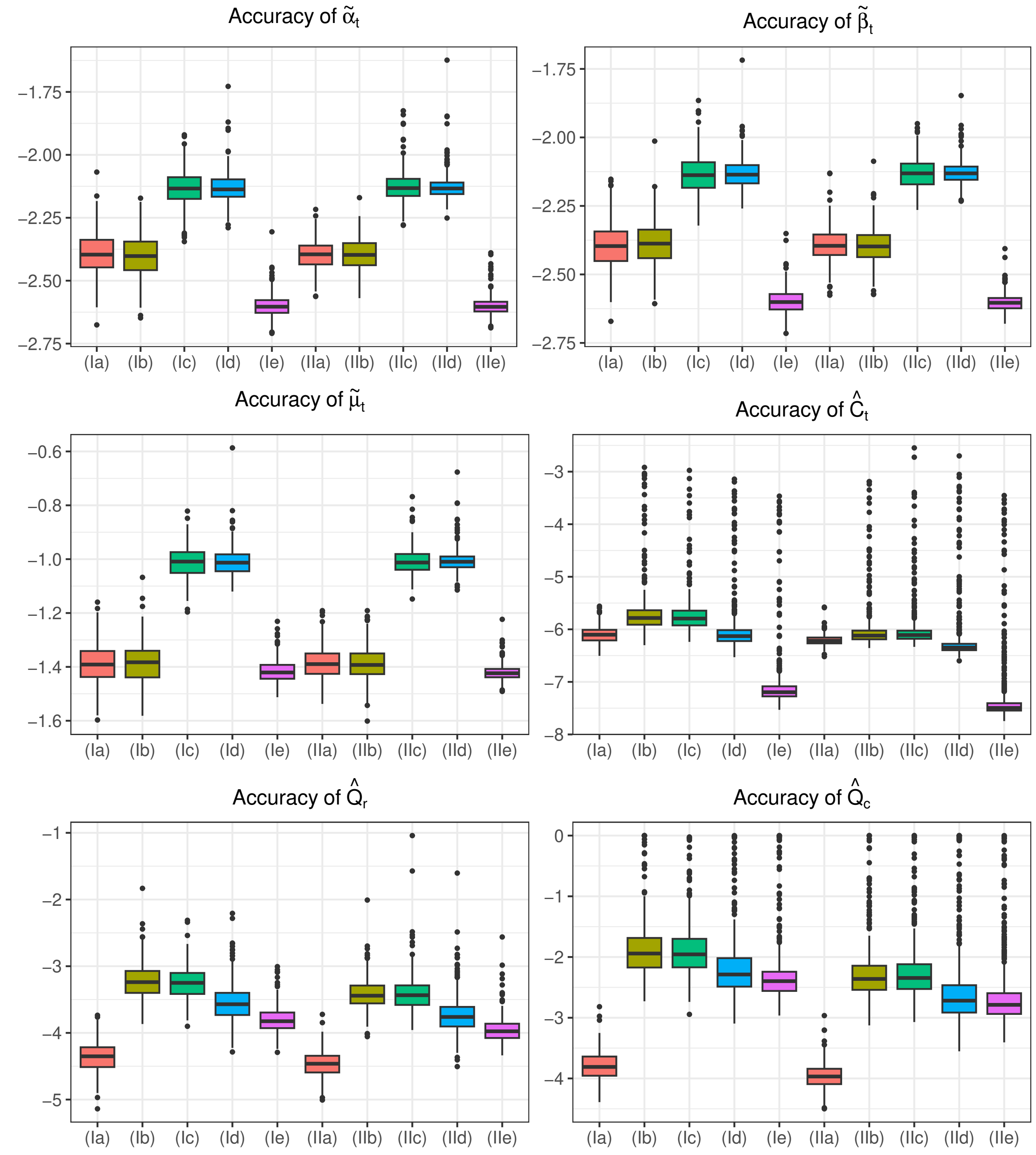

Figure 1: Box plots of the MSEs (in log-scale) for $\widetilde{\boldsymbol{\alpha}}_t$, $\widetilde{\boldsymbol{\beta}}_t$, $\widetilde{\mu}_t$, $\widehat{\mathbf{C}}_t$, $\widehat{\mathbf{Q}}_r$, $\widehat{\mathbf{Q}}_c$ against settings from (Ia) to (Ie), comparing with settings (IIa)–(IIe).

To motivate the different sparsity settings in (IIIb) and (IIIc), parameters are varied to feature different characteristics of sparsity. In detail, we have longer sparse blocks in setting (IIIb) compared to (IIIa) while the expected length of each sub-block, $\mathbb{E}(m_l)$ (see Proposition 1(3)), is the same. Similarly, setting (IIIc) facilitates longer sparse sub-blocks compared to (IIIb) but they have the same p_* controlling the total size of the sparse blocks. Experimenting different length of sparse sub-blocks in setting (IIIc) is necessary, so that we can examine the ability of the fused term in (2.7), designed to recover piecewise sparsity. For every replication, we retain the DAFL estimators $\widehat{\boldsymbol{\alpha}}_t$ and $\widehat{\boldsymbol{\beta}}_t$ under the optimal tuning parameter $\widehat{\lambda}_{C_p}$ selected according to Section 4. To quantify how well we recover the sparse and dense blocks, the sensitivity (true positive rate) and specificity (true negative rate) for the row main effects are defined as

$$\text{sensitivity}_{\boldsymbol{\alpha}} = \frac{\sum_{i=1}^{p} |\{t \in \mathcal{B}_{\alpha,i} : \widehat{\alpha}_{t,i} > 0\}|}{\sum_{i=1}^{p} |\mathcal{B}_{\alpha,i}|}, \quad \text{specificity}_{\boldsymbol{\alpha}} = \frac{\sum_{i=1}^{p} |\{t \in \mathcal{S}_{\alpha,i} : \widehat{\alpha}_{t,i} = 0\}|}{\sum_{i=1}^{p} |\mathcal{S}_{\alpha,i}|}.$$

The two measures for the column effects, sensitivity$_{\boldsymbol{\beta}}$ and specificity$_{\boldsymbol{\beta}}$, are analogously defined. From Table 1, the block selection of our final estimators attains superior performance across all the settings (IIIa)–(IIIg). In particular, the estimators are oracle in view of sensitivity, viz. every truly non-zero entry in both row and column effects is recovered perfectly. It benefits from our tuning parameter selection method, which is calibrated to enforce sparsity without over-penalizing and thus hardly shrinks true signals to zero. Specificity remains virtually unchanged when we lengthen the sparse blocks from settings (IIIa)–(IIIb), demonstrating that our estimator continues to correctly exclude zero entries even as true sparse segments grow larger. It is worth noticing that when the size of sparse sub-blocks increases in (IIIc), specificity improves appreciably. This reflects that the fused penalty in (2.7) enforces piecewise-constant and suppresses total variation, which is suitable for a piecewise-sparse scenario. A larger variation in (IIId) does not undermine the performance, but a weaker signal in (IIIe) leads to a pronounced decline in specificity, underscoring the critical role of signal-to-noise ratio in accurate sparsity recovery. Finally, it is not surprising that the increase in dimensions and length of the series improves the performance and reduces estimation variations.

To visualize the advantage of our final estimator, we compare MSE($\widehat{\boldsymbol{\alpha}}_t^{\mathcal{B}}$) and MSE($\widetilde{\boldsymbol{\alpha}}_t$), MSE($\widehat{\boldsymbol{\beta}}_t^{\mathcal{B}}$), and MSE($\widetilde{\boldsymbol{\beta}}_t$) under settings (IIIa)–(IIIg). The results are shown in Figure 2. In every simulation design, the final estimator achieves a markedly lower median error. The advantage is particularly pronounced when sparsity increases from setting (IIIa) to (IIIb) and (IIIc), where the penalization effectively leverages the sparse structure. However, our final estimator suffers very mildly from larger variation in setting (IIId), but still outperforms the initial estimator. As the dimensionality and the length of the time series grow in (IIIf) and (IIIg), the advantage widens.

Lastly, we further examine the sensitivity and specificity of the final estimators by tuning different

Table 1: Sensitivity and Specificity for final estimators $\widehat{\boldsymbol{\alpha}}_t^{\mathcal{B}}$ and $\widehat{\boldsymbol{\beta}}_t^{\mathcal{B}}$ across settings (IIIa)–(IIIg). Means and standard deviations (bracketed) over 500 replications are displayed.

Setting	sensitivity$_{\boldsymbol{\alpha}}$	sensitivity$_{\boldsymbol{\beta}}$	specificity$_{\boldsymbol{\alpha}}$	specificity$_{\boldsymbol{\beta}}$
(IIIa)	1 (0)	1 (0)	.977 (.008)	.976 (.008)
(IIIb)	1 (0)	1 (0)	.978 (.006)	.978 (.007)
(IIIc)	1 (0)	1 (0)	.994 (.002)	.994 (.002)
(IIId)	1 (0)	1 (0)	.986 (.005)	.986 (.005)
(IIIe)	1 (0)	1 (0)	.913 (.020)	.913 (.019)
(IIIf)	1 (0)	1 (0)	.982 (.006)	.982 (.005)
(IIIg)	1 (0)	1 (0)	.992 (.002)	.992 (.002)

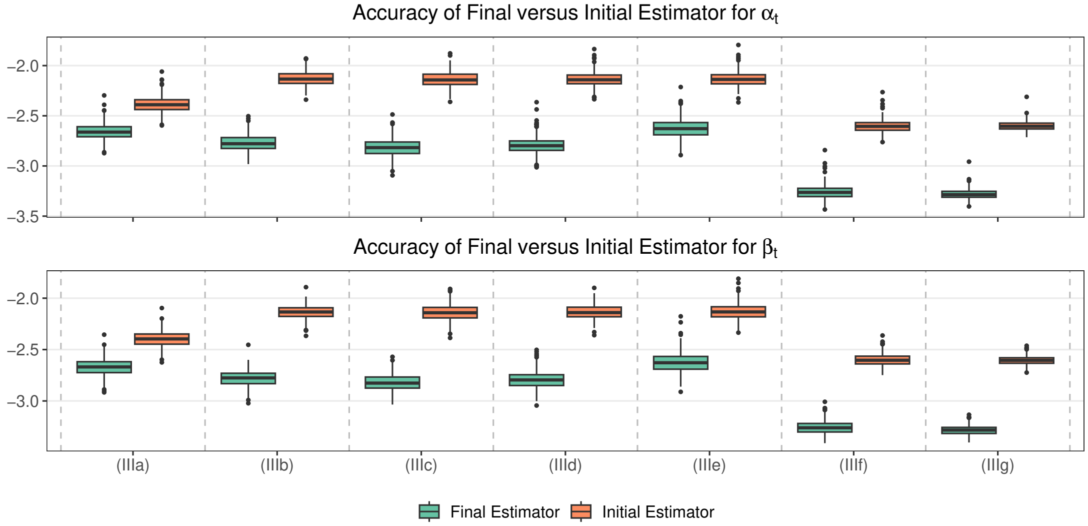

Figure 2: Box plots of the MSEs (in log-scale) of final estimators $\widehat{\boldsymbol{\alpha}}_t^{\mathcal{B}}$ and $\widehat{\boldsymbol{\beta}}_t^{\mathcal{B}}$ versus initial estimators $\widetilde{\boldsymbol{\alpha}}_t$ and $\widetilde{\boldsymbol{\beta}}_t$ against settings from (IIIa) to (IIIg).

parameters: (i) expected sub-block length $\mathbb{E}(m_l)$, which is inspired from the improved performance in setting (IIIc) compared to (IIIb); (ii) dimension p and q; and (iii) the series length T. We present the results for case (i) in Figure 3, and both results for (ii) and (iii) in Figure 4. Here, we only focus on the row main effects as the results for the column main effects are analogous and thus omitted. Sensitivity remains identically 1 in all settings, confirming the robustness of non-zero main effect selection. As $\mathbb{E}(m_l)$ grows, specificity rises monotonically and achieves near 99% when the expected sub-block length is larger, indicating highly accurate sparse-block detection. Likewise, specificity improves monotonically as either the dimensionality p and q or the series length T increases.

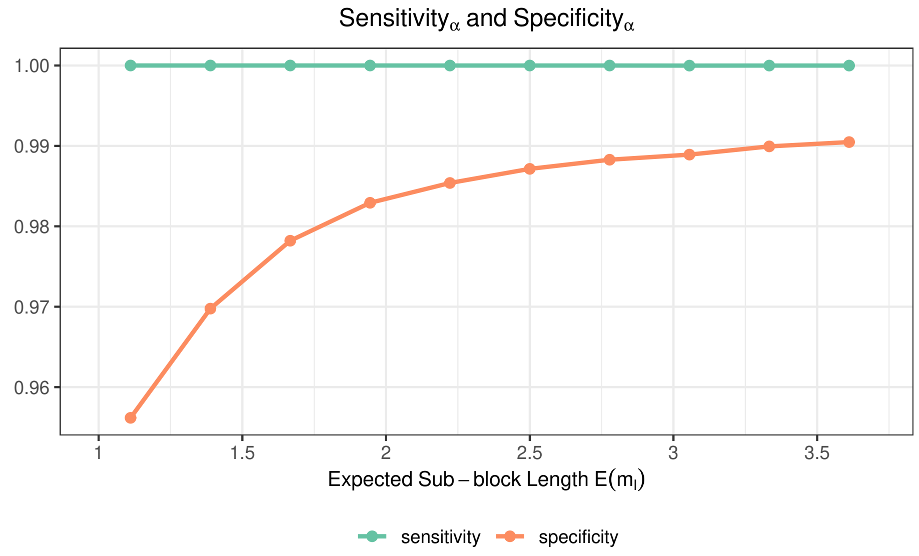

Figure 3: Sensitivity and specificity for the row main effects against different expected lengths of sparse sub-blocks. The setting is the same as (IIIb), but $\mathbb{E}(m_l)$ takes different values by adjusting $\pi^{\mathcal{S}}$ and $\pi^{\mathcal{B}}$.

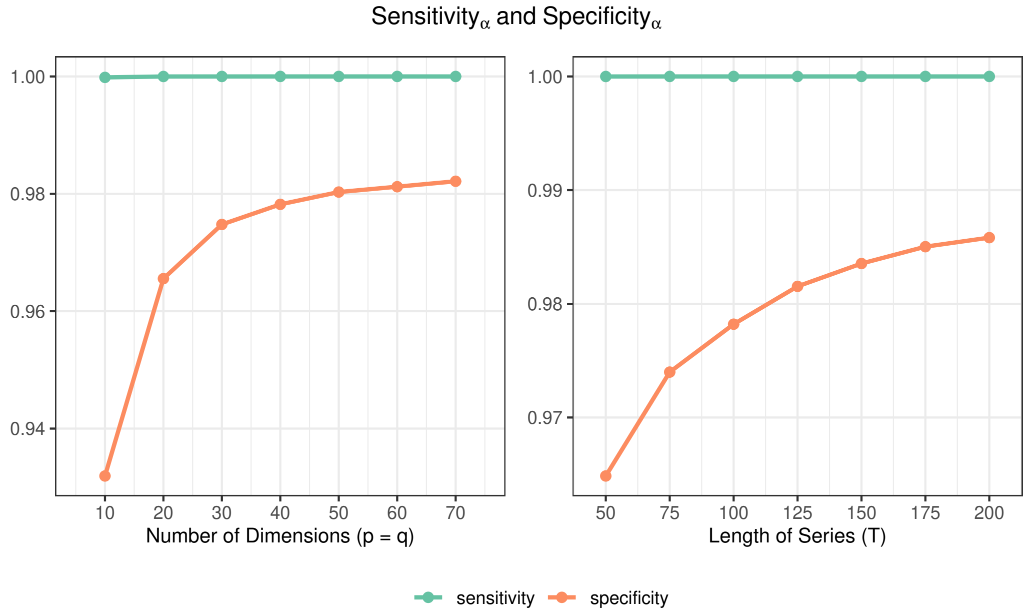

Figure 4: Sensitivity and specificity for the row main effects against different dimensionality p and q and series length T. Same as setting (IIIb) but with corresponding (p, q, T). We set $p = q$ for simplicity.

5.2 Real data analysis

We illustrate the proposed method for MEFM using the publicly available New York City (NYC) Yellow Taxi Trip Records (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 Covid-19 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, where tuning parameters are chosen via the modified C_p criterion described in Section 4.

In Figure 5, 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 Covid-19 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 6 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 Covid-19 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 7. 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, 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.

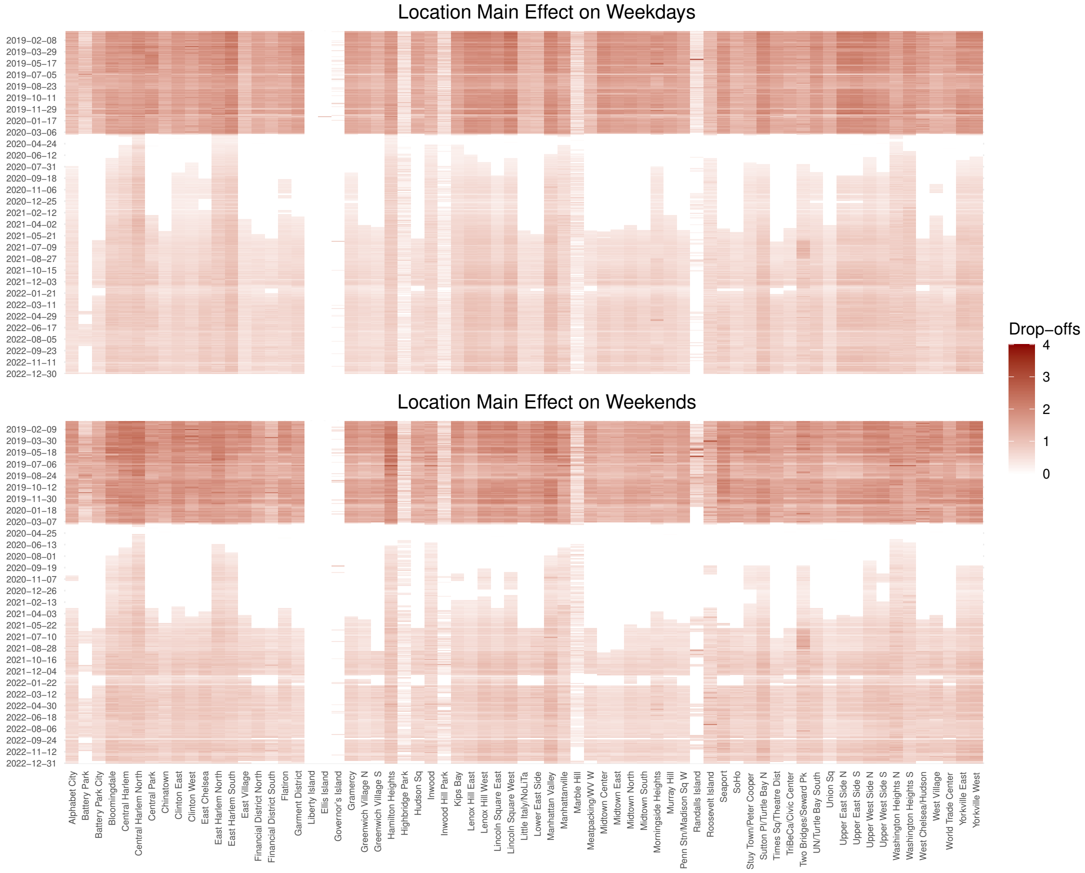

Figure 5: Heatmap of the location main effect $\widehat{\boldsymbol{\alpha}}_t^{\mathcal{B}}$.

6 Concluding Remarks

In this article, we address a critical gap in the literature of matrix factor models by developing methodologies to recover any sparsity structure within. In particular, we leverage MEFM by Lam and Cen (2024) and reconcile the dual demands of interpretability for e.g. applied econometricians, and structural parsimony inherent in matrix-valued time series. Through the flexible, data-driven identification scheme, our approach enables meaningful sparsity in the main effects without forcing unrealistic parameter configurations. This advancement broadens the applicability of matrix factor models to scenarios where localized effects exist,

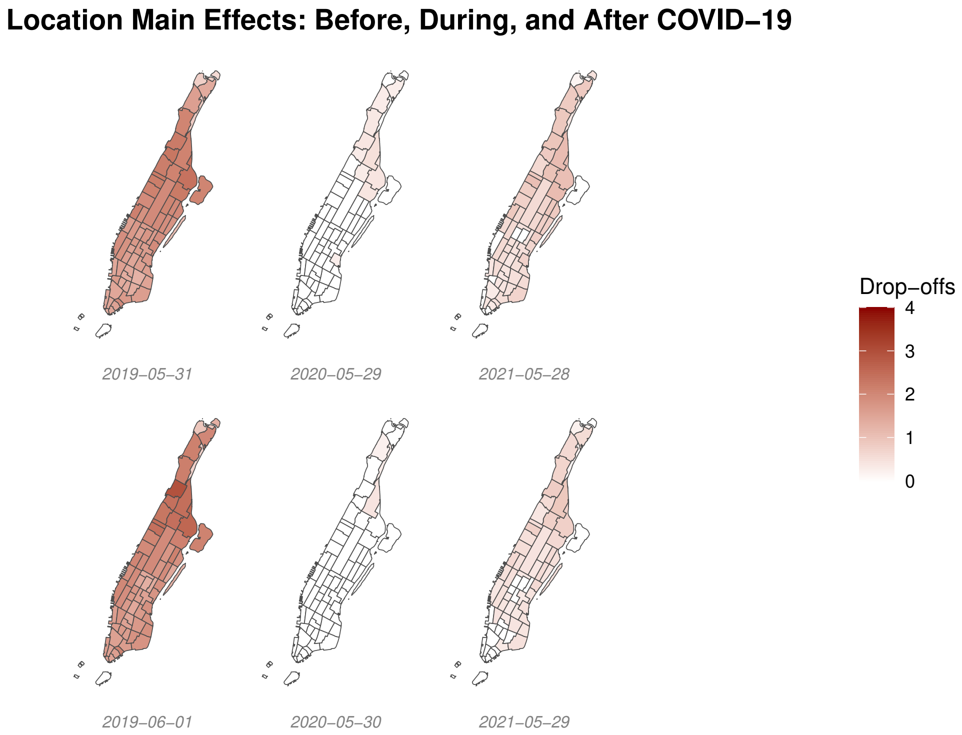

Figure 6: Snapshot heatmaps of location main effect. The top rows are weekdays, and the bottom rows are weekends. The color scale matches Figure 5.

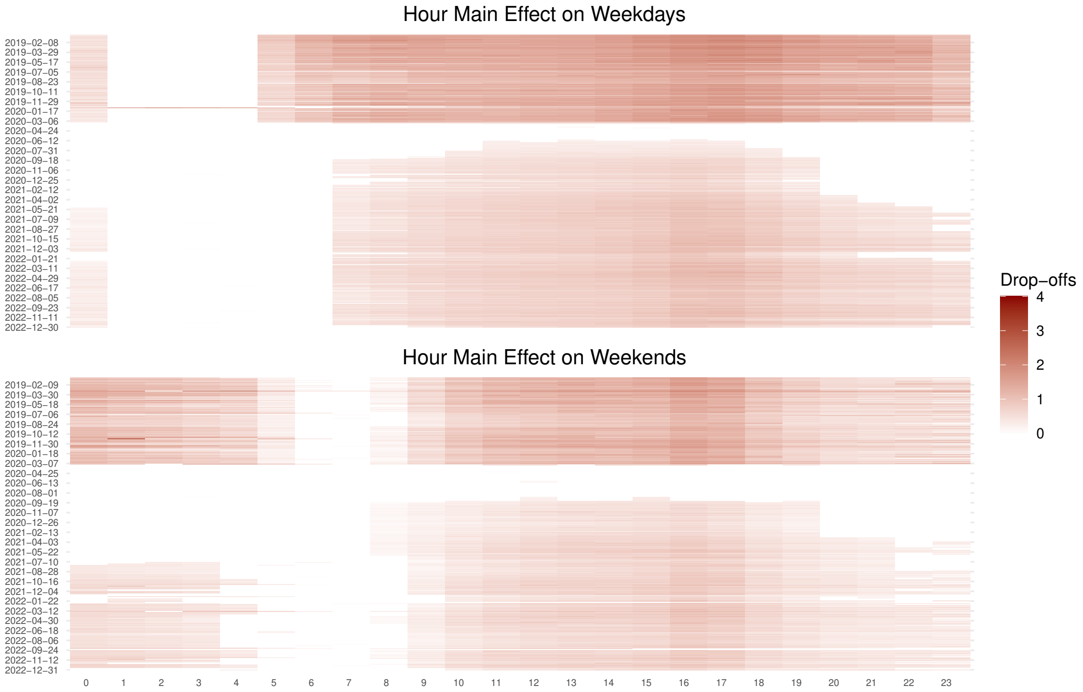

Figure 7: Heatmap of the hour main effect $\widehat{\boldsymbol{\beta}}_t^{\mathcal{B}}$.

as demonstrated by the real example to analyze traffic data. We hope that this article sheds light on the sparsity structure in data analysis for matrix-valued time series or even higher-order tensor time series, which shall provide new insights in addressing dependency in complicated data structures.

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 (2.5), 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.

Supplementary materials for the paper “Sparsity of the Main Effect Matrix Factor Models”

A Technical Proofs

A.1 Proof of Theorems

Proof of Theorem 1. To show (i), suppose we have another set of parameters, $\big(\ddot{\mu}_t, \ddot{\boldsymbol{\alpha}}_t, \ddot{\boldsymbol{\beta}}_t, \ddot{\mathbf{C}}_t\big)$ for $t \in [T]$, also satisfying (2.1). For each $t \in [T]$, right-multiply $\mathbf{1}_q$ on both sides of (2.1), by Assumption (IC1), it holds that

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

Notice that both $(q\mu_t + \mathbf{1}_q^\top\boldsymbol{\beta}_t^*)$ and $(q\ddot{\mu}_t + \mathbf{1}_q^\top\ddot{\boldsymbol{\beta}}_t)$ are scalars, we obtain $\boldsymbol{\alpha}_t^* - \ddot{\boldsymbol{\alpha}}_t = c\mathbf{1}_p$ for some scalar constant c. This implies the index sets $\arg\min_j\{\alpha_{t,j}^*\} = \arg\min_j\{\ddot{\alpha}_{t,j}\}$, which hence uniquely determine the main-effect vector up to a shift. Together with Assumption (IC1), we conclude that $c = 0$ and $\boldsymbol{\alpha}_t^* = \ddot{\boldsymbol{\alpha}}_t$ as a result. We also have $\boldsymbol{\beta}_t^* = \ddot{\boldsymbol{\beta}}_t$ by left-multiplying $\mathbf{1}_p^\top$ on both sides of (2.1) and use a similar argument. For each $t \in [T]$, by respectively left- and right-multiplying $\mathbf{1}_p^\top$ and $\mathbf{1}_q$ on both sides of (2.1), we obtain

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

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 $t \in [T]$.

For (ii), given $\mathbf{X}_t = \acute{\mathbf{A}}_r\mathbf{F}_t\acute{\mathbf{A}}_c^\top + \mathbf{E}_t =: \acute{\mathbf{C}}_t + \mathbf{E}_t$, we can construct for each $t \in [T]$,

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

By Remark 1 in Lam and Cen (2024), it holds immediately that

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

as in (2.1). It is easy to see $\min\{\boldsymbol{\alpha}_t^*\} = 0$ and $\min\{\boldsymbol{\beta}_t^*\} = 0$. Condition (IC1) is indeed satisfied by $\mathbf{M}_a \mathbf{1}_a = \mathbf{0}$ and hence

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

This completes the proof of the theorem. □

Proof of Theorem 2. By Condition (IC1), right-multiplying $\mathbf{1}_q$ on both sides of (2.1), we have

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

Then taking minimum over vector entries on both sides of (A.1) and left-multiplying $\mathbf{1}_p$ give us

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

Substitute $(q\mu_t \mathbf{1}_p)$ back to (A.1), we then have $q\boldsymbol{\alpha}_t^* = \mathbf{X}_t \mathbf{1}_q - \mathbf{1}_p \min\{\mathbf{X}_t \mathbf{1}_q\} - \mathbf{E}_t \mathbf{1}_q + \mathbf{1}_p \min\{\mathbf{E}_t \mathbf{1}_q\}$. Similarly, for $\boldsymbol{\beta}_t^*$, we have $q\boldsymbol{\beta}_t^* = \mathbf{X}_t^\intercal \mathbf{1}_p - \mathbf{1}_q \min\{\mathbf{X}_t^\intercal \mathbf{1}_p\} - \mathbf{E}_t^\intercal \mathbf{1}_p + \mathbf{1}_q \min\{\mathbf{E}_t^\intercal \mathbf{1}_p\}$.

Therefore, for the initial estimators for the main effects, we have

$$\begin{aligned}
\frac{1}{p}\|\widetilde{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}_t^*\|^2 &= \frac{1}{p}\|q^{-1}\mathbf{E}_t \mathbf{1}_q - q^{-1}\mathbf{1}_p \min\{\mathbf{E}_t \mathbf{1}_q\}\|^2, \\
\frac{1}{q}\|\widetilde{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^*\|^2 &= \frac{1}{q}\|p^{-1}\mathbf{E}_t^\intercal \mathbf{1}_p - p^{-1}\mathbf{1}_q \min\{\mathbf{E}_t^\intercal \mathbf{1}_p\}\|^2.
\end{aligned}$$

Furthermore, by Lemma 1,

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

Similarly, $\|p^{-1}\mathbf{E}_t^\intercal \mathbf{1}_p\|^2 = O_P(qp^{-1})$. Then, consider a series of sub-Gaussian random variables $\{X_i\}_{i=1}^p$ with variance proxy σ^2. For any $\lambda \geq 0$, by the Jensen's inequality,

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

implying that $\mathbb{E}(\max_{i\in[p]}\{X_i\}) \leq \log(p)/\lambda + \lambda\sigma^2/2 \leq \sqrt{2\sigma^2 \log(p)}$, where equality holds in the last inequality only when $\lambda = \sqrt{2\log(p)}/\sigma$. Therefore, we conclude $\mathbb{E}(\max_{i\in[p]} |X_i|) \leq \sqrt{2\sigma^2 \log(2p)}$.

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 $\sigma^2 \asymp q$. Subsequently, we have $\mathbb{E}(\max_{i\in[p]}\{|\sum_{j=1}^q E_{t,ij}|\}) = O(\sqrt{q\log(p)})$, yielding $\|q^{-1}\mathbf{1}_p \min\{\mathbf{E}_t\mathbf{1}_q\}\|^2 = O_P(q^{-1}p\log(p))$. Together with (A.2), we have

$$\frac{1}{p}\cdot\|\widetilde{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}_t^*\|^2 = O_P(q^{-1}\log(p)), \quad \frac{1}{q}\cdot\|\widetilde{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^*\|^2 = O_P(p^{-1}\log(q)).$$

In addition,

$$\begin{aligned}(\widetilde{\mu}_t - \mu_t)^2 &= \big\{p^{-1}\mathbf{1}_p^\intercal(\widetilde{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}_t^*) + q^{-1}\mathbf{1}_q^\intercal(\widetilde{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t^*) - (pq)^{-1}\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{1}_q\big\}^2\\ &= \big\{q^{-1}\min\{\mathbf{E}_t\mathbf{1}_q\} + p^{-1}\min\{\mathbf{E}_t^\intercal\mathbf{1}_p\} - (pq)^{-1}\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{1}_q\big\}^2.\end{aligned}$$

We have $(q^{-1}\min\{\mathbf{E}_t\mathbf{1}_q\})^2 = O_P(q^{-1}\log(p))$ from the above result, and similarly $(p^{-1}\min\{\mathbf{E}_t^\intercal\mathbf{1}_p\})^2 = O_P(p^{-1}\log(q))$. As shown in Theorem 2 in Lam and Cen (2024), $\big((pq)^{-1}\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{1}_q\big)^2 = O_P(p^{-1}q^{-1})$. Combining all the above results in

$$(\widetilde{\mu}_t - \mu_t)^2 = O_P\Big(\frac{\log(p)}{q}\Big) + O_P\Big(\frac{\log(q)}{p}\Big) + O_P\Big(\frac{1}{pq}\Big) = O_P\Big\{\max\Big(\frac{\log(p)}{q}, \frac{\log(q)}{p}\Big)\Big\}.$$

For the next step, we show the consistency of the factor loading estimators. Recall the notation that $\mathbf{M}_a = \mathbf{I}_a - a^{-1}\mathbf{1}_a\mathbf{1}_a^\intercal$, from (2.5), we have

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

where the second last equality used the definitions of $\widetilde{\boldsymbol{\alpha}}_t$ and $\widetilde{\boldsymbol{\beta}}_t$ in (2.2) and (2.3) respectively, together with the fact that $\mathbf{M}_a\mathbf{1}_a = \mathbf{0}$. Hence,

$$\begin{aligned}\widetilde{\mathbf{L}}_t\widetilde{\mathbf{L}}_t^\intercal &= \mathbf{M}_p\mathbf{X}_t\mathbf{M}_q\mathbf{M}_q^\intercal\mathbf{X}_t^\intercal\mathbf{M}_p^\intercal = \mathbf{M}_p\mathbf{X}_t\mathbf{M}_q\mathbf{X}_t^\intercal\mathbf{M}_p\\ &= \mathbf{M}_p(\mathbf{C}_t + \mathbf{E}_t)\mathbf{M}_q(\mathbf{C}_t^\intercal + \mathbf{E}_t^\intercal)\mathbf{M}_p = \mathbf{C}_t\mathbf{C}_t^\intercal + \mathbf{C}_t\mathbf{E}_t^\intercal\mathbf{M}_p + \mathbf{M}_p\mathbf{E}_t\mathbf{C}_t^\intercal + \mathbf{M}_p\mathbf{E}_t\mathbf{M}_q\mathbf{E}_t^\intercal\mathbf{M}_p\\ &= \mathbf{C}_t\mathbf{C}_t^\intercal + \mathbf{C}_t\mathbf{E}_t^\intercal + \mathbf{E}_t\mathbf{C}_t^\intercal + \mathbf{E}_t\mathbf{E}_t^\intercal + (pq)^{-1}\mathbf{E}_t\mathbf{1}_q\mathbf{1}_q^\intercal\mathbf{E}_t^\intercal\mathbf{1}_p\mathbf{1}_p^\intercal + (pq)^{-1}\mathbf{1}_p\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{1}_q\mathbf{1}_q^\intercal\mathbf{E}_t^\intercal\\ &\quad - p^{-1}\mathbf{C}_t\mathbf{E}_t^\intercal\mathbf{1}_p\mathbf{1}_p^\intercal - p^{-1}\mathbf{1}_p\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{C}_t^\intercal - p^{-1}\mathbf{E}_t\mathbf{E}_t^\intercal\mathbf{1}_p\mathbf{1}_p^\intercal - p^{-1}\mathbf{1}_p\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{E}_t^\intercal - q^{-1}\mathbf{E}_t\mathbf{1}_q\mathbf{1}_q^\intercal\mathbf{E}_t^\intercal\\ &\quad + p^{-2}\mathbf{1}_p\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{E}_t^\intercal\mathbf{1}_p\mathbf{1}_p^\intercal - (pq)^{-1}p^{-1}\mathbf{1}_p\mathbf{1}_p^\intercal\mathbf{E}_t\mathbf{1}_q\mathbf{1}_q^\intercal\mathbf{E}_t^\intercal\mathbf{1}_p\mathbf{1}_p^\intercal,\end{aligned}\tag{A.3}$$

which is exactly the same as the result in the proof of Theorem 2 in Lam and Cen (2024). Similarly, it holds for $\widetilde{\mathbf{L}}_t^\intercal \widetilde{\mathbf{L}}_t$ that

$$\begin{aligned}\widetilde{\mathbf{L}}_t^\intercal \widetilde{\mathbf{L}}_t = \mathbf{C}_t^\intercal \mathbf{C}_t + \mathbf{C}_t^\intercal \mathbf{E}_t + \mathbf{E}_t^\intercal \mathbf{C}_t + \mathbf{E}_t^\intercal \mathbf{E}_t + (pq)^{-1}\mathbf{E}_t^\intercal \mathbf{1}_p \mathbf{1}_p^\intercal \mathbf{E}_t \mathbf{1}_q \mathbf{1}_q^\intercal + (pq)^{-1}\mathbf{1}_q \mathbf{1}_q^\intercal \mathbf{E}_t^\intercal \mathbf{1}_p \mathbf{1}_p^\intercal \mathbf{E}_t \\ - q^{-1}\mathbf{C}_t^\intercal \mathbf{E}_t \mathbf{1}_q \mathbf{1}_q^\intercal - q^{-1}\mathbf{1}_q \mathbf{1}_q^\intercal \mathbf{E}_t^\intercal \mathbf{C}_t - q^{-1}\mathbf{E}_t^\intercal \mathbf{E}_t \mathbf{1}_q \mathbf{1}_q^\intercal - q^{-1}\mathbf{1}_q \mathbf{1}_q^\intercal \mathbf{E}_t^\intercal \mathbf{E}_t - p^{-1}\mathbf{E}_t^\intercal \mathbf{1}_p \mathbf{1}_p^\intercal \mathbf{E}_t \\ + q^{-2}\mathbf{1}_q \mathbf{1}_q^\intercal \mathbf{E}_t^\intercal \mathbf{E}_t \mathbf{1}_q \mathbf{1}_q^\intercal - (pq)^{-1}q^{-1}\mathbf{1}_q \mathbf{1}_q^\intercal \mathbf{E}_t^\intercal \mathbf{1}_p \mathbf{1}_p^\intercal \mathbf{E}_t \mathbf{1}_q \mathbf{1}_q^\intercal. \qquad \text{(A.4)}\end{aligned}$$

To ease the notation, define

$$\begin{aligned}\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, \qquad \text{(A.5)} \\ \mathbf{R}_{c,t} := \mathbf{C}_t^\intercal \mathbf{E}_t + \mathbf{E}_t^\intercal \mathbf{C}_t + \mathbf{E}_t^\intercal \mathbf{E}_t + (pq)^{-1}\mathbf{E}_t^\intercal \mathbf{1}_p \mathbf{1}_p^\intercal \mathbf{E}_t \mathbf{1}_q \mathbf{1}_q^\intercal + (pq)^{-1}\mathbf{1}_q \mathbf{1}_q^\intercal \mathbf{E}_t^\intercal \mathbf{1}_p \mathbf{1}_p^\intercal \mathbf{E}_t \\ - q^{-1}\mathbf{C}_t^\intercal \mathbf{E}_t \mathbf{1}_q \mathbf{1}_q^\intercal - q^{-1}\mathbf{1}_q \mathbf{1}_q^\intercal \mathbf{E}_t^\intercal \mathbf{C}_t - q^{-1}\mathbf{E}_t^\intercal \mathbf{E}_t \mathbf{1}_q \mathbf{1}_q^\intercal - q^{-1}\mathbf{1}_q \mathbf{1}_q^\intercal \mathbf{E}_t^\intercal \mathbf{E}_t - p^{-1}\mathbf{E}_t^\intercal \mathbf{1}_p \mathbf{1}_p^\intercal \mathbf{E}_t \\ + q^{-2}\mathbf{1}_q \mathbf{1}_q^\intercal \mathbf{E}_t^\intercal \mathbf{E}_t \mathbf{1}_q \mathbf{1}_q^\intercal - (pq)^{-1}q^{-1}\mathbf{1}_q \mathbf{1}_q^\intercal \mathbf{E}_t^\intercal \mathbf{1}_p \mathbf{1}_p^\intercal \mathbf{E}_t \mathbf{1}_q \mathbf{1}_q^\intercal, \qquad \text{(A.6)}\end{aligned}$$

so that from (A.3) and (A.4), we can write

$$\widetilde{\mathbf{L}}_t \widetilde{\mathbf{L}}_t^\intercal = \mathbf{C}_t \mathbf{C}_t^\intercal + \mathbf{R}_{r,t}, \quad \widetilde{\mathbf{L}}_t^\intercal \widetilde{\mathbf{L}}_t = \mathbf{C}_t^\intercal \mathbf{C}_t + \mathbf{R}_{c,t}.$$

The remaining steps are exactly the same as in the proof of Theorem 2 in Lam and Cen (2024). Hence,

$$\begin{aligned}\left\|\widehat{\mathbf{Q}}_r - \mathbf{Q}_r \mathbf{H}_r^\intercal\right\|_F^2 = O_P\Big(T^{-1}p^{2(1-\delta_{r,k_r})}q^{1-2\delta_{c,1}} + p^{1-2\delta_{r,k_r}}q^{2(1-\delta_{c,1})}\Big), \\ \left\|\widehat{\mathbf{Q}}_c - \mathbf{Q}_c \mathbf{H}_c^\intercal\right\|_F^2 = O_P\Big(T^{-1}q^{2(1-\delta_{c,k_c})}p^{1-2\delta_{r,1}} + q^{1-2\delta_{c,k_c}}p^{2(1-\delta_{r,1})}\Big),\end{aligned}$$

where $\mathbf{H}_r$ and $\mathbf{H}_c$ are asymptotically invertible by Lemma 5 in Lam and Cen (2024). Finally, the rates of the error of estimated factor series and individual common components are identical to Theorem 3 in Lam and Cen (2024), as well as their proofs, which completes the proof of Theorem 2. □

Proof of Theorem 4. 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 (2.7) in estimating the row main effects. Define $\widetilde{\boldsymbol{\alpha}}_{\cdot,i} = (\widetilde{\alpha}_{1,i}, \dots, \widetilde{\alpha}_{T,i})^\intercal$ and $\widehat{\boldsymbol{\alpha}}_{\cdot,i} = (\widehat{\alpha}_{1,i}, \dots, \widehat{\alpha}_{T,i})^\intercal$, for the ease of notation. By the KKT condition, $\widehat{\boldsymbol{\alpha}}_{\cdot,i}$ is a solution minimizing the loss function in (2.7) if and only if there exists a subgradient

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

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

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

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

$$h_t : \left\{\begin{array}{ll} h_t = u_{\alpha,t,i}\operatorname{sign}(\widehat{\alpha}_{t,i} - \widehat{\alpha}_{t-1,i}) - u_{\alpha,t+1,i}\operatorname{sign}(\widehat{\alpha}_{t+1,i} - \widehat{\alpha}_{t,i}), & \widehat{\alpha}_{t-1,i} \neq \widehat{\alpha}_{t,i} \neq \widehat{\alpha}_{t+1,i}; \\ \big|h_t + u_{\alpha,t+1,i}\operatorname{sign}(\widehat{\alpha}_{t+1,i} - \widehat{\alpha}_{t,i})\big| \leq u_{\alpha,t,i}, & \widehat{\alpha}_{t-1,i} = \widehat{\alpha}_{t,i} \neq \widehat{\alpha}_{t+1,i}; \\ \big|h_t - u_{\alpha,t,i}\operatorname{sign}(\widehat{\alpha}_{t,i} - \widehat{\alpha}_{t-1,i})\big| \leq u_{\alpha,t+1,i}, & \widehat{\alpha}_{t-1,i} \neq \widehat{\alpha}_{t,i} = \widehat{\alpha}_{t+1,i}; \\ \big|h_t\big| \leq u_{\alpha,t,i} + u_{\alpha,t+1,i}, & \text{otherwise.} \end{array}\right\}, \tag{A.7}$$

and also a subgradient

$$\mathbf{g} = \partial\Big(\sum_{t=1}^{T} \gamma_{\alpha,t,i}\big|\widehat{\alpha}_{t,i}\big|\Big) = \left\{\mathbf{g} \in \mathbb{R}^T : \begin{cases} g_t = \gamma_{\alpha,t,i}\operatorname{sign}(\widehat{\alpha}_{t,i}), & \widehat{\alpha}_{t,i} \neq 0; \\ |g_t| \leq \gamma_{\alpha,t,i}, & \text{otherwise.} \end{cases}\right\}, \tag{A.8}$$

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

$$\widehat{\boldsymbol{\alpha}}_{\cdot,i} - \widetilde{\boldsymbol{\alpha}}_{\cdot,i} = -\partial\Big(\lambda_\alpha \sum_{t=2}^{T} u_{\alpha,t,i}\big|\widehat{\alpha}_{t,i} - \widehat{\alpha}_{t-1,i}\big| + \lambda_\alpha \sum_{t=1}^{T} \gamma_{\alpha,t,i}\big|\widehat{\alpha}_{t,i}\big|\Big) = -\lambda_\alpha \mathbf{h} - \lambda_\alpha \mathbf{g}. \tag{A.9}$$

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 (A.7), 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. (A.9) implies that uniformly over $i \in [p]$ and $t \in \mathcal{S}_{\alpha,i}$,

$$\widetilde{\alpha}_{t,i} = \lambda_\alpha h_t + \lambda_\alpha g_t, \tag{A.10}$$

where $|g_t| \leq \gamma_{\alpha,t,i}$ according to (A.8). 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 (A.7) can only take values $\big|h_t\big| \leq u_{\alpha,t,i} + u_{\alpha,t+1,i}$. Hence for (A.10), we need to show that (uniformly over $i \in [p]$ and $t \in \overline{\mathcal{S}}_{\alpha,i}$),

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

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

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

where the second equality indeed holds true from Assumption (R2) and the first result of Lemma 3.

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

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

We only consider the first scenario above as the second follows a similar proof. Notice that $\widetilde{\alpha}_{t,i} \geq 0$, so with (A.10), it suffices to show that (uniformly over $i \in [p]$ and $t \in \overline{\mathcal{S}}_{\alpha,i}$ with $t+1 \in \mathcal{B}_{\alpha,i}$),

$$\begin{cases} -\lambda_\alpha u_{\alpha,t+1,i} - \lambda_\alpha u_{\alpha,t,i} - \lambda_\alpha \gamma_{\alpha,t,i} \leq 0, \\ \widetilde{\alpha}_{t,i} \leq -\lambda_\alpha u_{\alpha,t+1,i} + \lambda_\alpha u_{\alpha,t,i} + \lambda_\alpha \gamma_{\alpha,t,i}, \end{cases} \tag{A.11}$$

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

$$\max_{i\in[p]} \max_{t\in\mathcal{S}_{\alpha,i}:t+1\in\mathcal{B}_{\alpha,i}} \{\widetilde{\alpha}_{t,i}\} = o_P\big(\min_{i\in[p]} \min_{t\in\mathcal{S}_{\alpha,i}:t+1\in\mathcal{B}_{\alpha,i}} \{\widetilde{\alpha}_{t+1,i}\}\big). \tag{A.12}$$

In detail, we have

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

which, together with Assumption (R2) and Lemma 3, shows (A.12).

Lastly, consider Case (c). If $t = 1$, then we have $|h_1| \le u_{\alpha,2,i}$; otherwise if $t = T$, then $|h_T| \le u_{\alpha,T,i}$. Suppose $t = 1$, with (A.10), it is required that

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

Similar to Case (a), it is sufficient to show $\max_{i\in[p]}\{\widetilde{\alpha}^2_{1,i}\} \le \max_{i\in[p]} \|(\widetilde{\boldsymbol{\alpha}}_{\cdot,i})_{\mathcal{S}_{\alpha,i}}\|^2_{\max} = o_P(\lambda_\alpha)$, 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 $t \in \mathcal{S}^\circ_{\alpha,i}$ for consistency in the sparse block. Using (A.7), we require uniformly over $i \in [p]$ and $t \in \mathcal{S}^\circ_{\alpha,i}$ that

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

which holds true if we can show

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

The first equality is direct from Lemma 3 and Assumption (R2), while the second equality from (A.12).

For sign consistency in $\mathcal{B}_{\alpha,i}$, from (A.9), 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 (A.7) and (A.8), it suffices to show λ_α is stochastically dominated by $\min_{i\in[p]}\min_{t\in\mathcal{B}_{\alpha,i}}\{\widetilde{\alpha}^2_{t,i}\}$. This is direct by combining Assumption (R2), Lemma 3, and

$$\min_{i\in[p]}\min_{t\in\mathcal{B}_{\alpha,i}}\{\widetilde{\alpha}_{t,i}\} \geq \min_{i\in[p]}\min_{t\in\mathcal{B}_{\alpha,i}}\{\alpha^*_{t,i}\} - \max_{i\in[p]}\max_{t\in\mathcal{B}_{\alpha,i}}|\widetilde{\alpha}_{t,i}-\alpha^*_{t,i}|. \tag{A.13}$$

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 (A.9) and the subgradients from (A.7) and (A.8), we have

$$\begin{aligned}|\mathcal{I}_1| \leq \lambda_\alpha|u_{\alpha,t,i}+u_{\alpha,t+1,i}+\gamma_{\alpha,t,i}| &\leq \frac{\lambda_\alpha}{\min_{i\in[p]}\min_{t\in\mathcal{B}_{\alpha,i}}\{\widetilde{\alpha}_{t,i}\}} \\ &\leq \frac{\lambda_\alpha}{\min_{i\in[p]}\min_{t\in\mathcal{B}_{\alpha,i}}\{\alpha^*_{t,i}\} - \max_{i\in[p]}\max_{t\in\mathcal{B}_{\alpha,i}}|\widetilde{\alpha}_{t,i}-\alpha^*_{t,i}|},\end{aligned} \tag{A.14}$$

where the last inequality used (A.13). 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 (A.14), $|\mathcal{I}_1| = o_P(1)$. 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

$$\begin{aligned}\widetilde{\alpha}_{t,i}-\alpha^*_{t,i} &= q^{-1}\mathbf{E}^\intercal_{t,i\cdot}\mathbf{1}_q - q^{-1}\min\{\mathbf{E}_t\mathbf{1}_q\} \\ &= q^{-1}(\mathbf{A}_{e,r})^\intercal_{i\cdot}\mathbf{F}_{e,t}\mathbf{A}^\intercal_{e,c}\mathbf{1}_q + q^{-1}\sum_{j=1}^{q}\Sigma_{\varepsilon,ij}\varepsilon_{t,ij} - q^{-1}\min\{\mathbf{E}_t\mathbf{1}_q\}.\end{aligned} \tag{A.15}$$

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, $\mathbb{E}(q^{-1}(\boldsymbol{\Sigma}_\epsilon\circ\boldsymbol{\epsilon}_t)\mathbf{1}_q) = \mathbf{0}$ 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 $|q^{-1}((\boldsymbol{\Sigma}_\epsilon\circ\boldsymbol{\epsilon}_t)\mathbf{1}_q)_i| = O_P(q^{-1/2})$. 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 2. 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 (A.15) 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. □

A.2 Auxiliary results and proofs

Proof of Corollary 3. It is directly from Theorem 2.2 and Theorem 2.3. □

Proof of Corollary 5. Result 1 is direct from Theorem 4, while result 2 is immediate by Lemma 3, 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 $t = 2, \dots, T$, 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 $\{t_\ell+1, \dots, t_\ell+m_\ell\}$. Every run of consecutive zeros is a geometric string, i.e., for $k = 1, 2, \dots$, $\mathbb{P}(\text{block length} = k) = (\pi^{\mathcal{S}})^{k-1}(1-\pi^{\mathcal{S}})$. 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. □

As we adapt the setting of factor structure as in Cen and Lam (2025) and Lam and Cen (2024), we list Lemma 1 in the following for further use and refer readers to Cen and Lam (2025) for the proof in detail.

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

1. *(Weak correlation of noise $\mathbf{E}_t$ across different rows, columns and times). There exists some positive*

constant $C < \infty$ so that for any $t \in [T], i, j \in [p], h \in [q]$,

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

2. *(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,*

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

3. *(Further results on factor $\mathbf{F}_t$). 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 $\min\{T, p, q\} \to \infty$.

Analogously, we list (A.16) to (A.22) in Lemma 2 and see Cen and Lam (2025) for proofs. Further, we prove the rate of $\sum_{t=1}^{T}\mathbf{R}_{r,t}$ defined in (A.5).

Lemma 2 *(Bounding $\sum_{t=1}^{T}\mathbf{R}_{r,t}$). Under Assumptions (F1), (L1), (E1) and (E2), it holds that*

$$\Big\|\sum_{t=1}^{T}\mathbf{C}_t\mathbf{E}_t^\intercal\Big\|_F^2 = O_P(Tp^{1+\delta_{r,1}}q), \tag{A.16}$$

$$\Big\|\sum_{t=1}^{T}\mathbf{E}_t\mathbf{E}_t^\intercal\Big\|_F^2 = O_P(Tp^2q + T^2pq^2), \tag{A.17}$$

$$\Big\|\sum_{t=1}^{T}\mathbf{1}_q^\intercal\mathbf{E}_t^\intercal\mathbf{1}_p\mathbf{E}_t\mathbf{1}_q\mathbf{1}_p^\intercal\Big\|_F^2 = O_P(Tp^3q^2 + T^2p^2q^2), \tag{A.18}$$

$$\Big\|\sum_{t=1}^{T}\mathbf{C}_t\mathbf{E}_t^\intercal\mathbf{1}_p\mathbf{1}_p^\intercal\Big\|_F^2 = O_P(Tp^{3+\delta_{r,1}}q), \tag{A.19}$$

$$\Big\|\sum_{t=1}^{T}\mathbf{E}_t\mathbf{E}_t^\intercal\mathbf{1}_p\mathbf{1}_p^\intercal\Big\|_F^2 = O_P(Tp^4q + T^2p^3q^2), \tag{A.20}$$

$$\Big\|\sum_{t=1}^{T}\mathbf{E}_t\mathbf{1}_q\mathbf{1}_q^\intercal\mathbf{E}_t^\intercal\Big\|_F^2 = O_P(Tp^2q^2 + T^2pq^2), \tag{A.21}$$

$$\Big\|\sum_{t=1}^{T}(\mathbf{1}_q^\intercal\mathbf{E}_t^\intercal\mathbf{1}_p)^2\mathbf{1}_p\mathbf{1}_p^\intercal\Big\|_F^2 = O_P(T^2p^4q^2), \tag{A.22}$$

$$\Big\| \sum_{t=1}^{T} \mathbf{1}_p^\intercal \mathbf{E}_t \mathbf{E}_t^\intercal \mathbf{1}_p \mathbf{1}_p \mathbf{1}_p^\intercal \Big\|_F^2 = O_P(Tp^6 q + T^2 p^5 q^2). \tag{A.23}$$

Thus, with $\mathbf{R}_{r,t}$ defined in (A.5), we have

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

Lemma 3 *Let Assumption (IC1), (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*

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

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

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

Proof of Lemma 3. To show the first result, from (2.1) and Condition (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 (2.2), it holds for any $t \in [T]$,

$$\begin{aligned}
&(\widetilde{\boldsymbol{\alpha}}_{\cdot,i})_t \equiv \widetilde{\alpha}_{t,i} = q^{-1}(\mathbf{X}_t \mathbf{1}_q)_i - q^{-1} \min\{\mathbf{X}_t \mathbf{1}_q\} \\
&= \mu_t + q^{-1}\mathbf{1}_q^\intercal \boldsymbol{\beta}^*_t + \alpha^*_{t,i} + q^{-1}\mathbf{1}_q^\intercal \mathbf{E}_{t,i\cdot} - \min_{j\in[p]} \big\{ \mu_t + q^{-1}\mathbf{1}_q^\intercal \boldsymbol{\beta}^*_t + \alpha^*_{t,j} + q^{-1}\mathbf{1}_q^\intercal \mathbf{E}_{t,j\cdot} \big\} \\
&= \alpha^*_{t,i} + q^{-1}\mathbf{1}_q^\intercal \mathbf{E}_{t,i\cdot} - \min_{j\in[p]} \big\{ \alpha^*_{t,j} + q^{-1}\mathbf{1}_q^\intercal \mathbf{E}_{t,j\cdot} \big\} = \alpha^*_{t,i} + q^{-1}\mathbf{1}_q^\intercal \mathbf{E}_{t,i\cdot} - \min_{j\in[p]} \big\{ q^{-1}\mathbf{1}_q^\intercal \mathbf{E}_{t,j\cdot} \big\},
\end{aligned} \tag{A.24}$$

where the last equality used Assumption (IC1). From (A.24), we have

$$\begin{aligned}
&\big\| (\widetilde{\alpha}_{1,i} - \alpha^*_{1,i}, \ldots, \widetilde{\alpha}_{T,i} - \alpha^*_{T,i})^\intercal_{\mathcal{S}_{\alpha,i}} \big\|_{\max} = \max_{t\in\mathcal{S}_{\alpha,i}} \big| q^{-1}\mathbf{1}_q^\intercal \mathbf{E}_{t,i\cdot} - \min_{j\in[p]} \big\{ q^{-1}\mathbf{1}_q^\intercal \mathbf{E}_{t,j\cdot} \big\} \big| \\
&\le \max_{t\in\mathcal{S}_{\alpha,i}} \big| q^{-1}\mathbf{1}_q^\intercal \mathbf{E}_{t,i\cdot} \big| + \max_{t\in\mathcal{S}_{\alpha,i}} \max_{j\in[p]} \big| q^{-1}\mathbf{1}_q^\intercal \mathbf{E}_{t,j\cdot} \big| \le 2 \max_{t\in\mathcal{S}_{\alpha,i}} \max_{j\in[p]} \big| q^{-1}\mathbf{1}_q^\intercal \mathbf{E}_{t,j\cdot} \big|.
\end{aligned} \tag{A.25}$$

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

$$E_{t,jh} = \mathbf{A}^\intercal_{e,r,j\cdot} \Big(\sum_{w\ge0} a_{e,w} \mathbf{X}_{e,t-w} \Big) \mathbf{A}_{e,c,h\cdot} + \Sigma_{\epsilon,jh} \Big(\sum_{g\ge0} a_{\epsilon,g} X_{\epsilon,t-g,jh} \Big),$$

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 $\max_{i\in[p]} \max_{t\in\mathcal{S}_{\alpha,i}} \max_{j\in[p]} \big|q^{-1}\mathbf{1}_q^\intercal \mathbf{E}_{t,j\cdot}\big| = O_P\big\{q^{-1/2}\log^{1/2}\big(p\sum_{i=1}^{p}|\mathcal{S}_{\alpha,i}|\big)\big\}$. Together with the equation (A.25) 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 3. The remaining result of the lemma can be shown by repeating all previous arguments, except that $\alpha_{t,i}^*$ is non-zero in (A.24). This completes the proof of Lemma 3. □

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, 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.

Davidson, J. (2021). *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford University Press.

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. (2024). Matrix-valued factor model with time-varying main effects. *arXiv preprint arXiv:2406.00128*.

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.