# Sparse and Robust Multivariate Functional Principal Component Analysis for Passenger Flow Pattern Discovery in Metro Systems

Kai Wang and Fugee Tsung

***Abstract*— Modern metro systems in big cities have accumulated large amounts of passenger transit transaction data via the deployment of automatic fare collection (AFC) devices, which could facilitate a thorough analysis of passenger flow dynamics. To discover the underlying passenger flow patterns over all stations in an entire metro system, this paper proposes a multivariate functional principal component analysis (MFPCA) method. The functional integrity of a single daily passenger flow profile at each station is explicitly utilized, and the complex correlation among a multitude of daily passenger flow profiles from all stations is comprehensively modeled. Moreover, to simultaneously improve the interpretability of eigenvectors and mitigate the sensitivity to outliers, the MFPCA is formulated as a minimization problem with both a sparsity and a robustness penalty term. A computationally efficient algorithm is developed accordingly to obtain the eigenvectors. The superiority of our proposed sparse and robust MFPCA (SRMFPCA) is validated using a Hong Kong Mass Transit Railway (MTR) dataset. The derived sparse and smooth eigenvectors can be well interpreted as empirically meaningful passenger flow patterns. The results of our method can be further used as solid foundations for station clustering, correlation analysis and outlier identification.**

***Index Terms*— Functional data analysis (FDA), hierarchical clustering, model interpretability, robust principal component analysis (RPCA), sparsity regularization.**

## I. Introduction

WIDE applications of automatic fare collection (AFC) devices and smart cards in modern large-scale metro systems can generate sheer amounts of passenger transit transaction data every day [1], [2]. For example, the Mass Transit Railway (MTR), a public metro network serving Hong Kong city, can produce over 5.76 million daily patronage records [3]. Such massive transaction data can be processed to infer detailed information on passenger flows at any time during an entire day and at any station within a whole metro system, and are thus extremely useful in various facets of today's intelligent transportation systems, such as the route pattern recognition, travel demand prediction, train operation scheduling, station crowdedness monitoring, among many others [4].

Hinging on a large historical passenger transit transaction dataset, this paper aims to discover the underlying passenger flow patterns in a metro system, which could serve as fundamental inputs for prediction models, control rules and improvement schemes [5]. Note that our scope covers passenger flows in a global metro system rather than at only a few stations or lines, and passenger flows all day long rather than at several limited time points. Typically, the passenger flows at a station over the course of a complete day often exhibit inherent functional properties. For instance, at the entrance gates of a station close to residential areas, the passenger flow increases to a morning peak and then falls down gradually, while it stays evenly and reaches a late-afternoon peak at the exit gates [6]. Another example of a daily passenger inflow profile at a station near business districts in Hong Kong is given in Fig. 1(a), where the longitudinally recorded passenger counts (black dotted line) clearly reveal a hidden and smooth intraday temporal trend (red solid line).

The functional integrity of the daily passenger flow profiles in the time domain makes the functional data analysis (FDA), an ensemble of statistical techniques for analyzing data in forms of curves or functions [7], a proper and adequate tool in our context. In the FDA framework, the passenger flow (inflow or outflow) profiles in different days are regarded as different realizations (with noises) of a random functional variable. That is, a daily passenger flow profile is deemed to be a single functional datum with a continuity property instead of a collection of discrete data points. Such functional representations exploit maximum information in the intraday passenger flow dynamics [5], and eliminate the conventional efforts in manually defining many flow-related microscopic features, such as the average ridership, variance of ridership, morning/afternoon peak duration, etc [8], [9].

### A. Related Works and Research Gap

Metro passenger flow patterns have been characterized rather simply in some previous works by handcrafting different variables or indexes with different focuses, like the morning

Manuscript received May 25, 2020; revised November 29, 2020 and March 1, 2021; accepted April 27, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 71931006, in part by the Hong Kong RGC General Research Funds under Grant 16201718 and Grant 16216119, and in part by the Fellowship of China Post-Doctoral Science Foundation under Grant 2020M673430. The Associate Editor for this article was Q. Zhu. *(Corresponding author: Kai Wang.)*

Kai Wang is with the School of Management, Xi'an Jiaotong University, Xi'an 710049, China, and also with the State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: kwangai@xjtu.edu.cn).

Fugee Tsung is with the Department of Industrial Engineering and Decision Analytics, The Hong Kong University of Science and Technology, Hong Kong (e-mail: season@ust.hk).

This article has supplementary downloadable material available at https://doi.org/10.1109/TITS.2021.3078816, provided by the authors.

Digital Object Identifier 10.1109/TITS.2021.3078816

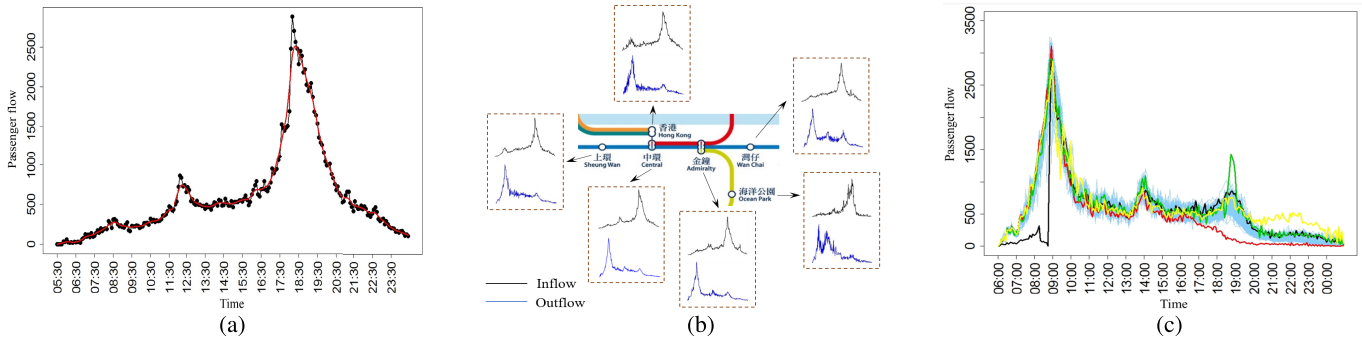



Fig. 1. (a) Passenger inflow profile at the Central station in MTR on Jan 5, 2017. (b) Passenger inflow and outflow profiles at several stations in a subgraph of MTR. (c) Passenger outflow profiles at the Central station in MTR in 2017.

or afternoon peak ridership measuring extreme scenarios of passenger flows [8] and the ratio of morning and afternoon ridership describing balance levels of passenger flows [10]. These summarized characteristics, however, are quite subjective and might cause great information losses as only certain aspects of passenger flows are covered.

A data-driven method that can leverage all the discrete data points in a daily traffic profile to discover flow patterns is the principal component analysis (PCA). Li *et al.* [11], [12] used PCA to extract the intrinsic variation trends from a collection of daily traffic time series which have provided more local and detailed information than the sample mean trends. Traffic flows have been shown to be with Gaussian-type fluctuations and PCA can be applied to retrieve flow features [13]–[15]. A probabilistic PCA is also derived there for missing data imputation. Jiang *et al.* [16] configured the most salient flow features by PCA for traffic node clustering. PCA can also be taken to perform the low-rank decomposition of a traffic data matrix and to clearly identify some telling traffic patterns (e.g., morning and afternoon peaks, school in session, seasonal variation) [17]. A dynamic factor model that is similar to PCA is developed to recognize the latent factors that determine the underlying traffic state [18]. Nevertheless, these PCA-based works all ignore the fact that the traffic flows are measured densely in the time domain and would possess a smoothness or continuity nature as discussed before.

To utilize the hidden property in the daily traffic profiles, FDA is beginning to gain increasing research interest for traffic data analysis. The first work is Chiou [19], where the functional PCA (FPCA), a variant of PCA that considers the continuity nature of functional data and whose eigenfunctions refer to the smooth flow variation patterns, is adopted to cluster the heterogeneous vehicle flow trajectories. After that, a traffic flow trajectory in a new day is dynamically classified over time and its rest part in that day is repeatedly predicted. In light of the independence between the principal component (PC) scores, Wagner-Muns *et al.* [20] developed separate SARIMA time series models for each of the PC scores such that the traffic volume prediction can be greatly simplified.

Besides for traffic flow prediction, FPCA is also employed in [21] for missing data imputation and outlier detection. In this case, the authors computed the expectations of the PC scores. Moreover, by adding to or subtracting from the traffic mean function with the derived eigenfunctions, FPCA has been demonstrated to be powerful in recognizing the underlying temporal patterns in daily traffic flow profiles and can yield clusters with better discrimination [5]. A multivariate control chart using PC scores can be built therein to monitor the day-to-day traffic evolution. The use of FDA in current transportation literature, to the best of our knowledge, has mainly been investigated in the above works, but the revealed utility of FDA definitely deserves much more research attention.

The preceding FPCA methods are all confined to a univariate or single traffic flow profile at a certain location. By contrast, in this paper we would consider passenger flows at all stations in a metro system, and each station has an inflow and an outflow profile (see Fig. 1(b) for example). The discovery of passenger flow patterns thus needs to account for a much more complex data correlation strcuture accordingly. To be specific, the passenger flow at a time point in an inflow/outflow profile might be related to the passenger flow at another time point within the same profile, and it could also be affected by the passenger flows in other profiles (the outflow/inflow profile at the same station, or the inflow and outflow profiles at different stations) when considering all stations collectively. To deal with such coupled within-profile and across-profile correlations, a multivariate FPCA (MFPCA) is required.

### *B. Challenges and Contributions*

For multivariate profiles, different variants of MFPCA have been proposed in the statistics community. In [22], [23], all profiles are first vectorized and then concatenated into a long vector before the regular PCA is applied. This procedure actually enlarges the original data dimension even further. To resolve the pitfall, another commonly used MFPCA decomposes all profiles using the same set of eigenfunctions, and the across-profile interaction is captured by the correlations of the PC scores on each eigenfunction [24]–[28]. Note that these works all first define (e.g., [24], [25]) or separate out (e.g., [26]–[28]) a particular covariance matrix and then apply the eigendecomposition to retrieve the eigenfunctions. However, when applying this MFPCA to our problem, we are still faced with two major technical challenges as described below.

The first challenge concerns model interpretability. As our goal is to explore the underlying passenger flow patterns,

the derived eigenfunctions should be of clear interpretation with understandable practical meanings. The results of conventional PCA, unfortunately, are often difficult to interpret [29]. To solve this shortcoming, a sparse PCA (SPCA) has been developed in [30]–[32] where only the significant elements in an eigenvector are kept and the other negligible elements are zeroed out. Allen [33] and Chen and Lei [34] exploited this sparsity structure in the FPCA for univariate profile, but the works for multivariate profiles are still quite scarce. Two exceptions are Zhang *et al.* [35] and Wang and Tsung [36] where a sparsity in PC scores and a hierarchical sparsity in eigenfunctions are analyzed respectively.

The second challenge lies in model robustness. In reality, due to malfunctioning hardware, extreme weather and special events, outliers can be observed in the collected passenger flow dataset. More severely, the anomaly can occur at any time with an arbitrary duration in a daily passenger flow profile (see various colored abnormal lines superimposed on a group of normal cyan lines in Fig. 1(c) for example). Such complicated outliers of diverse formats corrupt the capability of the existing robust PCA (RPCA) methods in [37]–[40] which rely on some robust covariance statistics. Assuming a low-rank structure of data matrix, RPCA can also be achieved by pursuing outliers in the residual matrix using an $L_1$-norm penalty [41]–[43]. The extension of these RPCA methods to our multivariate profiles and high interpretability context has not yet been studied.

This paper proposes a sparse and robust MFPCA as an integrative method to simultaneously solve the above challenges. We regard the daily passenger inflow and outflow profiles at each station as a bivariate profile, and these bivariate profiles at all stations are decomposed with a shared set of orthonormal eigenfunctions. Then our MFPCA is formulated as a reconstruction error minimization problem, which paves the way for further model improvement. To enhance the interpretability of eigenfunctions and mitigate the sensitivity to outliers, two additional penalty terms inducing the sparisty in eigenfunctions and in residual matrix are introduced together into the objective function. An efficient optimization algorithm based on the coordinate descent method is also developed. The proposed method is finally applied to a large-scale Hong Kong MTR dataset to validate its superiority.

To sum up, the contributions of this paper are highlighted as below:

- A metro-system-level discovery of passenger flow patterns is performed, where the massive passenger inflow and outflow data gathered at all stations and in a complete day are analyzed collectively. The extracted patterns are of clear meanings to reflect the metro-level passenger flow dynamics, and can be further used for station clustering, correlation analysis and outlier identification.
- An MFPCA which can capture both the within-profile and across-profile correlations is developed. Unlike the previous two-step formulation that first configures a particular covariance matrix and then conducts eigendecomposition [24]–[28], we theoretically transform the MFPCA into a unified minimization problem which is able to integrate additional penalty terms to induce favorable properties in the derived eigenfunctions.
- A holistic PCA method which inherits the advantages of MFPCA, SPCA and RPCA jointly is finally proposed. It enjoys high interpretability by producing sparse and smooth eigenfunctions, and is free of diverse outliers that could occur in the real passenger flow profiles. As such, our method is gifted with more capabilities in model interpretation and robustness.

Following the introduction, Section II provides a preliminary description of the Hong Kong MTR passenger flow dataset that acts as a specific background of this paper. Section III elaborates the derivation of the sparse and robust MFPCA. The results of our method in the real dataset studies are presented in Section IV. Section V arrives at the conclusions. More technical details and application results are available in a supplementary material.

## II. Data

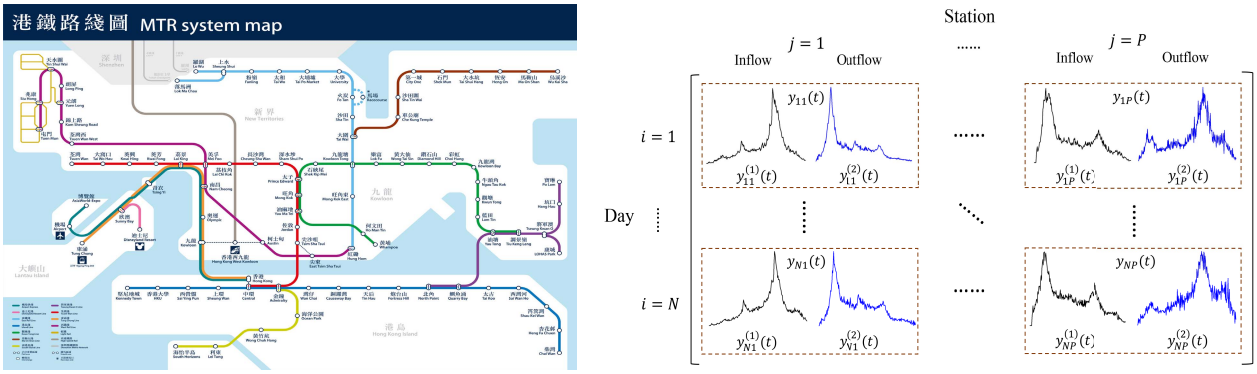
The dataset that motivates the development of our method contains passenger flows in the entire Hong Kong MTR system and in 2017 all year around. The MTR system, which consists of 12 rail lines and 94 stations (see Fig. 2(a) and http://www.mtr.com.hk), has become the top option for Hong Kong citizens' daily trip and commute. As the raw data collected by the AFC devices are transaction records including smart card ID, transaction date, in-gate time, in-gate station, out-gate time, out-gate station and fare, we have transformed them into tap-in (inflow) and tap-out (outflow) passenger counts at each station in a 5-minute time interval. The time periods for the passenger inflow and outflow every day are set to be 5:30AM–0:30AM(+1) and 6:00AM–1:00AM(+1), respectively, as almost no transactions take place in the other time slots. Hence, each station has 228 equidistant data points in its daily passenger inflow and outflow profiles. Note that a few stations are discarded since they suffer from severe data-missing problems in this dataset. In addition, we remove public holidays, weekends, typhoon days (MTR is out of service) and some days with missing data from our calendar as here we would focus on the analysis of normal behaviors of the MTR system on weekdays. The basic information of the dataset under study is summarized in Table I.

## III. Methodology

This section first introduces an MFPCA that is applicable to multivariate profiles and its formulation as a minimization problem. Then a sparse and robust MFPCA is developed to improve the model interpretability and robustness. A computationally efficient algorithm is also derived to obtain the eigenfunctions.

### A. MFPCA

Suppose there are $N$ days and $P$ stations in the passenger flow dataset. Let $y_{ij}^{(1)}(t)$ and $y_{ij}^{(2)}(t)$ denote the passenger inflow and outflow profile at the $j$th station, $j = 1, \ldots, P$, in the $i$th day, $i = 1, \ldots, N$, respectively, where $t \in \mathcal{T}$ is the time domain within a day (see the data structure in Fig. 2(b)).




(a) (b)

Fig. 2. (a) MTR route map. (b) Structure of daily passenger flow profile dataset.

TABLE I
BASIC INFORMATION OF DATASET

| | | |
|---|---|---|
| Number of days $N$ | 233 | Jan: 18, Feb: 18, Mar: 20, Apr: 17, May: 18, Jun: 21, Jul: 21, Aug: 18, Sept: 21, Oct: 20, Nov: 22, Dec: 19. |
| Number of stations $P$ | 79 | Island Line: 17, South Island Line: 4, Tseung Kwan O Line: 5, Kwun Tong Line: 11, East Rail Line: 9, Ma On Shan Line: 8, Tsuen Wan Line: 13, West Rail Line: 9, Tung Chung Line: 3. |
| Number of time intervals $T$ | 228 | Inflow: 5-minute time interval from 5:30AM to 0:30AM (+1, next day), Outflow: 5-minute time interval from 6:00AM to 1:00AM (+1, next day). |

Then $\mathbf{y}_{ij}(t) = \left(y_{ij}^{(1)}(t), y_{ij}^{(2)}(t)\right)^T$ is a bivariate profile and can be modeled in the FDA framework as

$$\mathbf{y}_{ij}(t) = \boldsymbol{\mu}_j(t) + \boldsymbol{\gamma}_{ij}(t), \quad i = 1, \ldots, N, \quad j = 1, \ldots, P, \quad (1)$$

where $\boldsymbol{\mu}_j(t) = \left(\mu_j^{(1)}(t), \mu_j^{(2)}(t)\right)^T$ is the mean function at the $j$th station, and $\boldsymbol{\gamma}_{ij}(t) = \left(\gamma_{ij}^{(1)}(t), \gamma_{ij}^{(2)}(t)\right)^T$ is the realization in the $i$th day of a zero-mean squared-integrable random process $\boldsymbol{\gamma}_j(t) = \left(\gamma_j^{(1)}(t), \gamma_j^{(2)}(t)\right)^T$ associated with the $j$th station.

To represent the random process functions $\boldsymbol{\gamma}_j(t)$'s across all of the $P$ stations, we define a shared set of orthonormal eigenfunctions

$$\{\boldsymbol{v}_k(t) = \left(v_k^{(1)}(t), v_k^{(2)}(t)\right)^T\}_{k=1}^{\infty},$$

which satisfy

$$\int_{\mathcal{T}} \boldsymbol{v}_k(t)^T \boldsymbol{v}_{k'}(t)dt = \int_{\mathcal{T}} v_k^{(1)}(t)v_{k'}^{(1)}(t)dt + \int_{\mathcal{T}} v_k^{(2)}(t)v_{k'}^{(2)}(t)dt = \begin{cases} 1, & \text{if } k = k', \\ 0, & \text{if } k \neq k'. \end{cases}$$

Additionally, the majority of data variation is usually contained in a few leading eigenfunctions. In light of this, Eq. (1) becomes

$$\mathbf{y}_{ij}(t) = \boldsymbol{\mu}_j(t) + \sum_{k=1}^{K} s_{ij,k}\boldsymbol{v}_k(t) + \boldsymbol{\epsilon}_{ij}(t), \quad i = 1, \ldots, N, \quad j = 1, \ldots, P, \quad (2)$$

where $s_{ij,k}$ is the $k$th PC score of $\mathbf{y}_{ij}(t)$ calculated as

$$\begin{aligned} s_{ij,k} &= \int_{\mathcal{T}} \boldsymbol{\gamma}_{ij}(t)^T \boldsymbol{v}_k(t)dt = \int_{\mathcal{T}} \left(\mathbf{y}_{ij}(t) - \boldsymbol{\mu}_j(t)\right)^T \boldsymbol{v}_k(t)dt \\ &= \int_{\mathcal{T}} \left(y_{ij}^{(1)}(t) - \mu_j^{(1)}(t)\right)v_k^{(1)}(t)dt \\ &\quad + \int_{\mathcal{T}} \left(y_{ij}^{(2)}(t) - \mu_j^{(2)}(t)\right)v_k^{(2)}(t)dt, \end{aligned}$$

and $\boldsymbol{\epsilon}_{ij}(t)$ is the remaining noise of magnitude $\sigma^2$. More importantly, for each eigenfunction $\boldsymbol{v}_k(t)$, the corresponding PC scores of the $P$ stations in the $N$ days, denoted by $\mathbf{s}_{i,k} = (s_{i1,k}, \ldots, s_{iP,k})^T \in \mathbf{R}^P$, $i = 1, \ldots, N$, are regarded to be $N$ independent samples generated from a $P$-dimensional normal distribution $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_k)$. In this setting, the relationship between the $j$th and $j'$th station is modeled by the covariance of their PC scores $s_{ij,k}$ and $s_{ij',k}$ under each $k$, i.e., $\mathrm{Cov}(s_{ij,k}, s_{ij',k}) = \sigma_{k,jj'}^2$ where $\sigma_{k,jj'}^2$ is the $(j, j')$ element of $\boldsymbol{\Sigma}_k$.

Based on Eq. (2) and the defined covariance structure among PC scores, we can derive the covariance of passenger flow profiles, either inflow or outflow, at any two stations evaluated at any two time points, i.e.,

$$\begin{aligned} &\mathrm{Cov}\left(y_{ij}^{(l)}(t), y_{ij'}^{(l')}(t')\right) \\ &= \mathrm{Cov}\Big(\mu_j^{(l)}(t) + \sum_{k=1}^{K} s_{ij,k}v_k^{(l)}(t) + \epsilon_{ij}^{(l)}(t), \\ &\quad \mu_{j'}^{(l')}(t') + \sum_{k=1}^{K} s_{ij',k}v_k^{(l')}(t') + \epsilon_{ij'}^{(l')}(t')\Big) \end{aligned}$$

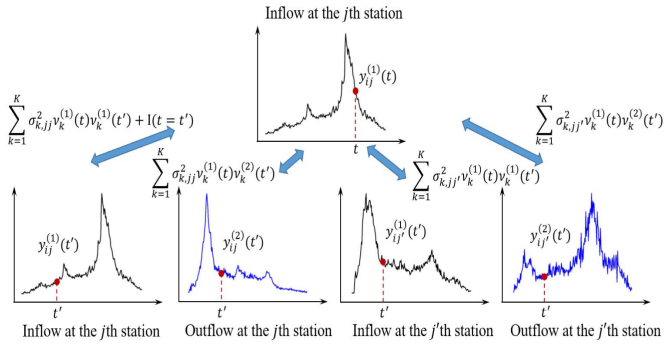



Fig. 3. Within-profile and across-profile covariances in MFPCA.

$$= \sum_{k=1}^{K} \mathrm{Cov}(s_{ij,k}, s_{ij',k}) v_k^{(l)}(t) v_k^{(l')}(t') + \sigma^2 \mathrm{I}(j = j', l = l', t = t') = \sum_{k=1}^{K} \sigma^2_{k,jj'} v_k^{(l)}(t) v_k^{(l')}(t') + \sigma^2 \mathrm{I}(j = j', l = l', t = t'), \quad (3)$$

where $t, t' \in \mathcal{T}$, $l, l' = 1, 2$, $j, j' = 1, \ldots, P$, and $\mathrm{I}(\cdot)$ is the indicator function. From Eq. (3), the within-profile ($l = l'$, $j = j'$) and the across-profile ($l \neq' l$ or $j \neq j'$) covariance can be both easily computed as illustrated in Fig. 3.

### B. MFPCA: A Minimization Formulation

In practice, the daily passenger flow profiles are measured at a finite grid of time points, i.e., $t = 1, \ldots, T$. We let $\mathbf{y}_{ij}^{(1)} = \left(y_{ij}^{(1)}(1), \ldots, y_{ij}^{(1)}(T)\right)^T \in \mathbf{R}^T$, $\mathbf{y}_{ij}^{(2)} = \left(y_{ij}^{(2)}(1), \ldots, y_{ij}^{(2)}(T)\right)^T \in \mathbf{R}^T$ and $\mathbf{y}_{ij} = \left(\mathbf{y}_{ij}^{(1)^T}, \mathbf{y}_{ij}^{(2)^T}\right)^T \in \mathbf{R}^{2T}$ denote the passenger inflow vector, outflow vector and flow vector at the $j$th station in the $i$th day, respectively, and $\boldsymbol{Y}_i = (\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iP})^T \in \mathbf{R}^{P \times 2T}$ the passenger flow matrix stacking all passenger flow vectors from the $P$ stations in the $i$th day. Analogously, we have $\boldsymbol{\mu}_j = \left(\boldsymbol{\mu}_j^{(1)^T}, \boldsymbol{\mu}_j^{(2)^T}\right)^T \in \mathbf{R}^{2T}$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_P)^T \in \mathbf{R}^{P \times 2T}$, $\boldsymbol{v}_k = \left(\boldsymbol{v}_k^{(1)^T}, \boldsymbol{v}_k^{(2)^T}\right)^T \in \mathbf{R}^{2T}$, and $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K) \in \mathbf{R}^{2T \times K}$. Then the matrix version of Eq. (2) is

$$\boldsymbol{Y}_i = \boldsymbol{\mu} + \boldsymbol{S}_i \boldsymbol{V}^T + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, N, \quad (4)$$

where $\boldsymbol{S}_i = (\boldsymbol{Y}_i - \boldsymbol{\mu})\boldsymbol{V} = (\mathbf{s}_{i,1}, \ldots, \mathbf{s}_{i,K}) \in \mathbf{R}^{P \times K}$ is the PC score matrix in the $i$th day, and $\boldsymbol{\epsilon}_i$ is the error matrix.

We now formulate a minimization problem as below:

$$\widehat{\boldsymbol{A}} = \arg\min_{\boldsymbol{A}} \sum_{i=1}^{N} ||\boldsymbol{X}_i - \boldsymbol{X}_i \boldsymbol{A}\boldsymbol{A}^T||_F^2 = ||\boldsymbol{X} - \boldsymbol{X}\boldsymbol{A}\boldsymbol{A}^T||_F^2, \quad \text{s.t. } \boldsymbol{A}^T\boldsymbol{A} = \mathbf{I}, \quad (5)$$

where $\boldsymbol{A} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K) \in \mathbf{R}^{2T \times K}$, $||\cdot||_F$ is the Frobenius norm, $\boldsymbol{X}_i = \boldsymbol{Y}_i - \boldsymbol{\mu} \in \mathbf{R}^{P \times 2T}$ is the centered sample, and $\boldsymbol{X} = (\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_N^T)^T \in \mathbf{R}^{NP \times 2T}$ is the whole data matrix. Then we have

*Proposition 1: As* $N \to \infty$, $|\langle \widehat{\boldsymbol{\alpha}}_k, \boldsymbol{v}_k \rangle| = |\widehat{\boldsymbol{\alpha}}_k^T \boldsymbol{v}_k| \xrightarrow{a.s.} 1$, *for* $k = 1, \ldots, K$.

*Proof:* See Section S-I in the supplementary material. □

Proposition 1 implies that the MFPCA in Eq. (2), which is designed particularly for multivariate profiles, can also be formulated as a minimization problem in (5). The minimum solutions $\widehat{\boldsymbol{\alpha}}_k$'s are consistent estimates of the true eigenvectors $\boldsymbol{v}_k$'s since the absolute value of their inner product converges to one almost surely. Finally, before delivering our sparse and robust MFPCA, we transform Model (5) into an alternative with two decision variables:

$$\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{B}} = \arg\min_{\boldsymbol{A}, \boldsymbol{B}} ||\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^T||_F^2, \quad \text{s.t. } \boldsymbol{A}^T\boldsymbol{A} = \mathbf{I}. \quad (6)$$

*Proposition 2: Models (6) and (5) are equivalent.*

*Proof:* See Section S-I in the supplementary material. □

Actually, $\widehat{\boldsymbol{B}} = \widehat{\boldsymbol{A}}$ in Model (6), but we introduce such an extra variable $\boldsymbol{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) \in \mathbf{R}^{2T \times K}$ as it is free of constraint. This property makes it convenient to impose penalty terms on $\boldsymbol{B}$ to induce certain appealing structures, e.g., sparsity and smoothness, in the derived eigenvectors as will be shown in the following subsection.

### C. Sparse and Robust MFPCA

First, to improve the model interpretability, we incorporate the sparsity idea into the eigenvectors. Specifically, we expect the small-valued elements in an eigenvector to be automatically shrunken as zero. In this way, the remaining non-zero elements would reveal some significant time points and temporal trends within a day of particular meanings (e.g., early-morning and late-afternoon peaks), which can assist the interpretation of the extracted passenger flow patterns. To promote this sparsity, we employ an $L_1$-norm or LASSO penalty term

$$h_1(\boldsymbol{B}) = \lambda_1 ||\boldsymbol{B}||_1 = \lambda_1 \sum_{k=1}^{K} ||\boldsymbol{\beta}_k||_1 = \lambda_1 \sum_{k=1}^{K} \left(||\boldsymbol{\beta}_k^{(1)}||_1 + ||\boldsymbol{\beta}_k^{(2)}||_1\right) = \lambda_1 \sum_{k=1}^{K} \sum_{t=1}^{T} \left(|\beta_k^{(1)}(t)| + |\beta_k^{(2)}(t)|\right) \quad (7)$$

with $\lambda_1 \geq 0$ as the tuning parameter.

Next, we consider the model robustness. It can be observed that if outliers exist, the magnitudes of some elements in the residual matrix $\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^T$ in Model (6) would become abnormally large. In addition, due to the diversity of outliers as discussed in Section I-B, these extreme elements can occur arbitrarily at any locations in the residual matrix. In light of this fact, we define an outlier matrix $\boldsymbol{M} = (\boldsymbol{M}_1^T, \ldots, \boldsymbol{M}_N^T)^T \in \mathbf{R}^{NP \times 2T}$ of the same structure as the data matrix $\boldsymbol{X}$, and assume that the data matrix $\boldsymbol{X}$ can be decomposed into three parts as

$$\boldsymbol{X} = \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^T + \boldsymbol{M} + \boldsymbol{\epsilon}. \quad (8)$$

Generally, the potential outliers in a dataset could only occupy a small fraction, which indicates most elements in $\boldsymbol{M}$ being zero. Therefore, we explore the sparsity in $\boldsymbol{M}$ by considering

another $L_1$-norm penalty term

$$h_2(\boldsymbol{M}) = \lambda_2||\boldsymbol{M}||_1 = \lambda_2 \sum_{i=1}^{N} ||\boldsymbol{M}_i||_1 = \lambda_2 \sum_{i=1}^{N}\sum_{j=1}^{P} ||\mathbf{m}_{ij}||_1$$
$$= \lambda_2 \sum_{i=1}^{N}\sum_{j=1}^{P} \left(||\mathbf{m}_{ij}^{(1)}||_1 + ||\mathbf{m}_{ij}^{(2)}||_1\right)$$
$$+ \lambda_2 \sum_{i=1}^{N}\sum_{j=1}^{P}\sum_{t=1}^{T} \left(|m_{ij}^{(1)}(t)| + |m_{ij}^{(2)}(t)|\right), \quad (9)$$

where $\lambda_2 \geq 0$ is the tuning parameter.

Last but not the least, recall that in reality the daily passenger flow profiles are measured discretely at only a finite number of time points, e.g., 228 time points from 5:30AM to 0:30AM(+1) in the passenger inflow profiles in our MTR dataset, so the eigenfunctions derived from these data are also discretely evaluated. To retain the continuity or smoothness nature of these derived eigenfunctions in the time domain, as in [7], [44], we devise a penalty term which measures the roughness of eigenvectors using the squared values of the numerical second-order derivatives. A smaller roughness indicates a higher degree of smoothness in the eigenvectors. The penalty is defined as

$$h_3(\boldsymbol{B}) = \lambda_3||\boldsymbol{D}\boldsymbol{B}||_F^2 = \lambda_3 \sum_{k=1}^{K} ||\boldsymbol{D}\boldsymbol{\beta}_k||_2^2$$
$$= \lambda_3 \sum_{k=1}^{K} \left(||\widetilde{\boldsymbol{D}}\boldsymbol{\beta}_k^{(1)}||_2^2 + ||\widetilde{\boldsymbol{D}}\boldsymbol{\beta}_k^{(2)}||_2^2\right), \quad (10)$$

where $\lambda_3 \geq 0$ is the tuning parameter, $\boldsymbol{D} = \text{diag}\{\widetilde{\boldsymbol{D}}, \widetilde{\boldsymbol{D}}\}$ is a block diagonal matrix, and

$$\widetilde{\boldsymbol{D}} = \begin{pmatrix} 1 & -2 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & -2 & 1 \end{pmatrix}_{(T-2)\times T}$$

is the numerical second-order derivative operator. A larger $\lambda_3$ induces more smooth $\widehat{\boldsymbol{\beta}}_k$'s.

By combining Eqs. (6)-(10), our sparse and robust MFPCA (SRMFPCA) is finally proposed as

$$\min_{\boldsymbol{A},\boldsymbol{B},\boldsymbol{M}} \; ||\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^T - \boldsymbol{M}||_F^2 + \lambda_1||\boldsymbol{B}||_1 + \lambda_2||\boldsymbol{M}||_1$$
$$+ \lambda_3||\boldsymbol{D}\boldsymbol{B}||_F^2,$$
$$\text{s.t.} \;\; \boldsymbol{A}^T\boldsymbol{A} = \mathbf{I}, \quad (11)$$

and the derived eigenvectors are $\widehat{\boldsymbol{v}}_k = \widehat{\boldsymbol{\beta}}_k/||\widehat{\boldsymbol{\beta}}_k||_2$, $k = 1, \ldots, K$. The tuning parameters $\lambda_1$, $\lambda_2$ and $\lambda_3$ can be properly selected by the AIC, BIC or cross-validations (see more discussion in Section S-II in the supplementary material).

### D. Algorithm

To solve the SRMFPCA in Eq. (11), we develop a coordinate descent algorithm where an efficient update is available for each variable when the other ones are fixed.

- *Step 1:* Initialization. Calculate $\boldsymbol{X}^T\boldsymbol{X}$ and let the initial $\boldsymbol{A}$ and $\boldsymbol{B}$ contain the first $K$ eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$.
- *Step 2:* Given $\boldsymbol{A}$ and $\boldsymbol{B}$,

  $$\widehat{\boldsymbol{M}} = \text{sign}(\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^T) \cdot \left(|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^T| - \frac{\lambda_2}{2}\right)_+,$$

  where $\text{sign}(\cdot)$, $|\cdot|$ and $(\cdot)_+$ are applied in an element-wise manner.
- *Step 3:* Given $\boldsymbol{B}$ and $\boldsymbol{M}$, let $(\boldsymbol{X} - \boldsymbol{M})^T\boldsymbol{X}\boldsymbol{B} = \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Lambda}}\widetilde{\boldsymbol{V}}^T$, and then

  $$\widehat{\boldsymbol{A}} = \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{V}}^T.$$

- *Step 4:* Given $\boldsymbol{A}$ and $\boldsymbol{M}$, let $\mathbf{Z}^T\mathbf{Z} = \boldsymbol{X}^T\boldsymbol{X} + \lambda_3\boldsymbol{D}^T\boldsymbol{D}$, $\mathbf{r}_k = \mathbf{Z}^{-T}\boldsymbol{X}^T(\boldsymbol{X} - \boldsymbol{M})\boldsymbol{\alpha}_k$, and then

  $$\widehat{\boldsymbol{\beta}}_k = \arg\min_{\boldsymbol{\beta}_k} ||\mathbf{r}_k - \mathbf{Z}\boldsymbol{\beta}_k||_2^2 + \lambda_1||\boldsymbol{\beta}_k||_1,$$

  which can be easily obtained by using any LASSO solvers [45]–[47].
- Repeat Steps 2-4 until convergence.

The derivation of our algorithm is detailed in Section S-III in the supplementary material.

The time complexity of Step 1 is $O(NPT^2)+O(T^3)$, where the first term arises in the calculation of $\boldsymbol{X}^T\boldsymbol{X}$ and the second in the eigendecomposition. Step 2 involves a time complexity $O(NPTK)$ to obtain $\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^T$, and the soft thresholding $(\cdot)_+$ can be performed parallelly to get each element of $\widehat{\boldsymbol{M}}$ individually. The calculation of $(\boldsymbol{X}-\boldsymbol{M})^T\boldsymbol{X}\boldsymbol{B}$ and its singular value decomposition in Step 3 induce $O(NPT^2) + O(T^2K)$ and $O(TK^2)$, respectively. Step 4 requires a time complexity $O(NPT^2)+O(T^3)$, $O(NPT^2)+O(T^2K)$, and $O(T^3\ K)$ in computing $\mathbf{Z}$, $\mathbf{r}_k$'s and using a LASSO solver $K$ times. To sum up, as $K$ is typically far less than $T$, the dominant time complexity of our algorithm is $\min\{(O(NPT^2), O(T^3\ K)\}$. Since the objective function value is monotonically decreasing in the coordinate descent method and it is obviously lower-bounded by zero, our algorithm will converge to a stationary point. We have achieved an acceptable running time when using this algorithm to handle the massive Hong Kong MTR dataset (see Section IV-A).

## IV. Results

In this section, we apply the SRMFPCA proposed in Section III to the Hong Kong MTR dataset described in Section II. We first demonstrate the superiority of our method in terms of reconstruction errors and passenger flow patterns. The results of the SRMFPCA are further used to cluster stations, analyze correlation and identify outliers.

### A. Reconstruction Errors

To use the SRMFPCA in Eq. (11), we first center the daily passenger flow profiles at each station by subtracting their corresponding mean functions, i.e., $\mathbf{x}_{ij} = \mathbf{y}_{ij} - \widehat{\boldsymbol{\mu}}_j$. To suppress the influence of outliers on the mean function estimation, we use the sample median instead of the sample average statistics, i.e., $\widehat{\mu}_j^{(l)}(t) = \text{Median}\left(y_{1j}^{(l)}(t), \ldots, y_{Nj}^{(l)}(t)\right)$, $t = 1, \ldots, T$, $l = 1, 2$, $j = 1, \ldots, P$. We then apply the eigendecomposition to $\boldsymbol{X}^T\boldsymbol{X}$. The scree plot of the eigenvalues is given in Fig. 4(a), where we observe an elbow point at $k = 9$

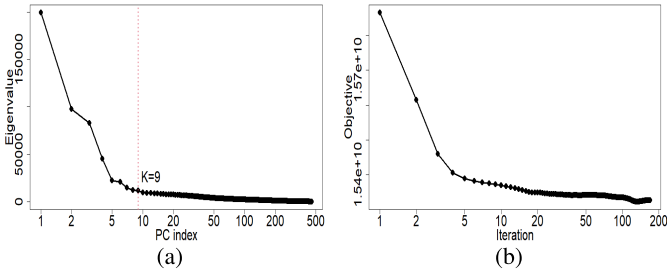



Fig. 4. (a) Scree plot. (b) Convergence analysis of our algorithm.

since the eigenvalues after that seem to level off without sharp drops anymore. Therefore, we determine the number of PCs as $K = 9$.

We shall now compare the proposed SRMFPCA with many benchmark methods regarding reconstruction errors. If a method yields an eigenvector matrix $\widehat{\boldsymbol{V}}$, then $\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}} - \boldsymbol{X}_i \widehat{\boldsymbol{V}}\, \widehat{\boldsymbol{V}}^T$ is the reconstruction errors for the $i$th sample, $i = 1, \ldots, N$. The benchmarks include: (1) the regular PCA, (2) the standard MFPCA which only produces smooth eigenvectors and considers no outliers [24]–[28], (3) the RPCA which takes the minimum covariance determination to select normal samples to compute the covariance matrix [37]–[40], and (4) a sequential version of our SRMFPCA (SSRMFPCA) which first applies the RPCA and then zeros small elements and smoothes large elements in the derived eigenvectors.

We randomly divide the whole dataset into 20 parts, and select each part as the test dataset and the remaining ones as the associated training dataset. For one training dataset, when the tuning parameters are given, our algorithm converges very quickly as shown in Fig. 4(b) and 20 iterations are already enough. The search of the optimal tuning parameters $\lambda_1$ and $\lambda_3$ in a $10 \times 10$ grid consumes 0.74 hours when programmed by Python 3.7 in a personal computer with 1.60-GHz i5-10210U CPUs and 8GB RAM. After the eigenvectors are estimated, we use the normalized mean squared error (NMSE) defined as below to measure the overall reconstruction errors:

$$\text{NMSE} = \frac{1}{N'} \sum_{i=1}^{N'} \left( ||\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}} - \boldsymbol{X}_i \widehat{\boldsymbol{V}}\ \widehat{\boldsymbol{V}}^T||_F^2 / ||\boldsymbol{Y}_i||_F^2 \right).$$

It is called the in-sample NMSE if $\boldsymbol{Y}_i$ is from the training dataset and $N' = 19N/20$, and the out-sample NMSE if $\boldsymbol{Y}_i$ is from the test one and $N' = N/20$. We repeat the above procedure for each of the 20 training-test datasets, and list the mean and standard deviation of the 20 NMSEs obtained by different methods in Table II.

Some observations from Table II deserve our comments.

- The reconstruction errors, both in-sample and out-sample, are reduced from the PCA to the MFPCA when the continuity or smoothness property of the daily passenger flow profiles is utilized.
- The RPCA does not perform well due to the complicated and diverse outliers as shown in Fig. 1(c). Some profiles with only partial abnormal segments might be mistaken as normal samples in the RPCA.

TABLE II
NMSEs OF COMPETING METHODS ($\times 10^{-2}$)

| | | PCA | MFPCA | RPCA | SSRMFPCA | SRMFPCA |
|---|---|---|---|---|---|---|
| In-sample | Mean | 1.8853 | 1.6784 | 1.9738 | 2.4588 | 1.4038 |
| | Std | 0.0042 | 0.0056 | 0.0215 | 0.0227 | 0.0024 |
| Out-sample | Mean | 1.8892 | 1.6829 | 1.9915 | 2.4804 | 1.4053 |
| | Std | 0.0850 | 0.1146 | 0.4372 | 0.4615 | 0.0480 |

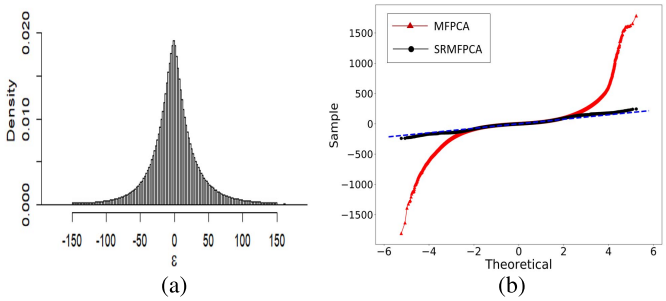



Fig. 5. Distribution of error term $\epsilon$: (a) histogram and (b) normal probability (Q-Q) plot.

- Though conceptually simple, the SSRMFPCA which removes outliers and refines eigenvectors in a step-by-step manner can not produce an optimal result.
- By contrast, our SRMFPCA, which integrates the benefits of smoothness, sparsity and robustness together using a unified optimization model in Eq. (11), behaves best with the smallest reconstruction errors.
- The closeness of the out-sample and in-sample NMSEs of our SRMFPCA also indicates the stableness of its generalization performance.

After completing the in-sample and out-sample comparison, we finally retrain our SRMFPCA, making use of the whole dataset, to better estimate the eigenvectors as a basis for further analysis in the following subsections. We plot the histogram of noises or errors $\epsilon$'s in Eq. (2) to check our model assumption (see Fig. 5(a)). The error term is unimodal-distributed with mean $-0.05$ ($\approx 0$) and variance $38.07^2$. The normal probability (Q-Q) plot is exhibited in Fig. 5(b), where we can see that our error term is more normally-distributed with less outliers than that from the conventional MFPCA. We also group these error terms by different stations and times, and visualize their standard deviation, i.e., $\widehat{\sigma}$'s, in Fig. 6 (left panel). Compared to the conventional MFPCA whose $\widehat{\sigma}$'s are also visualized in Fig. 6 (right panel), our SRMFPCA generates much smaller $\widehat{\sigma}$'s which are at the almost same order of magnitude. The histograms of these $\widehat{\sigma}$'s from the two methods are also given in Fig. 7. The $\widehat{\sigma}$ from our SRMFPCA has a shorter and lighter right tail, and its maximum value is 109.64, much smaller than that from the MFPCA (i.e., 391.65). To sum up, since we identify the outliers explicitly in SRMFPCA, our assumption on the homogeneous variances over stations and times without the effects of outliers can be more easily satisfied to lead to a credible estimation results in the MTR data analysis.

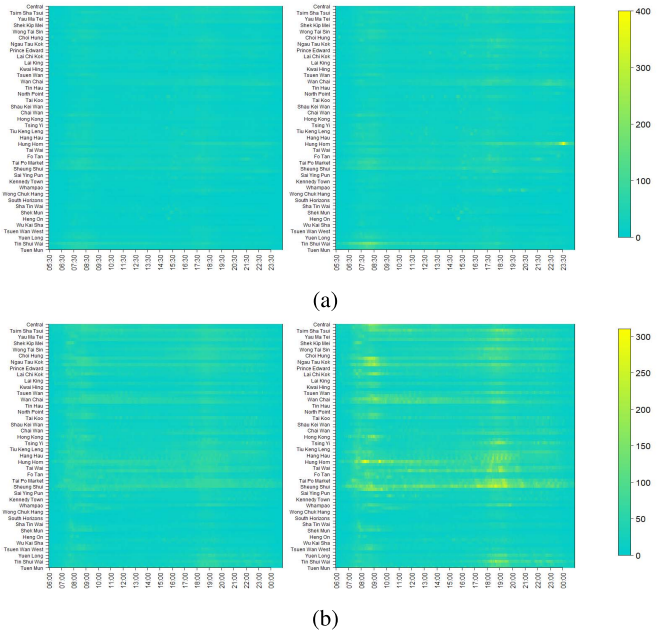


Fig. 6. Magnitude of error term $\epsilon$: $\widehat{\sigma}$ over different stations and times in (a) passenger inflow profiles and (b) passenger outflow profiles. Each panel contains two heatmaps from our SRMFPCA (left) and the standard MFPCA (right), and the same color scale is used for fair comparison.

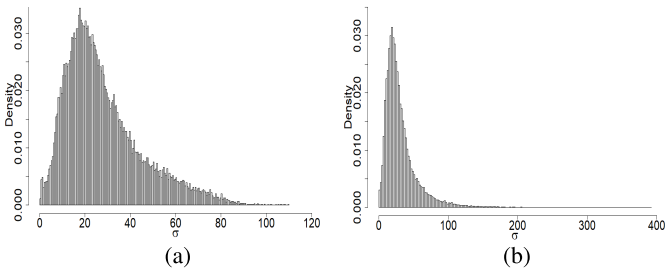


Fig. 7. Histograms of $\widehat{\sigma}$ obtained by (a) SRMFPCA and (b) MFPCA.

## B. Passenger Flow Patterns

We here demonstrate the advantages of our SRMFPCA in producing interpretable and robust passenger flow patterns. The first two eigenvectors by different methods are shown in Fig. 8. It can be clearly seen that our eigenvectors are the most clean and concise ones as they are both sparse and smooth in the inflow and outflow time domain. Compared to those from the conventional PCA, the eigenvectors from the standard MFPCA are more smooth, but they are still dense without filtering small elements and neglect the effects of outliers. Due to the existing diverse outliers in our passenger flow profiles, the RPCA and SSRMFPCA seem to be quite unstable with pretty strong fluctuations. Moreover, the sequential manner of the SSRMFPCA can only lead to a local optimum.

We now interpret our derived eigenvectors in terms of the passenger flow patterns. Note that our eigenvectors are ranked by their abilities in explaining the total data variation, which is defined as $1-||\boldsymbol{X}-\boldsymbol{X}\widehat{\beta}_k\widehat{\alpha}_k^T-\widehat{\boldsymbol{M}}||_F^2/||\boldsymbol{X}||_F^2$, for $k=1,\ldots,K$. In the first eigenvector in Fig. 8(a), which explains 16.5% of the total data variation, one outflow peak occurs during 18:00–21:00 while one inflow peak happens later at time 21:30–23:30. This means passengers exit stations in the late

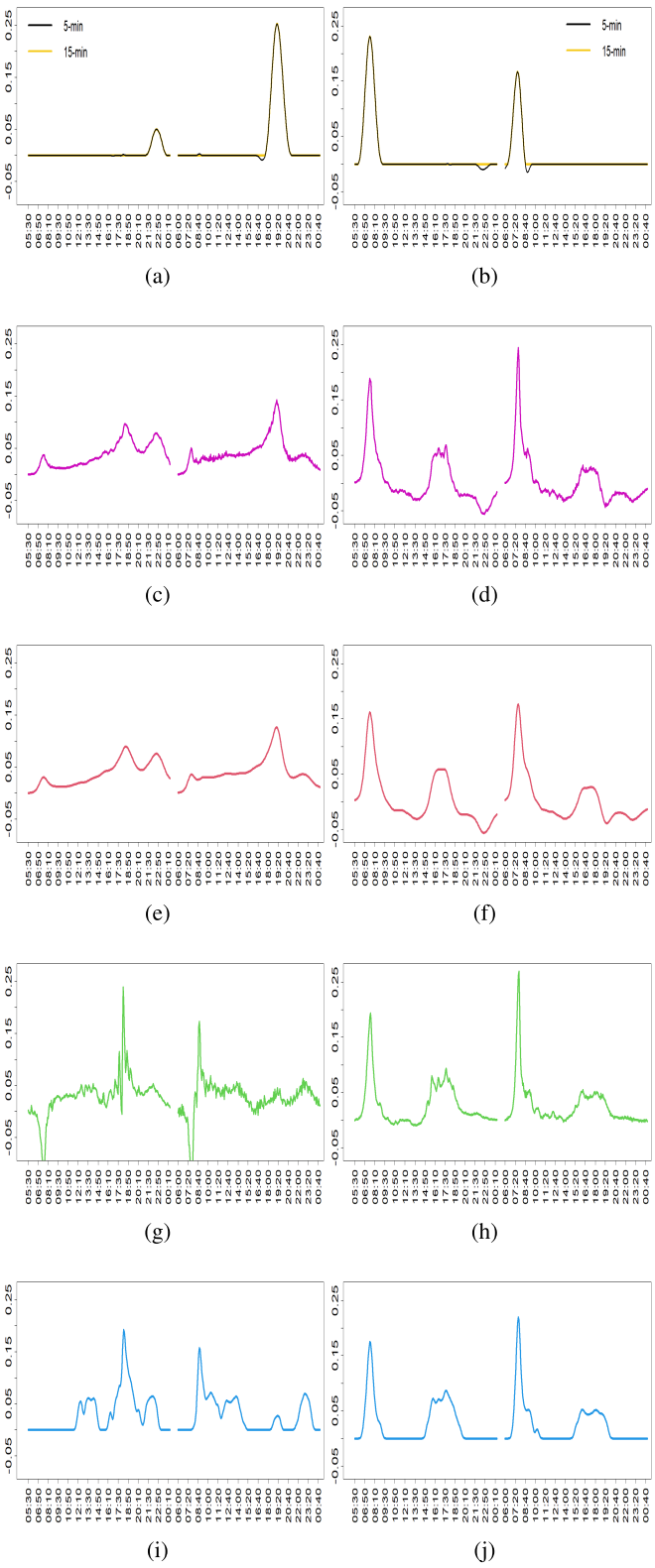
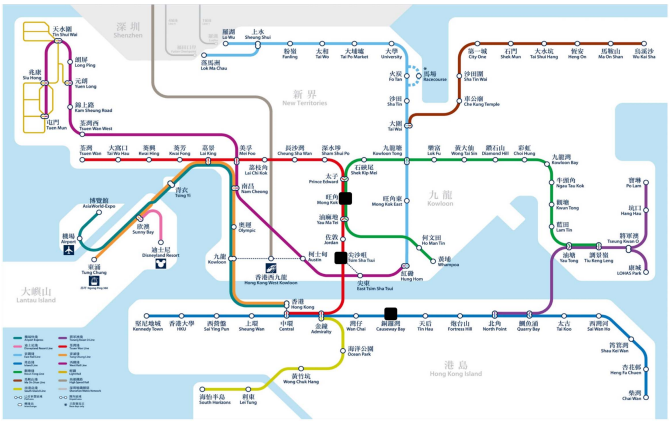
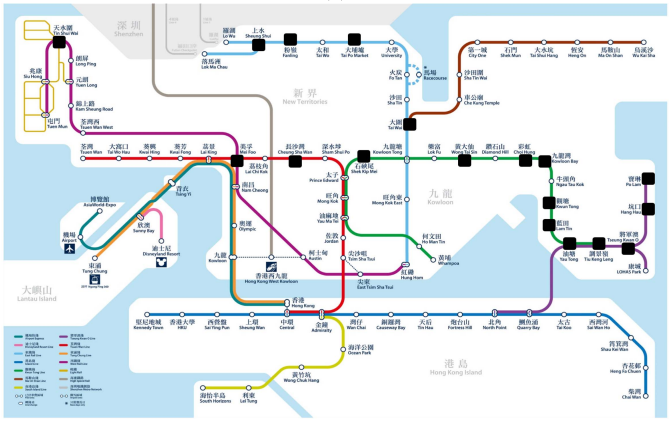


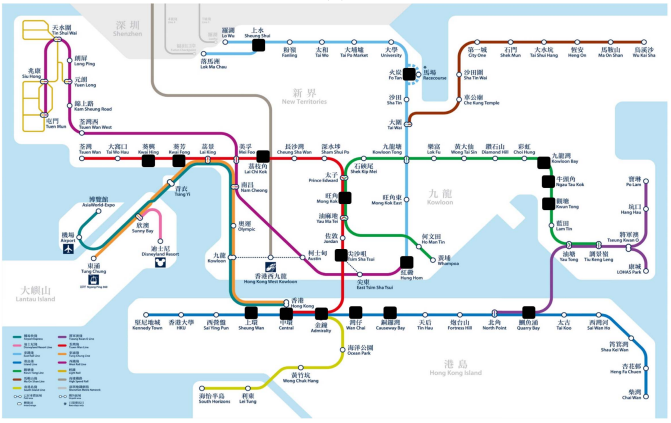
Fig. 8. Estimated eigenvectors. Rows 1 to 5 correspond to the SRMFPCA, PCA, MFPCA, RPCA and SSRMFPCA, and Columns 1 to 2 refer to the first to the second eigenvector. Each panel includes two curves representing the inflow part (left) and the outflow part (right) of an eigenvector.



(a)



(b)



(c)

Fig. 9. Station clusters with higher PC scores on the (a) first eigenvector (b) second eigenvector and (c) third eigenvector.

afternoon and early evening, and then enter stations in the late evening. The reason for such a passenger flow pattern might be that passengers visit some places for dining, shopping or entertainment and then leave those places.

To justify the above conjecture, the PC scores of all stations on the first eigenvector are calculated, upon which all stations are separated into two groups using the hierarchical clustering method. The group with a larger magnitude of PC scores includes the Mong Kok, Tsim Sha Tsui and Causeway Bay stations as marked in Fig. 9(a). To show the difference between the above three stations (cluster one) and the other ones (cluster two), we retrieve the passenger outflow data points or the check-out ridership of these two clusters during 18:00–21:00 and compare their boxplots in Fig. 10(a). The boxplots of the passenger inflows during 21:30–23:30 are compared in Fig. 10(b). We see that cluster one attracts much more ridership during these two time periods. It is worth mentioning that the three stations are all near the busiest commercial places in Hong Kong with many restaurants, malls and recreation facilities around, which validates our interpretation of the first passenger flow pattern and indicates the popularity of evening activities in these commercial places for Hong Kong citizens.

TABLE III

NMSEs OF SRMFPCA USING DIFFERENT TIME INTERVALS ($\times 10^{-2}$)

| | | 5-min | 15-min | | | 5-min | 15-min |
|---|---|---|---|---|---|---|---|
| In-sample | Mean | 1.4038 | 1.4037 | Out-sample | Mean | 1.4053 | 1.4055 |
| | Std | 0.0024 | 0.0023 | | Std | 0.0480 | 0.0481 |

Similarly, we can investigate the practical meanings of the other derived eigenvectors. In the second eigenvector in Fig. 8(b), passengers go to and leave the metro stations in the early morning. This pattern indicates that passengers might leave their homes for business districts, and is verified in Fig 9(b) by the marked station group with higher PC scores that covers the stations close to both residential and working areas in Hong Kong. The second eigenvector captures 8.10% of the total data variation. The remaining seven eigenvectors and their interpretation are available in Section S-IV in the supplementary material. For example, the third eigenvector in Fig. S-1(a) represents a passenger inflow peak in the late afternoon and early evening, and the identified station group in Fig. 9(c) with higher PC scores includes stations in the working places in Hong Kong. As such, this pattern implies that passengers finish their daily work and leave their workplaces.

As a side note, we have explored the sensitivity of the discovered passenger flow patterns to the use of different time intervals. Figs. 8(a)-(b) also exhibit the derived eigenvectors when the time interval is shifted to 15 minutes. It can be seen that the results from the 15-minute time interval are just slightly more smooth than those from the 5-minute one, but they are overlapped in most parts with the same intra-day temporal trend. The reconstruction errors of our SRMFPCA defined in Section IV-A using these two time intervals are listed in Table III, where both the in-sample and out-sample errors have no significant differences. Therefore, the influence of time interval is not notable in our MTR data analysis and the 5-minute one is proper here since this interval already covers many passenger counts due to the overall massive MTR ridership in a single day.

### C. Station Clustering

The derived eigenvectors can be utilized to cluster the MTR stations. Unlike the binary clustering in Section IV-B using

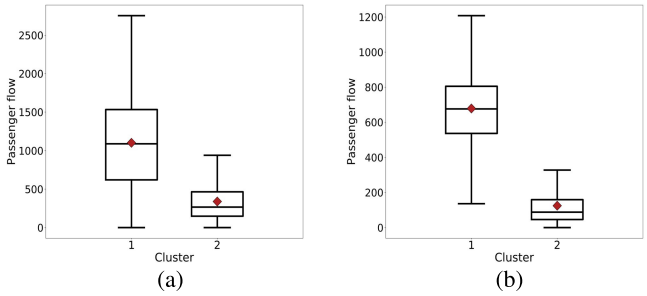



Fig. 10. Cluster comparison: (a) check-out ridership boxplots, (b) check-in ridership boxplots.

TABLE IV
CLUSTERING RESULTS OF DIFFERENT METHODS

| | Original | PCA | MFPCA | RPCA | SSRMFPCA | SRMFPCA |
|---|---|---|---|---|---|---|
| SW | 0.1521 | 0.2013 | 0.2184 | 0.1717 | 0.2287 | 0.2363 |
| No. of clusters | 5 | 11 | 11 | 13 | 8 | 12 |

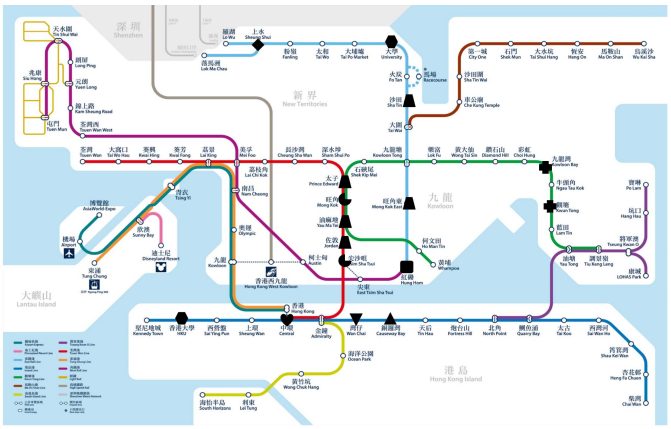


Fig. 11. Nine station clusters in the MTR system: different symbols represent different clusters.

only a single eigenvector, here we take the PC scores of all $K = 9$ eigenvectors. Since different eigenvectors represent different passenger flow patterns, the use of $K = 9$ eigenvectors actually considers the overall behaviors of each station. This also incorporates all significant signals in the passenger flow dataset to discriminate stations and discards noises as the data dimension is greatly reduced. Specifically, for the $j$th station, we use all of its PC scores $\{s_{ij,k}\}_{k=1,i=1}^{K,N}$ as its features, and in the hierarchical clustering, we employ the Ward's linkage that measures the total within-cluster sum of squares when merging two clusters. The optimal number of clusters is determined by the average silhouette width criterion [48], which has been justified as appropriate for traffic data clustering in [5], [49], [50]. Note that the silhouette width for the $j$th station is

$$\mathrm{sw}(j) = \frac{b(j) - a(j)}{\max\{a(j), b(j)\}},$$

where $a(j)$ is the mean distance between the $j$th station and all other stations in the same cluster, and $b(j)$ is the mean distance between the $j$th station and the stations in its nearest cluster. The average silhouette width is $\mathrm{SW} = \sum_{j=1}^{P} \mathrm{sw}(j)/P$.

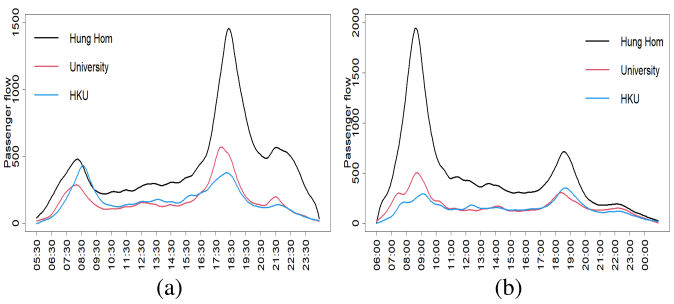



Fig. 12. Station comparison: (a) average of the daily passenger inflow profiles, (b) average of the daily passenger outflow profiles.

TABLE V
NUMBER OF CLUSTERS OVER NUMBER OF EIGENVECTORS

| No. of Eigenvectors | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| No. of clusters | 2 | 3 | 4 | 4 | 10 | 11 | 12 | 13 | 12 |

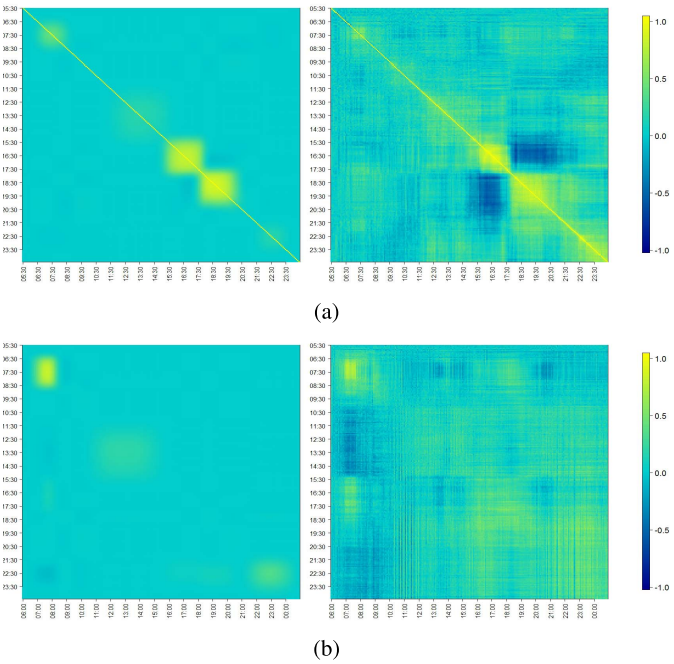



Fig. 13. (a) Within-profile correlation in the passenger inflow profile at the Admiralty station (x-axis: inflow time domain; y-axis: inflow time domain). (b) Across-profile correlation between the passenger outflow profile at the Admiralty station and the passenger inflow profile at the Choi Hung station (x-axis: outflow time domain; y-axis: inflow time domain). Each panel contains two heatmaps from our SRMFPCA (left) and the conventional method (right).

Clearly, a larger SW implies a better separation of different clusters, and the optimal number of clusters is decided as the one with the maximum average silhouette width.

The clustering results of different methods have been tabulated in Table IV. We can first see that compared to the clustering that directly relies on the original passenger flow profile data, all PCA-based methods, performing a data dimension reduction before clustering, gets better SW values. In addition, the largest SW is achieved by our SRMFPCA,

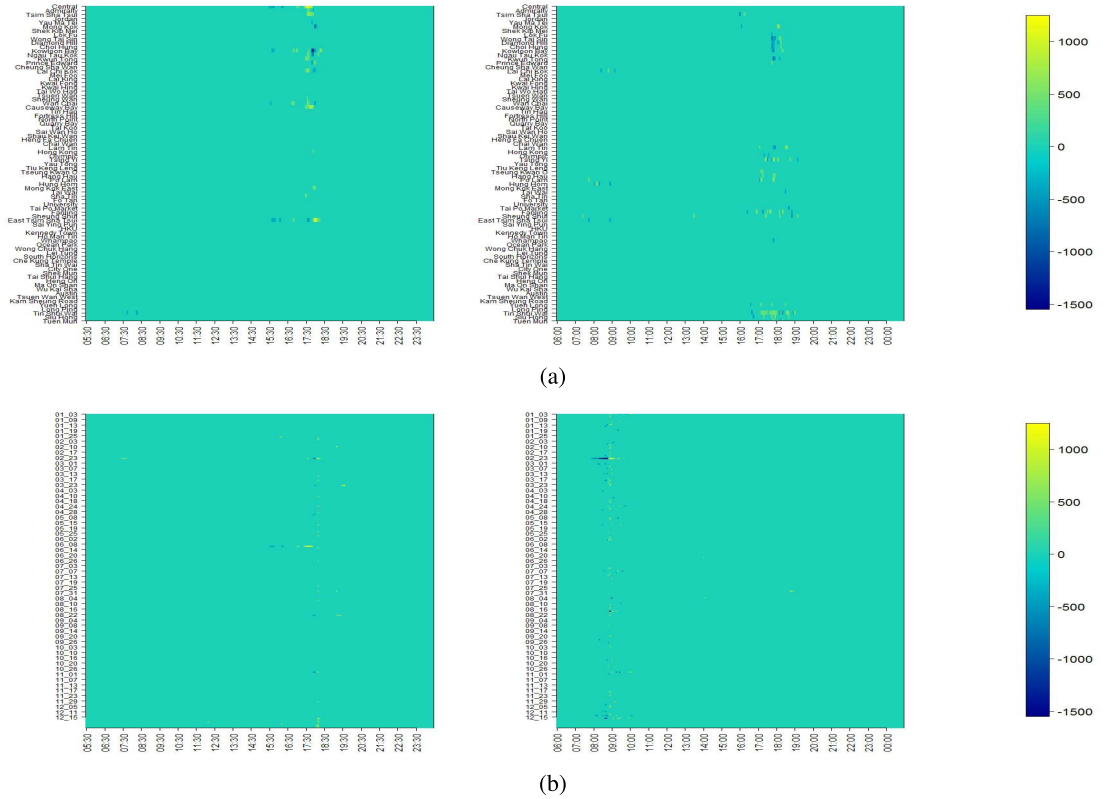



Fig. 14. (a) Passenger flow outliers on June 12, 2017. (b) Passenger flow outliers at the Central station. Each figure contains two heatmaps: the left one is the passenger inflow part while the right one the outflow part.

which indicates that our interpretable and robust eigenvectors can lead to a better station discrimination.

Our SRMFPCA generates 12 station clusters. Nine clusters are highlighted with different symbols in Fig. 11 and more information is provided in the supplementary material. Several key clusters containing only one station (e.g., Central, Causeway Bay, Hung Hom, Sheung Shui in Fig. 11) are identified, which could better help practitioners focus on the critical few stations. Additionally, our clustering is also more reasonable as, e.g., it correctly merges the two stations (University and HKU) that are near universities into one cluster. Though the Hung Hum station also connects the Hong Kong Polytechnic University, its passenger inflow and outflow profiles are quite different from those of the University and HKU stations (see Fig. 12). The Hung Hom station has much more ridership, especially in the late-afternoon and evening inflow peak and the early-morning outflow peak as it is also near many other public facilities such as Hong Kong Coliseum, long-distance train stations to mainland China, Tsim Sha Tsui Promenade, etc.

Finally, the sensitivity of the number of eigenvectors and clusters is investigated in Table V. A general trend is that as more eigenvectors are used, more clusters will be obtained. It is not surprising since more aspects of the daily passenger flow profiles will be incorporated when using more eigenvectors, which could facilitate a further station discrimination. The number of stations becomes kind of stable when a sufficient number of eigenvectors ($K \geq 7$) are utilized.

### *D. Correlation Analysis*

As illustrated in Section III-A, using the MFPCA, we can compute both the within-profile and the across-profile covariance. These covariances can be further standardized to yield the correlations. We first explore the within-profile correlation. An example is given in Fig. 13(a), where the correlations of passenger inflows at different time points at the Admiralty station are visualized in two heatmaps. The left heatmap is obtained by applying Eq. (3) to the results of our SRMFPCA, while the right one by performing the correlation calculation directly on the original noisy passenger flow data. It can be seen that our SRMFPCA can filter out noises significantly and thus makes the correlation hotspots stand out clearly. As the Admiralty station is located at business districts in Hong Kong, major inflow correlation occurs in the afternoon in the left heatmap of Fig. 13(a), which is within our expectation as passengers leave their workplaces during this time period.

Next, we investigate the across-profile correlation by considering one more station, Choi Hung, which is close to residential areas in Hong Kong. We calculate the correlation between the passenger outflow profile at the Admiralty station and the passenger inflow profile at the Choi Hung station. Our heatmap at the left part of Fig. 13(b) is again better at isolating a few critical hotspots than the conventional one at the right part of Fig. 13(b). We recognize a high correlation block in the early morning, which makes sense as more passengers leaving from their homes (near the Choi Hung station) could produce

more outflows at the working places (near the Admiralty station).

### E. Outlier Identification

As a byproduct of our proposed SRMFPCA, we can easily identify the outliers in the Hong Kong MTR dataset by looking at the nonzero elements in the estimated outlier matrix $\widehat{\boldsymbol{M}}$. The outliers can be viewed from different perspectives to gain different insights. When a day is given, we can seek the passenger inflow and outflow outliers at different stations in that day. For example, some abnormal stations such as Central, Admiralty, East Tsim Sha Tsui and Tin Shui Wai can be found on June 12, 2017 in Fig. 14(a). When focusing on a particular station, we can uncover the passenger flow anomalies across different days in a year. An example is given in Fig. 14(b) on the Central station, where we can see that as a station within the commercial and office areas in Hong Kong, the Central station suffers from unstable passenger inflows at around 18:00 and fluctuating passenger outflows at around 9:00. Such discovered knowledges about the passenger flow abnormality are quite useful for guiding the resources allocation in the MTR's daily operation.

## V. Conclusion

The large-scale passenger transit transaction data collected in modern metro systems contain extremely rich information for passenger flow analytics. Utilizing the functional property of daily passenger flow profiles, this paper proposes a novel MFPCA as a holistic method to systematically investigate the underlying passenger flow patterns in an entire metro system. To enhance the interpretability of discovered patterns and tackle the presence of diverse outliers in profile data, we incorporate a composite penalty function considering both sparsity and robustness in estimating the eigenvectors. The developed SRMFPCA is successfully applied to the Hong Kong MTR dataset, generating smaller reconstruction errors and producing sparse and smooth eigenvectors of clear interpretations. The results of our method can effectively support station clustering, correlation analysis and outlier identification.

Our method is also subject to a few limitations. First, our passenger flow pattern discovery and station clustering are conducted step by step. A more advanced model that can cluster stations and recognize cluster-wise patters simultaneously might perform better and thus needs more research attention. Second, a probabilistic version of our SRMFPCA that can deal with missing data is another promising improvement. The outliers in the real dataset could also exhibit certain spatial and temporal characteristics, such as lying at several adjacent stations and sequential time points, which requires a more tailored $L_1$-norm penalty term for $\boldsymbol{M}$ in Eq. (11). In addition, as passenger transfer activities can not be recorded by the AFC devices, our SRMFPCA based only on the tap-in and tap-out records can not provide passenger transfer patterns. An inference model of the passengers' transfer choices from their travel time data, using a probabilistic model and integrating any possible external information (e.g., train timetables, weather, social activities, camera and GPS data) is a new and valuable research topic. Finally, PCA is limited to the linear pattern extraction. To discover the nonlinear passenger flow patterns, the kernel PCA can be taken, but its modification to utilize the smoothness property of multivariate functional data and to address the model interpretation and diverse outliers deserves our future research efforts.

## References


[1] M.-P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: A literature review," *Transp. Res. C, Emerg. Technol.*, vol. 19, no. 4, pp. 557–568, Aug. 2011.

[2] L. Zhu, F. R. Yu, Y. Wang, B. Ning, and T. Tang, "Big data analytics in intelligent transportation systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 1, pp. 383–398, Jan. 2019.

[3] *Annual Report 2017*. MTR, Hong Kong, 2017.

[4] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1624–1639, Dec. 2011.

[5] I. G. Guardiola, T. Leon, and F. Mallor, "A functional approach to monitor and recognize patterns of daily traffic profiles," *Transp. Res. B, Methodol.*, vol. 65, pp. 119–136, Jul. 2014.

[6] Y. Sun, J. Shi, and P. M. Schonfeld, "Identifying passenger flow characteristics and evaluating travel time reliability by visualizing AFC data: A case study of Shanghai metro," *Public Transp.*, vol. 8, no. 3, pp. 341–363, Dec. 2016.

[7] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*. New York, NY, USA: Springer, 2005.

[8] C. Chen, J. Chen, and J. Barry, "Diurnal pattern of transit ridership: A case study of the New York City subway system," *J. Transp. Geography*, vol. 17, no. 3, pp. 176–186, May 2009.

[9] B. Du, Y. Cui, Y. Fu, R. Zhong, and H. Xiong, "SmartTransfer: Modeling the spatiotemporal dynamics of passenger transfers for crowdedness-aware route recommendations," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 6, p. 70, 2018.

[10] W. L. Wang, S. M. Lo, and S. B. Liu, "Aggregated metro trip patterns in urban areas of Hong Kong: Evidence from automatic fare collection records," *J. Urban Planning Develop.*, vol. 141, no. 3, Sep. 2015, Art. no. 05014018.

[11] L. Li, X. Su, Y. Zhang, J. Hu, and Z. Li, "Traffic prediction, data compression, abnormal data detection and missing data imputation: An integrated study based on the decomposition of traffic time series," in *Proc. 17th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2014, pp. 282–289.

[12] L. Li, X. Su, Y. Zhang, Y. Lin, and Z. Li, "Trend modeling for traffic time series analysis: An integrated study," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3430–3439, Dec. 2015.

[13] L. Qu, J. Hu, L. Li, and Y. Zhang, "PPCA-based missing data imputation for traffic flow volume: A systematical approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 3, pp. 512–522, Sep. 2009.

[14] C. Chen, Y. Wang, L. Li, J. Hu, and Z. Zhang, "The retrieval of intra-day trend and its influence on traffic prediction," *Transp. Res. C, Emerg. Technol.*, vol. 22, pp. 103–118, Jun. 2012.

[15] L. Li, Y. Li, and Z. Li, "Efficient missing data imputing for traffic flow by considering temporal and spatial dependence," *Transp. Res. C, Emerg. Technol.*, vol. 34, pp. 108–120, Sep. 2013.

[16] S. Jiang, S. Wang, Z. Li, W. Guo, and X. Pei, "Fluctuation similarity modeling for traffic flow time series: A clustering approach," in *Proc. IEEE 18th Int. Conf. Intell. Transp. Syst.*, Sep. 2015, pp. 848–853.

[17] S. Coogan, C. Flores, and P. Varaiya, "Traffic predictive control from low-rank structure," *Transp. Res. B, Methodol.*, vol. 97, pp. 1–22, Mar. 2017.

[18] T. Ma, Z. Zhou, and C. Antoniou, "Dynamic factor model for network traffic state forecast," *Transp. Res. B, Methodol.*, vol. 118, pp. 281–317, Dec. 2018.

[19] J.-M. Chiou, "Dynamical functional prediction and classification, with application to traffic flow prediction," *Ann. Appl. Statist.*, vol. 6, no. 4, pp. 1588–1614, Dec. 2012.

[20] I. M. Wagner-Muns, I. G. Guardiola, V. A. Samaranayke, and W. I. Kayani, "A functional data analysis approach to traffic volume forecasting," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 878–888, Mar. 2018.

[21] J.-M. Chiou, Y.-C. Zhang, W.-H. Chen, and C.-W. Chang, "A functional data approach to missing value imputation and outlier detection for traffic flow data," *Transportmetrica B, Transp. Dyn.*, vol. 2, no. 2, pp. 106–129, May 2014.
[22] M. Grasso, B. M. Colosimo, and M. Pacella, "Profile monitoring via sensor fusion: The use of PCA methods for multi-channel data," *Int. J. Prod. Res.*, vol. 52, no. 20, pp. 6110–6135, Oct. 2014.
[23] X. Fang, N. Z. Gebraeel, and K. Paynabar, "Scalable prognostic models for large-scale condition monitoring applications," *IISE Trans.*, vol. 49, no. 7, pp. 698–710, Jul. 2017.
[24] J. A. Dubin and H.-G. Müller, "Dynamical correlation for multivariate longitudinal data," *J. Amer. Stat. Assoc.*, vol. 100, no. 471, pp. 872–881, Sep. 2005.
[25] K. Paynabar, C. Zou, and P. Qiu, "A change-point approach for phase-I analysis in multivariate profile monitoring and diagnosis," *Technometrics*, vol. 58, no. 2, pp. 191–204, Apr. 2016.
[26] C.-Z. Di, C. M. Crainiceanu, B. S. Caffo, and N. M. Punjabi, "Multilevel functional principal component analysis," *Ann. Appl. Statist.*, vol. 3, no. 1, p. 458, Mar. 2009.
[27] A.-M. Staicu, C. M. Crainiceanu, and R. J. Carroll, "Fast methods for spatially correlated multilevel functional data," *Biostatistics*, vol. 11, no. 2, pp. 177–194, Apr. 2010.
[28] K. Hasenstab *et al.*, "A multi-dimensional functional principal components analysis of EEG data," *Biometrics*, vol. 73, no. 3, pp. 999–1009, Sep. 2017.
[29] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
[30] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graph. Statist.*, vol. 15, no. 2, pp. 265–286, Jun. 2006.
[31] H. Shen and J. Z. Huang, "Sparse principal component analysis via regularized low rank matrix approximation," *J. Multivariate Anal.*, vol. 99, no. 6, pp. 1015–1034, Jul. 2008.
[32] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–534, Jul. 2009.
[33] G. I. Allen and M. Weylandt, "Sparse and functional principal components analysis," 2013, *arXiv:1309.2895*. [Online]. Available: http://arxiv.org/abs/1309.2895
[34] K. Chen and J. Lei, "Localized functional principal component analysis," *J. Amer. Stat. Assoc.*, vol. 110, no. 511, pp. 1266–1275, Jul. 2015.
[35] C. Zhang, H. Yan, S. Lee, and J. Shi, "Weakly correlated profile monitoring based on sparse multi-channel functional principal component analysis," *IISE Trans.*, vol. 50, no. 10, pp. 878–891, Oct. 2018.
[36] K. Wang and F. Tsung, "Hierarchical sparse functional principal component analysis for multistage multivariate profile data," *IISE Trans.*, vol. 53, no. 1, pp. 58–73, Jan. 2021.
[37] P. J. Rousseeuw and K. V. Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, no. 3, pp. 212–223, Aug. 1999.
[38] M. Hubert, P. J. Rousseeuw, and K. V. Branden, "ROBPCA: A new approach to robust principal component analysis," *Technometrics*, vol. 47, no. 1, pp. 64–79, Feb. 2005.
[39] R. Maronna, "Principal components and orthogonal regression based on robust scales," *Technometrics*, vol. 47, no. 3, pp. 264–273, Aug. 2005.
[40] A. M. Zoubir, V. Koivunen, Y. Chakhchoukh, and M. Muma, "Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts," *IEEE Signal Process. Mag.*, vol. 29, no. 4, pp. 61–80, Jul. 2012.
[41] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis," *J. ACM*, vol. 58, no. 3, p. 11, 2011.
[42] X. Xing, X. Zhou, H. Hong, W. Huang, K. Bian, and K. Xie, "Traffic flow decomposition and prediction based on robust principal component analysis," in *Proc. IEEE 18th Int. Conf. Intell. Transp. Syst.*, Sep. 2015, pp. 2219–2224.
[43] X. Wang, Y. Zhang, H. Liu, Y. Wang, L. Wang, and B. Yin, "An improved robust principal component analysis model for anomalies detection of subway passenger flow," *J. Adv. Transp.*, vol. 2018, pp. 1–12, Aug. 2018.
[44] J. S. Morris, "Functional regression," *Annu. Rev. Statist. Appl.*, vol. 2, pp. 321–359, Apr. 2015.
[45] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, "Pathwise coordinate optimization," *Ann. Appl. Statist.*, vol. 1, no. 2, pp. 302–332, Dec. 2007.
[46] R. J. Tibshirani and J. Taylor, "The solution path of the generalized lasso," *Ann. Statist.*, vol. 39, no. 3, pp. 1335–1371, Jun. 2011.
[47] T. B. Arnold and R. J. Tibshirani, "Efficient implementations of the generalized lasso dual path algorithm," *J. Comput. Graph. Statist.*, vol. 25, no. 1, pp. 1–27, Jan. 2016.
[48] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987.
[49] M. Yildirimoglu and N. Geroliminis, "Experienced travel time prediction for congested freeways," *Transp. Res. B, Methodol.*, vol. 53, pp. 45–63, Jul. 2013.
[50] J. Zhao, Q. Qu, F. Zhang, C. Xu, and S. Liu, "Spatio-temporal analysis of passenger travel patterns in massive smart card data," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 11, pp. 3135–3146, Nov. 2017.

**Kai Wang** received the B.S. degree in industrial engineering from Xi'an Jiaotong University, Xi'an, China, in 2014, and the Ph.D. degree in industrial engineering and logistics management from The Hong Kong University of Science and Technology, Hong Kong, in 2018.

He is currently an Assistant Professor with the School of Management, Xi'an Jiaotong University. His research interests include statistical process control, industrial big data analytics, statistical machine learning, and transfer learning.

**Fugee Tsung** received the B.Sc. degree from National Taiwan University, Taipei, Taiwan, and the M.Sc. and Ph.D. degrees from the University of Michigan, Ann Arbor, MI, USA.

He is currently a Chair Professor with the Department of Industrial Engineering and Decision Analytics and the Director of the Quality and Data Analytics Laboratory, The Hong Kong University of Science and Technology, Hong Kong. His research interests include quality analytics in advanced manufacturing and service processes, industrial big data, and statistical process control, monitoring, and diagnosis.