

Dimension Reduction Methods for Graph Signals

Kyusoon Kim

Department of Statistics and Actuarial Science
Soongsil University

2026 NUU – SSU G-LAMP International Workshop

October 3, 2026

1. Introduction
2. Preliminaries
3. PCA in the Graph Frequency Domain
4. Graph Frequency-Domain Factor Modeling
5. Concluding Remarks

1. Introduction
2. Preliminaries
3. PCA in the Graph Frequency Domain
4. Graph Frequency-Domain Factor Modeling
5. Concluding Remarks

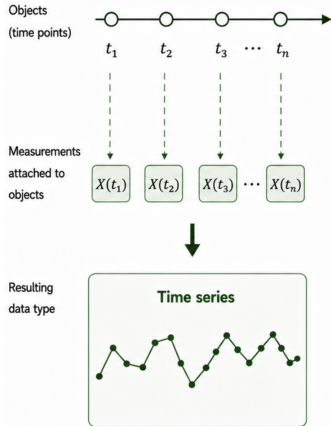
A Unified View of Statistical Data

Primary object of interest	Object / index domain	Attached values	Data type
Objects themselves	Points in $\mathbb{R}^p$	—	Euclidean data
	Functions in $L^2(\mathcal{T})$	—	Functional data
	Points on S^d	—	Spherical data
	Points on a manifold $\mathcal{M}$	—	Manifold-valued data
	Points in Kendall's shape space	—	Shape data
	Point configurations in a spatial domain $\mathcal{D} \subset \mathbb{R}^d$	—	Spatial point patterns
	Graphs $\mathcal{G} = (\mathcal{V}, \mathcal{E})$	—	Network data
Measurements on structured domains	Time domain $t \in \mathcal{T}$	$X(t) \in \mathbb{R}^p$	Time series
	Spatial domain $s \in \mathcal{D} \subset \mathbb{R}^d$	$X(s) \in \mathbb{R}^p$	Spatial data
	Graph node domain $v \in \mathcal{V}$	$X(v) \in \mathbb{R}^p$	Graph signals

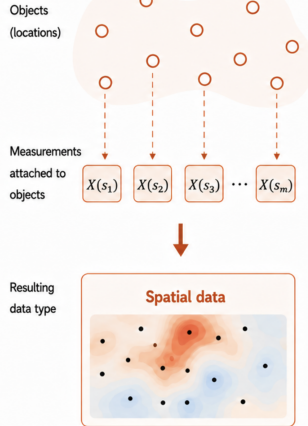
— indicates that the object itself is the observation.

Relationships Among Observations

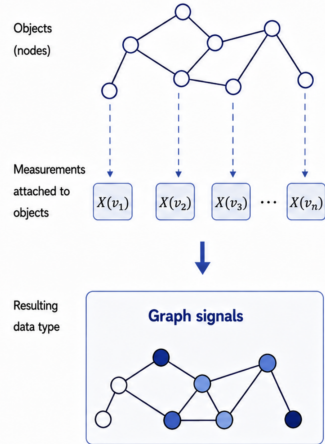
1. Temporal relationships



2. Spatial relationships



3. Graph-encoded relationships



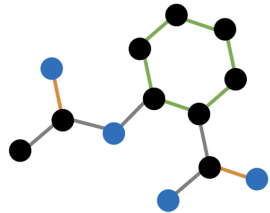
Examples of Graphs



Logical networks
(social networks, web pages)



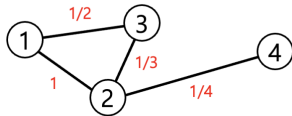
Physical networks
(metro systems, sensor networks)



Biological and chemical networks
(molecular graphs, proteins)

Notation

- $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$: a weighted graph
 - $\mathcal{V}$: a set of n nodes
 - $\mathcal{E}$: a set of edges
 - W : weighted adjacency matrix
 - The nonnegative element w_{ij} represents the **strength of the connection** between nodes i and j .
- We consider only graphs that are finite, connected, and simple, that is, graphs with no loops or multiple edges.



$\mathcal{G}$

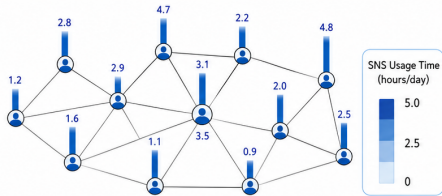
$$\begin{array}{c} \begin{array}{cccc} & 1 & 2 & 3 & 4 \end{array} \\ \begin{array}{l} 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{pmatrix} 0 & 1 & 1/2 & 0 \\ 1 & 0 & 1/3 & 1/4 \\ 1/2 & 1/3 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \end{pmatrix} \end{array}$$

W

- **Graph signals** refer to measurements observed at the nodes of a graph.
- The analysis of graph signals considers both the observed values themselves and the relational structure on which they are observed.

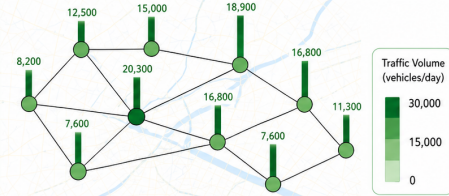
Examples of Graph Signals

1. Social Network (SNS Usage Time)



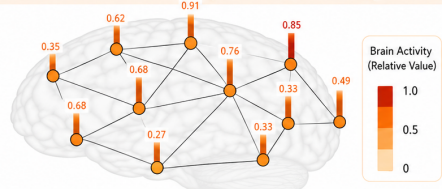
Bar height = SNS usage time (hours/day)

2. Urban Transportation Network (Traffic Volume)



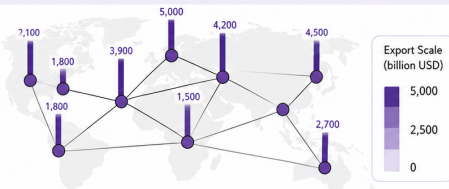
Bar height = Traffic volume observed on roads/sections (vehicles/day)

3. Brain Functional Connectivity Network (Brain Activity)



Bar height = Brain activity level in brain regions (relative value)

4. Global Trade Network (Export Scale)



Bar height = Export scale of countries (billion USD)

- **Random graph processes** refer to stochastic processes indexed by the nodes of a graph $\mathcal{G}$.

$$X : \Omega \times \mathcal{V} \rightarrow \mathbb{R} \text{ (or } \mathbb{C})$$

or equivalently,

$$\{X(v) : v \in \mathcal{V}\}, \quad X(v) : \Omega \rightarrow \mathbb{R} \text{ (or } \mathbb{C})$$

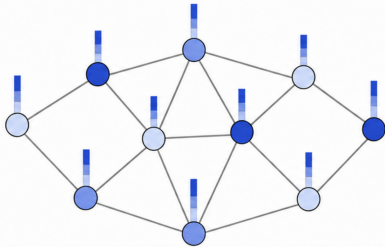
- A realization of a random graph process is a graph signal.

$$X = (X(1), \dots, X(n))^{\top} \in \mathbb{R}^n \xrightarrow{\text{realization}} x = (x(1), \dots, x(n))^{\top} \in \mathbb{R}^n$$

From Univariate to Multivariate Graph Signals

Univariate Graph Signal

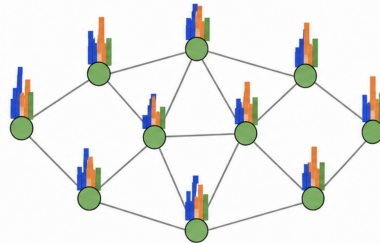
One variable is observed at each node



Example: PM2.5 concentration (univariate)
Bar height = value of the observation

Multivariate Graph Signal

Multiple variables are observed at each node



Example: ■ PM2.5 ■ Temperature ■ Humidity ...
Number of bars = number of variables observed

- **Multivariate graph process** refer to vector-valued stochastic processes indexed by the nodes of a graph $\mathcal{G}$.

$$X : \Omega \times \mathcal{V} \rightarrow \mathbb{R}^p \text{ (or } \mathbb{C}^p)$$

- We write

$$X = (X_1 | \cdots | X_p) \in \mathbb{R}^{n \times p}$$

$$j\text{th component graph process: } X_j = (X_j(1), \dots, X_j(n))^{\top} \in \mathbb{R}^n$$

- In this talk, we develop **dimension reduction methods** for multiple variables observed at graph nodes, while **incorporating the underlying graph structure**.
- To this end, we develop **PCA** and **factor model** in the graph frequency domain.

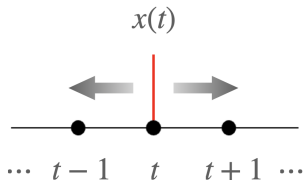
1. Introduction
2. Preliminaries
3. PCA in the Graph Frequency Domain
4. Graph Frequency-Domain Factor Modeling
5. Concluding Remarks

Times Series vs Graph Signals

Time series

Domain: **Infinite domain**

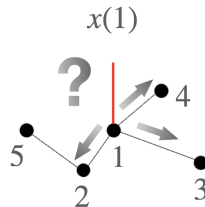
Shift direction: **Forward / Backward**



Graph signals

Domain: **Finite domain**

Shift direction: **No natural direction**

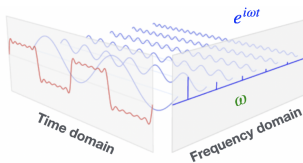


- In time series analysis, the shift operator plays a fundamental role in **spectral analysis** and **stationarity**.

$$(Bx)(t) = x(t - 1)$$

- **Spectral analysis:** Fourier modes are eigenfunctions of B .

$$Be^{i\omega t} = e^{-i\omega} e^{i\omega t}$$



- **Stationarity:** statistical properties are invariant under time shifts.

Graph Shift Operator (GSO)

- The **graph shift operator** (GSO), denoted by an $n \times n$ matrix S , defines the **notion of signal shift on a graph** (Sandryhaila and Moura, 2014; Segarra et al., 2018).

$$x = (x(1), \dots, x(5))^T$$

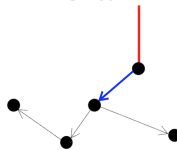
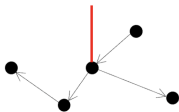
Forward

Backward

x

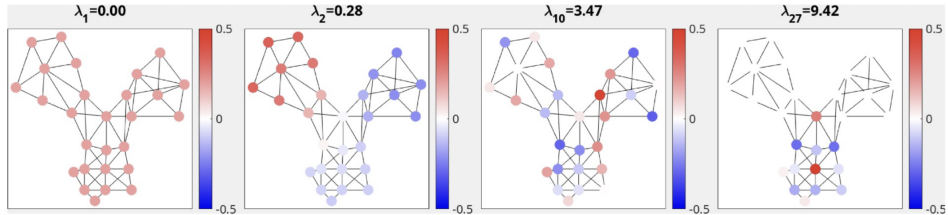
Sx

$S^T x$



Graph shift $\neq$ Node translation

- The GSO is usually assumed to be normal, i.e., $SS^H = S^H S$.
 - $S = V\Lambda V^H$, where Λ is the eigenvalue matrix, and V is a unitary eigenvector matrix
 - λ_ℓ : **graph frequency** (low: smooth variation / high: rapid variation)
 - v_ℓ : **basis signal**, representing a characteristic pattern of variation over the graph nodes.
- For undirected graphs, the graph Laplacian is the most common example of a GSO.



- The classical Fourier transform is defined as

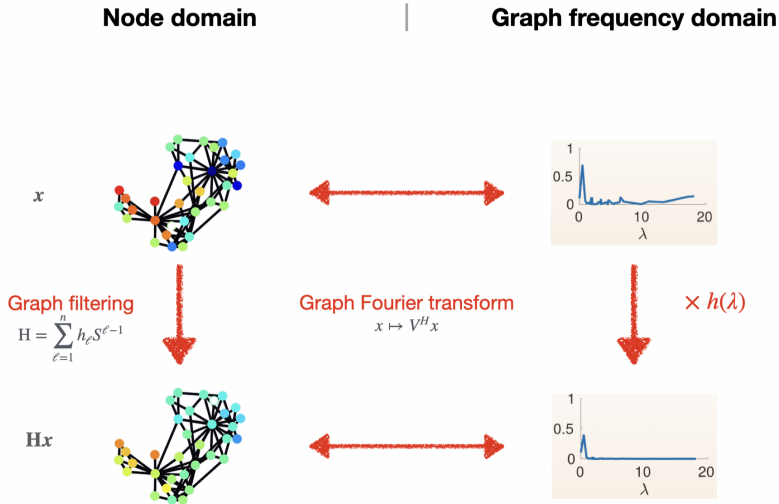
$$\tilde{f}(\omega) = \langle f(t), e^{i\omega t} \rangle,$$

where the complex exponentials are the **eigenfunctions of the shift operator**, corresponding to the **eigenvalues** $e^{-i\omega}$.

- The **graph Fourier transform (GFT)** of a graph signal $\mathbf{x}$ is analogously defined as the inner product between the **eigenvectors of the GSO** and the signal:

$$\tilde{x}_\ell = \tilde{x}(\lambda_\ell) = \langle \mathbf{x}, \mathbf{v}_\ell \rangle = \mathbf{v}_\ell^H \mathbf{x}, \quad 1 \leq \ell \leq n.$$

- $S = V\Lambda V^H = (\mathbf{v}_1 \mid \cdots \mid \mathbf{v}_n) \text{diag}(\{\lambda_\ell\}_{\ell=1}^n)(\mathbf{v}_1 \mid \cdots \mid \mathbf{v}_n)^H$.
- $\tilde{x}_\ell$: GFT coefficient corresponding to the eigenvalue λ_ℓ .



Definition (Stationarity)

Given a normal GSO S , a zero-mean random graph process X is **stationary with respect to S** if one of the following holds (Marques et al., 2017; Segarra et al., 2018):

- For any set of non-negative integers a , b , and $c \leq b$, it holds that

$$\mathbb{E}\left[(S^a X)((S^H)^b X)^H\right] = \mathbb{E}\left[(S^{a+c} X)((S^H)^{b-c} X)^H\right]$$

cf. Weak stationarity for time series: $\mathbb{E}\left[X(t+a)X(t-b)^H\right] = \mathbb{E}\left[X(t+(a+c))X(t-(b-c))^H\right]$

- The covariance matrix $\Sigma_X = \mathbb{E}\left[XX^H\right]$ and S are simultaneously diagonalizable.

We define the stationarity of a multivariate random graph process.

Definition (Stationarity in multivariate graph processes)

Given a normal graph shift operator S , consider a zero-mean p -dimensional random graph process $X = (X_1 \mid \cdots \mid X_p)$, where each X_i ($1 \leq i \leq p$) is an n -dimensional vector. The graph process X is said to be weakly stationary if one of the following holds:

- The cross-covariance matrices $\Sigma_{ij}^X := \text{Cov}(X_i, X_j)$ and S are simultaneously diagonalizable for all $1 \leq i, j \leq p$.

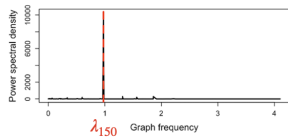
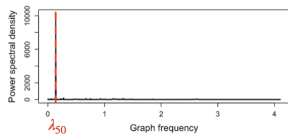
We define a **graph cross-spectral density (GCSD)**.

Definition (Graph cross-spectral density)

The **graph cross-spectral density (GCSD)** of zero-mean jointly weakly stationary graph processes X and Y is defined as an $n \times 1$ vector $p_{XY} = \text{diag}(V^H \Sigma_{XY} V)$.

- The GCSD, $p_{XY} = (p_{XY}(\lambda_1), \dots, p_{XY}(\lambda_n))^T$, is a function of graph frequency.
- It describes the **graph frequencies shared between two graph-structured variables** and provides information about the correlation of the two signals in the graph frequency domain.
- If $X = Y$, the GCSD reduces to the **graph power spectral density (GPSD)** as defined by Marques et al. (2017), which indicates the **energy distribution over graph frequencies**.

GPSD



GCSD

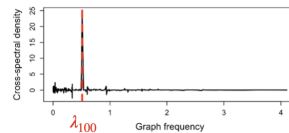


Figure 2: GPSD vs GCSD

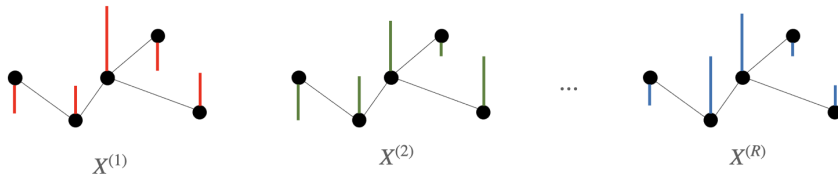
Estimation of Spectral Densities

- We consider R independent realizations $X^{(1)}, \dots, X^{(R)}$.
- The graph cross-spectral density is estimated by

$$\hat{p}_{ij}^X = \text{diag}(V^H \widehat{\text{Cov}}(X_i, X_j) V),$$

which is called the **graph cross-periodogram**.

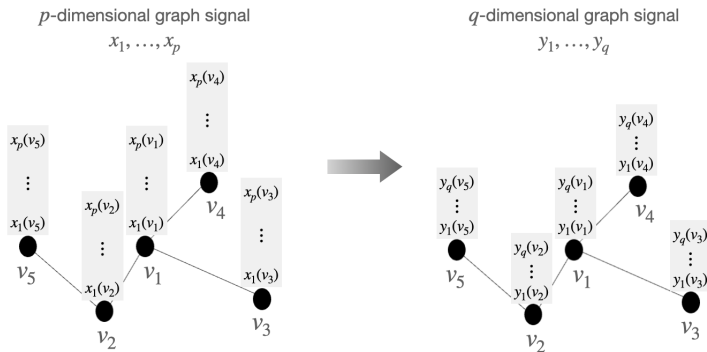
- When only a few realizations are available, graph windowing provides additional windowed signals for estimation.



1. Introduction
2. Preliminaries
3. PCA in the Graph Frequency Domain
4. Graph Frequency-Domain Factor Modeling
5. Concluding Remarks

Dimension Reduction of Multivariate Graph Signals

- We aim to reduce the dimension of data on graphs from p to q ($< p$).
- It is essential to **consider the inherent graph structures** during dimension reduction.



Motivating Example: G20 Economic Data

- Overlooking trade-based connections between countries during dimension reduction can **obscure significant shared trends across countries**.

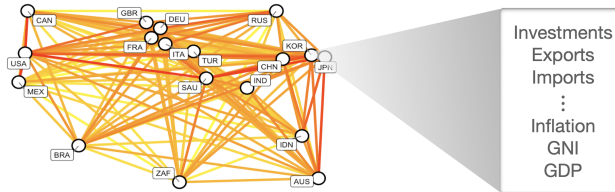
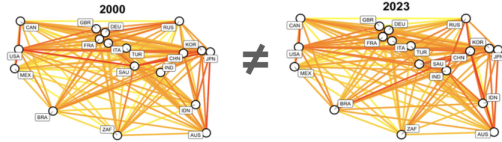
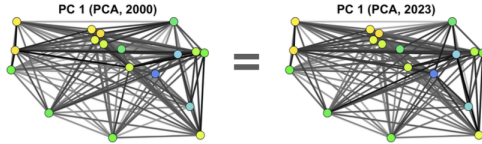


Figure 3: Multiple economic and growth indicators for G20 countries.

G20 trading network



Classical PCA



Proposed method

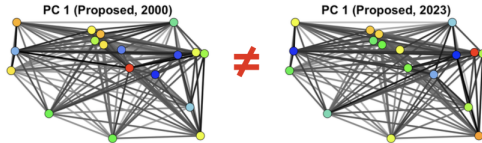
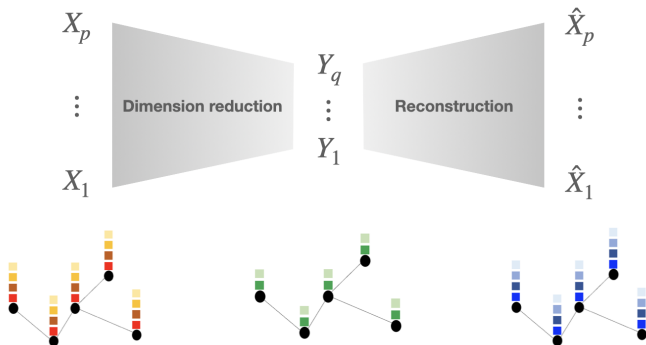


Figure 4: First principal component results from classical PCA and the proposed method under the 2000 and 2023 G20 trading networks.

Proposed Method: gFreqPCA

- Let $X = (X_1 \mid \cdots \mid X_p)$ denote a zero-mean p -dimensional stationary graph process with respect to the GSO S .
- We seek to construct q principal components $Y_1, \dots, Y_q$.



- Each principal component Y_i is defined as a **sum of graph-filtered outputs of the signal**

$$Y_i = H_{i1}X_1 + \cdots + H_{ip}X_p, \quad 1 \leq i \leq q,$$

where H_{ij} are graph filters.

- We reconstruct the graph signal as

$$\hat{X}_i = \mu_i + G_{i1}Y_1 + \cdots + G_{iq}Y_q,$$

where G_{ij} are graph filters.

- **We need to estimate the filters H_{ij} and G_{ij} .**

Theorem

The graph signal $\hat{X}_i$, which minimizes the mean squared error $\sum_{i=1}^p E[(X_i - \hat{X}_i)^H(X_i - \hat{X}_i)]$, is given by

$$\hat{X}_i = \hat{\mu}_i + \hat{A}_{i1}X_1 + \cdots + \hat{A}_{ip}X_p, \quad 1 \leq i \leq p,$$

where $\hat{A}_{ij} = V \text{diag}(\{\hat{a}_{ij}(\lambda_\ell)\}_{\ell=1}^n) V^H$,

$$\hat{\mu}_i = \mu_i^X - \sum_{j=1}^p \hat{G}_{ij} \left(\sum_{k=1}^p \hat{H}_{jk} \mu_k^X \right),$$

$$\hat{A}(\lambda_\ell) = \hat{G}(\lambda_\ell) \hat{H}(\lambda_\ell),$$

$$\hat{G}(\lambda_\ell) = (u_1(\lambda_\ell) | \cdots | u_q(\lambda_\ell)) = (\hat{H}(\lambda_\ell))^H.$$

Here, $\hat{H}_{ij} = V \text{diag}(\{\hat{h}_{ij}(\lambda_\ell)\}_{\ell=1}^n) V^H$ and $\hat{G}_{ij} = V \text{diag}(\{\hat{g}_{ij}(\lambda_\ell)\}_{\ell=1}^n) V^H$. Moreover, $\hat{A}(\lambda_\ell) = (\hat{a}_{ij}(\lambda_\ell))_{p \times p}$, $\hat{H}(\lambda_\ell) = (\hat{h}_{ij}(\lambda_\ell))_{q \times p}$, and $\hat{G}(\lambda_\ell) = (\hat{g}_{ij}(\lambda_\ell))_{p \times q}$.

The (minimum mean squared) reconstruction error is obtained as $\sum_{\ell=1}^n \sum_{i>q} \tau_i(\lambda_\ell)$.

Why Call It PCA in the Graph Frequency Domain?

- Obtaining the column vectors of $\hat{G}(\lambda_\ell)$ corresponds to sequentially determining vectors $\beta(\lambda_\ell)$ at each λ_ℓ such that

$$\max_{\beta(\lambda_\ell) \neq \mathbf{0}} \frac{\beta(\lambda_\ell)^H P_X(\lambda_\ell) \beta(\lambda_\ell)}{\beta(\lambda_\ell)^H \beta(\lambda_\ell)}$$

$$\text{Classical PCA: } \max_{\beta \neq \mathbf{0}} \frac{\beta^\top \text{Cov}(X) \beta}{\beta^\top \beta}$$

- $P_X(\lambda_\ell) := \left(p_{ij}^X(\lambda_\ell) \right)_{p \times p} = U(\lambda_\ell) T(\lambda_\ell) U(\lambda_\ell)^H$: the spectral matrix of X at graph frequency λ_ℓ .
- It is equal to the covariance matrix of the p -dimensional graph Fourier coefficient vector $(\tilde{X}_1(\lambda_\ell), \dots, \tilde{X}_p(\lambda_\ell))^\top$.

- In this sense, the proposed method performs classical PCA separately at each graph frequency.

- We can interpret each principal component at each graph frequency λ_ℓ :

$$\tilde{Y}_i(\lambda_\ell) = u_i(\lambda_\ell)^H (\tilde{X}_1(\lambda_\ell), \dots, \tilde{X}_p(\lambda_\ell))^T, \quad \ell = 1, \dots, n.$$

- **Low frequency:** loading vector $u_i(\lambda_\ell)$ highlights which variables contribute to the **coarse-scale variation in the component**.
- **High frequency:** loading vector $u_i(\lambda_\ell)$ highlights which variables drive **fine-scale fluctuations**.

Theorem

The following signal pairs exhibit zero graph coherence at all graph frequencies:

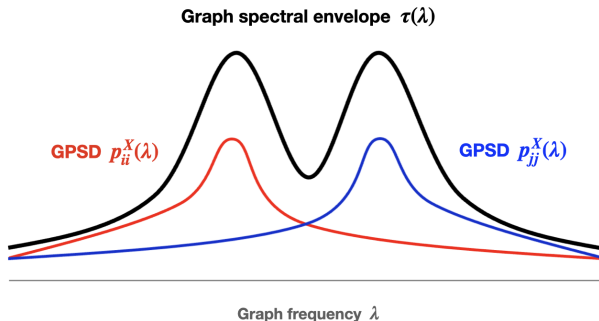
- principal components Y_i and Y_j ($i \neq j$)
- principal component Y_i and error signal $\epsilon_j = X_j - \hat{X}_j$
- ϵ_j and $\hat{X}_k$

Graph Spectral Envelope

- The graph spectral envelope is defined as

$$\tau(\lambda_\ell) := \max_{\beta(\lambda_\ell)} \frac{\beta(\lambda_\ell)^H P_X(\lambda_\ell) \beta(\lambda_\ell)}{\beta(\lambda_\ell)^H \beta(\lambda_\ell)}, \quad \ell = 1, \dots, n$$

- It bounds GPSDs of $X_1, \dots, X_p$, making it a useful tool for detecting common graph frequencies.



- The optimal graph frequency scaling is defined as

$$u(\lambda_\ell) := \arg \max_{\beta(\lambda_\ell)} \frac{\beta(\lambda_\ell)^H P_X(\lambda_\ell) \beta(\lambda_\ell)}{\beta(\lambda_\ell)^H \beta(\lambda_\ell)}, \quad \ell = 1, \dots, n$$

- The optimal graph frequency scaling can be used to estimate the normalized amplitudes of the basis signals.

- For the US sensor network, we generate a 12-dimensional stationary graph signal as follows:

$$X_i = c_{i,1}Z_1v_{50}^{\text{US}} + c_{i,2}Z_2v_{100}^{\text{US}} + c_{i,3}Z_3v_{150}^{\text{US}} + \delta_i, \quad i = 1, \dots, 11$$

$$X_{12} = 2\delta_{12}.$$



Figure 5: US sensor network.

Graph Periodograms for the Generated Signals

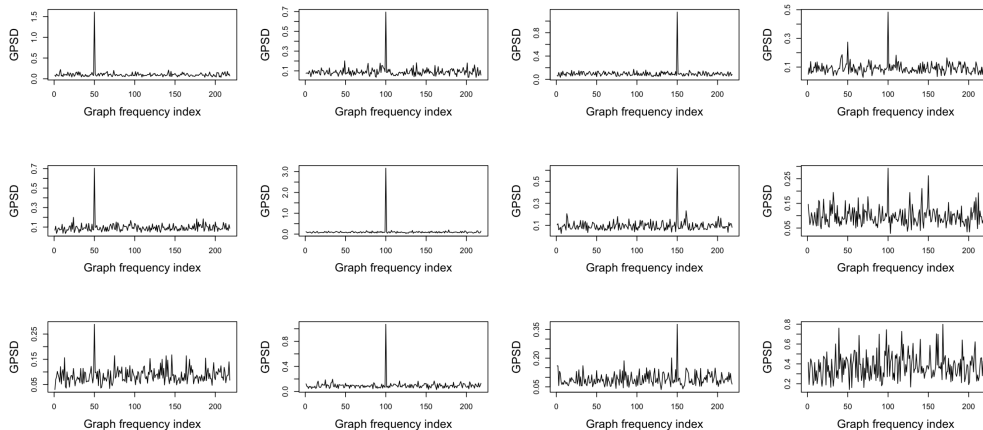


Figure 6: Windowed average graph periodograms of signal X_i .

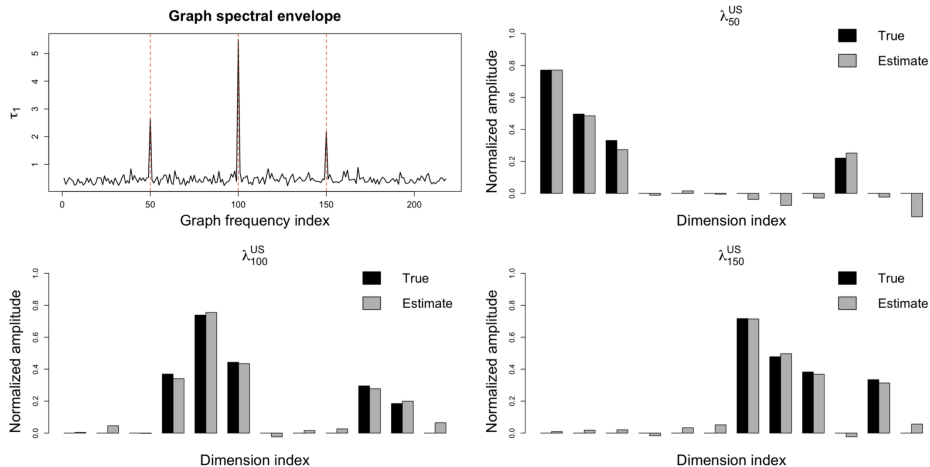


Figure 7: Graph spectral envelope and the optimal graph frequency scalings at different frequencies.

1. Introduction
2. Preliminaries
3. PCA in the Graph Frequency Domain
4. Graph Frequency-Domain Factor Modeling
5. Concluding Remarks

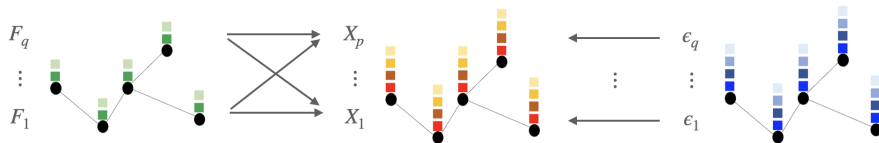
Factor Model in the Graph Frequency Domain

- Let $X = (X_1 \mid \cdots \mid X_p)$ denote a zero-mean p -dimensional stationary graph-structured data with respect to the GSO S .
- We introduce the following factor model on graphs using graph filters:

$$X_i = B_{i1}F_1 + \cdots + B_{iq}F_q + \epsilon_i, \quad 1 \leq i \leq p,$$

where $B_{ij} = V \text{diag}(\{b_{ij}(\lambda_\ell)\}_{\ell=1}^n) V^H$ are graph filters.

- $F = (F_1 \mid \cdots \mid F_q)$: a zero-mean, q -dimensional stationary **factor**.
- $\epsilon = (\epsilon_1 \mid \cdots \mid \epsilon_p)$: a zero-mean, p -dimensional stationary **error**.
- $\text{Cov}(\epsilon_i, F_j) = O$ for $1 \leq i \leq p, 1 \leq j \leq q$.
- $\text{Cov}(F_k, F_m) \neq O$ and $\text{Cov}(\epsilon_i, \epsilon_j) \neq O$ for $1 \leq k, m \leq q$ and $1 \leq i, j \leq p$.



- We can rewrite the equation in matrix form as

$$P_X(\lambda_\ell) = B(\lambda_\ell)P_F(\lambda_\ell)B(\lambda_\ell)^H + P_\epsilon(\lambda_\ell), \quad \ell = 1, \dots, n.$$

- $P_X(\lambda_\ell) = (p_{ij}^X(\lambda_\ell))_{p \times p}, \quad P_F(\lambda_\ell) = (p_{ij}^F(\lambda_\ell))_{q \times q}, \quad P_\epsilon(\lambda_\ell) = (p_{ij}^\epsilon(\lambda_\ell))_{p \times p}$
- $B(\lambda_\ell) = (b_{ij}(\lambda_\ell))_{p \times q}$

- Consequently, we can separate the model into smaller models for each graph frequency level ℓ .

$$\tilde{X}_\ell = B(\lambda_\ell)\tilde{F}_\ell + \tilde{\epsilon}_\ell, \quad \ell = 1, \dots, n,$$

where $\tilde{X}_\ell = (\tilde{X}_{\ell 1}, \dots, \tilde{X}_{\ell p})^\top$, $\tilde{F}_\ell = (\tilde{F}_{\ell 1}, \dots, \tilde{F}_{\ell q})^\top$, and $\tilde{\epsilon}_\ell = (\tilde{\epsilon}_{\ell 1}, \dots, \tilde{\epsilon}_{\ell p})^\top$ are graph Fourier coefficient vectors at graph frequency λ_ℓ .

- Each model is analogous to the classical factor model or static factor model.

- To estimate the factors and factor loadings, we aim to **minimize the mean squared error** $\mathbb{E}[\|\tilde{X}_\ell - B(\lambda_\ell)\tilde{F}_\ell\|_2^2] = \text{tr}[P_X(\lambda_\ell) - B(\lambda_\ell)P_F(\lambda_\ell)B(\lambda_\ell)^H]$ for each ℓ .
- However, since **$P_X(\lambda_\ell)$ is unknown**, we need to estimate it. This can be achieved by utilizing either the graph cross-periodogram or the windowed average graph cross-periodogram.
- We adopt the graph cross-periodogram, **assuming that we have R i.i.d. copies $X^{(1)}, \dots, X^{(R)}$ of the random graph process X** , to develop the theoretical asymptotic results for consistent estimation.
- Denote the vectors of $\tilde{X}_\ell$, $\tilde{F}_\ell$, and $\tilde{\epsilon}_\ell$ corresponding to $X^{(r)}$ by $\tilde{X}_\ell^{(r)}$, $\tilde{F}_\ell^{(r)}$, and $\tilde{\epsilon}_\ell^{(r)}$, respectively, for $1 \leq r \leq R$, which results in $\tilde{X}_\ell^{(r)} = B(\lambda_\ell)\tilde{F}_\ell^{(r)} + \tilde{\epsilon}_\ell^{(r)}$ for $1 \leq \ell \leq n$.

- For $\mathcal{X}_{\ell}^{\top} = (X_{\ell}^{(1)} \mid \cdots \mid X_{\ell}^{(R)})$, $\mathcal{F}_{\ell}^{\top} = (F_{\ell}^{(1)} \mid \cdots \mid F_{\ell}^{(R)})$, and $\mathcal{E}_{\ell}^{\top} = (\epsilon_{\ell}^{(1)} \mid \cdots \mid \epsilon_{\ell}^{(R)})$,

$$\mathcal{X}_{\ell}^{\top} = B(\lambda_{\ell})\mathcal{F}_{\ell}^{\top} + \mathcal{E}_{\ell}^{\top}, \quad \ell = 1, \dots, n.$$

- $\mathcal{X}_{\ell}^{\top} \mathcal{X}_{\ell}^{*} / R = \frac{1}{R} \sum_{r=1}^R X_{\ell}^{(r)} (X_{\ell}^{(r)})^H = \hat{P}_X(\lambda_{\ell})$ estimates $P_X(\lambda_{\ell})$.
- We estimate the factor loadings and latent factors using ordinary least squares (OLS).

- In addition to the factors and factor loadings, the **true number of factors, q , is also unknown**.
- When we use k factors, the factors and the factor loadings are represented by $\mathcal{F}_{\sim \ell}^k$ and $B^k(\lambda_{\ell})$, respectively.
- Denote by $(\bar{\mathcal{F}}_{\sim \ell}^k)^{\top} = (\bar{F}_{\sim \ell}^{(1),k} \mid \cdots \mid \bar{F}_{\sim \ell}^{(R),k})$ and $(\check{\mathcal{F}}_{\sim \ell}^k)^{\top} = (\check{F}_{\sim \ell}^{(1),k} \mid \cdots \mid \check{F}_{\sim \ell}^{(R),k})$.
- Referring to Bai and Ng (2002), the following criteria can be used to determine the number of factors:

$$PC(k) = \Theta(k, \hat{G}^k) + kg(p, R),$$

$$IC(k) = \log(\Theta(k, \hat{G}^k)) + kg(p, R),$$

where $g(p, R)$ is a penalty function, $\hat{G}^k$ indicates the factor estimator, and

$$\Theta(k, \mathcal{F}_{\sim \ell}^k) = \frac{1}{pR} \min_{B^k(\lambda_{\ell})} \|\mathcal{X}_{\sim \ell}^{\top} - B^k(\lambda_{\ell})(\mathcal{F}_{\sim \ell}^k)^{\top}\|_F^2.$$

Assumption 1 (Factors and factor loadings)

- (a) $B(\lambda_\ell)^H B(\lambda_\ell)/p \rightarrow D$ as $p \rightarrow \infty$ for $1 \leq \ell \leq n$, where D is a $q \times q$ positive definite matrix.
- (b) $|b_{ij}(\lambda_\ell)| \leq \bar{b}(\lambda_\ell) < \infty$ for $1 \leq \ell \leq n$.
- (c) $\frac{1}{R} \sum_{r=1}^R \tilde{F}_\ell^{(r)} (\tilde{F}_\ell^{(r)})^H \xrightarrow{p} P_F(\lambda_\ell)$ as $R \rightarrow \infty$, and $P_F(\lambda_\ell)$ is positive definite for $1 \leq \ell \leq n$.
- (d) $\mathbb{E}[\|\tilde{F}_\ell^{(r)}\|^4] < \infty$ for $1 \leq \ell \leq n$ and $1 \leq r \leq R$.

Assumption 2 (Moments of the error)

- (a) $\lim_{p \rightarrow \infty} \frac{1}{p} \|P_\epsilon(\lambda_\ell)\|_F^2 < \infty$ for $1 \leq \ell \leq n$, where $\|\cdot\|_F$ denotes the Frobenius norm.
- (b) $\mathbb{E}[|\epsilon_{\ell,i}^{(r)}|^4] \leq M$, for a positive constant M , where $1 \leq i \leq p$ and $1 \leq r \leq R$.
- (c) $\mathbb{E}[\frac{1}{\sqrt{p}} \sum_{i=1}^p ((\epsilon_{\ell,i}^{(r)})^* \epsilon_{\ell,i}^{(r')} - \mathbb{E}[(\epsilon_{\ell,i}^{(r)})^* \epsilon_{\ell,i}^{(r')}])^4] \leq M$, for a positive constant M , where $1 \leq r, r' \leq R$.

Theorem 1 (Consistency of factor estimators)

Let $C_{pR} = \min\{\sqrt{p}, \sqrt{R}\}$ and suppose Assumption 1(b), (c), (d), and Assumption 2 hold. We assume that k factors are estimated and q is the true number of factors, where k may satisfy $k \leq q$ or $k > q$. Then, for each $1 \leq \ell \leq n$, there exists a $q \times k$ matrix $\mathcal{H}_\ell^k$ with $\text{rank}(\mathcal{H}_\ell^k)$ such that

$$C_{pR}^2 \left(\frac{1}{R} \sum_{r=1}^R \|\check{F}_{\sim \ell}^{(r),k} - (\mathcal{H}_\ell^k)^\top F_{\sim \ell}^{(r)}\|^2 \right) = O_p(1) \quad \text{as } p, R \rightarrow \infty.$$

Theorem 2 (Consistency of factor number estimation)

Suppose Assumptions 1 and 2 hold. Let $C_{pR} = \min\{\sqrt{p}, \sqrt{R}\}$ and $\hat{k} = \arg \min_{1 \leq k \leq K} PC(k)$. If $g(p, R) \rightarrow \infty$ and $C_{pR}^2 \cdot g(p, R) \rightarrow \infty$ as $p, R \rightarrow \infty$, then $\lim_{p, R \rightarrow \infty} \Pr(\hat{k} = q) = 1$. $IC(k)$ also consistently estimates the true number of factor q .

- We have n individual models, each corresponding to a graph frequency λ_ℓ :

$$\tilde{X}_\ell = B(\lambda_\ell)\tilde{F}_\ell + \tilde{\epsilon}_\ell, \quad \ell = 1, \dots, n.$$

- The loading $b_{ij}(\lambda_\ell)$ captures the **relationship** between the i th graph-structured variable X_i and the j th factor F_j **at the graph frequency λ_ℓ** .
- Its magnitude $|b_{ij}(\lambda_\ell)|$ indicates **how much the factor contributes** at that frequency.
- Therefore, the same factor can be interpreted differently across graph frequencies:
 - **Low frequency**: contribution to **coarse-scale** patterns.
 - **High frequency**: contribution to **fine-scale** patterns.
- For complex-valued loadings, the argument captures the **directional relationship** between the factor and the variable.

① Finite sample performance

- Evaluate the accuracy of factor estimation and factor number selection across different values of p and R , and across different graph structures.

② Comparison with classical factor analysis

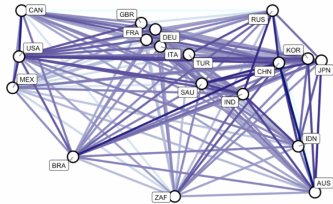
- Compare reconstruction performance to show the benefit of incorporating the underlying graph structure.

③ Effect of the graph structure

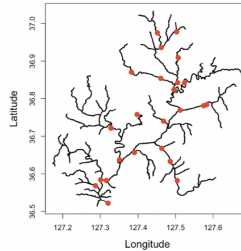
- Examine how misspecified or perturbed graph structures affect factor estimation.

④ Remarks on other graph-aware methods

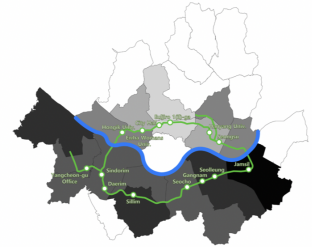
- Discuss differences from graph-based PCA and graph autoencoders in terms of reconstruction and interpretability.



G20 economic data



Miho-Cheon Catchment Data



Seoul Subway Data

Figure 8: Real data used in the study.

- **Graph:** 50 stations on Seoul Metropolitan Subway Line 2 $\rightarrow n = 50$.
- **Variables:** boarding and alighting passenger counts over 19 hourly intervals $\rightarrow p = 19$.
- **Observations:** weekly averaged data from 2019 to 2022 $\rightarrow R = 208$.
- **Preprocessing:** log transformation, $x \mapsto \log(1 + x)$, followed by centering.
- **GSO:** graph Laplacian, $S = L$.

Factor Interpretation

- We focus on the first estimated factor as an illustrative example.
- **Lowest frequency:** loadings have a **similar direction** across time intervals.
 - Common **coarse-scale** relationship.
- **Higher frequencies:** morning and evening intervals show **contrasting loading patterns**.
 - Finer-scale differences in commuting-related patterns.
- This illustrates the **frequency-wise interpretation** of factor loadings.

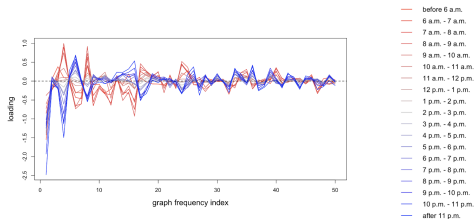


Figure 9: Loadings $b_{i1}^{\text{metro}}(\lambda_{\ell}^{\text{metro}})$ ($i = 1, \dots, 19$, $\ell = 1, \dots, 50$) for the first factor.

Two-Dimensional Loading Patterns

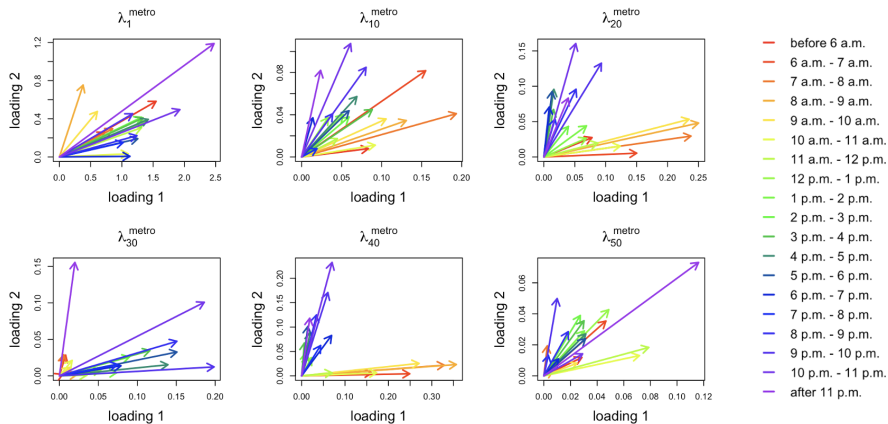


Figure 10: Two-dimensional plot of loadings for the two factors.

- We applied the k -means clustering algorithm with $k = 2$ using the first factor.
- The clusters correspond to periods before and after the outbreak of COVID-19.

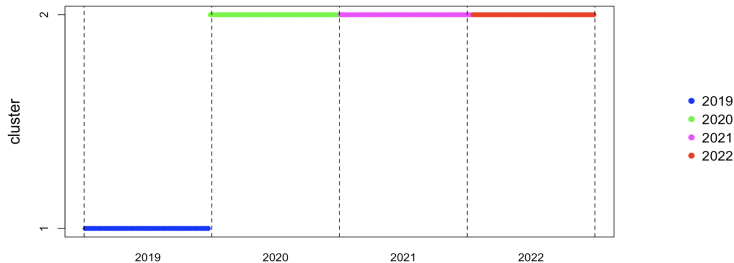


Figure 11: Clustering results using the first factor in the Seoul Metropolitan Subway Line 2 data.

1. Introduction
2. Preliminaries
3. PCA in the Graph Frequency Domain
4. Graph Frequency-Domain Factor Modeling
5. Concluding Remarks

- We introduced graph signals as data observed over interconnected objects, where both the observed values and the underlying graph structure are essential for statistical analysis.
- The irregular structure of graphs poses unique challenges, particularly in defining notions such as frequency and stationarity.
- The graph frequency domain provides a useful framework for characterizing variation over graphs at different scales.
- Among various statistical tasks for graph signals, we focused on dimension reduction, developing graph frequency-domain extensions of PCA and factor model.

- Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.
- Marques, A. G., Segarra, S., Leus, G., and Ribeiro, A. (2017). Stationary graph processes and spectral estimation. *IEEE Transactions on Signal Processing*, 65(22):5911–5926.
- Sandryhaila, A. and Moura, J. M. (2014). Discrete signal processing on graphs: Frequency analysis. *IEEE Transactions on Signal Processing*, 62(12):3042–3054.
- Segarra, S., Chepuri, S. P., Marques, A. G., and Leus, G. (2018). Statistical graph signal processing: Stationarity and spectral estimation. In *Cooperative and Graph Signal Processing*, pages 325–347. Elsevier.
- Shumway, R. and Stoffer, D. (2017). *Time Series Analysis and Its Applications: With R Examples*. Springer.

Additional Slides

- Constructive example: filtered output of a white graph process, $H\xi$.
- Diffusion-type processes: local linear dynamics $X^{(m+1)} = (I_n - \gamma^{(m)}S)X^{(m)}$ yield stationary graph processes.
- Finite periodic time series: graph stationarity on a directed cycle recovers weak stationarity.
- Interpretation: a common graph-spectral mechanism operates across the entire graph.

Statistical Replicates in Spectral Estimation

- Unlike stationarity in time or space, graph stationarity does not provide natural within-realization replication.
- Increasing the number of nodes changes the underlying graph itself.
- Nodes do not naturally act as statistical replicates.

Time

$$\text{Cov}(X(t), X(t+h)) = \gamma(h)$$

$$\underbrace{(X(1), X(1+h)), \dots, (X(T-h), X(T))}_{\text{same temporal lag } h} \longrightarrow \hat{\gamma}(h)$$

Space

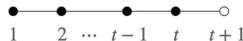
$$\text{Cov}(X(s), X(s+h)) = C(h)$$

$$\underbrace{\{(X(s_i), X(s_j)) : s_j - s_i \approx h\}}_{\text{same spatial lag } h} \longrightarrow \hat{C}(h)$$

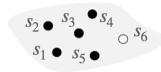
Graph

No analogous node translation

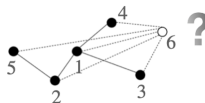
Time



Space



Graph



We briefly overview of the work by Shumway and Stoffer (2017).

- Let $Y_t = (y_{1t}, \dots, y_{pt})^\top$ denote a zero-mean p -dimensional stationary time series.
- The dynamic factor model can be expressed as

$$Y_t = \sum_{j=-\infty}^{\infty} A_j F_{t-j} + \xi_t,$$

where $\{A_j\}$ is a $p \times q$ filter and F_t is a q -dimensional factor.

- In the frequency domain,

$$f_{YY}(\omega) = A(\omega) f_{FF}(\omega) A(\omega)^H + f_{\xi\xi}(\omega),$$

where $A(\omega) = \sum_{t=-\infty}^{\infty} A_t \exp(-2\pi i t \omega)$.

- Assumption 1(a) and (b) concern the factor loadings.
 - They ensure that each factor contributes meaningfully to the spectral behavior of the observed graph-structured data, preventing trivial or degenerate contributions.
- Assumption 1(c) discusses the stability of the factors' variance and covariance structure across multiple realizations, which is necessary to guarantee consistency of the estimators.
- Assumption 1(d) is a technical condition that prevents factors from becoming excessively volatile or being dominated by extreme outliers.
- Assumption 2 allows for limited dependence among the errors and requires a finite moment, ensuring that the noise does not exhibit excessively large fluctuations.

- Weak stationarity level = $\|\text{diag}(V^H \hat{\Sigma} V)\|_2 / \|V^H \hat{\Sigma} V\|_F$.
- The closer this measure is to 1, the more weakly stationary the graph process becomes.

Table 1: Proportion of correct identification of the true factor number and MSE of rescaled estimated factors for varying p and R in the Karate club network 3-factor model. The values in each cell represent the percentage of correct identification, followed by the MSE in parentheses.

	$R = 100$	$R = 500$	$R = 1000$	$R = 2000$
$p = 100$	58.8 (705.276)	61.8 (321.274)	61.8 (227.046)	61.8 (161.191)
$p = 500$	61.8 (111.415)	64.7 (50.740)	67.6 (35.860)	67.6 (25.459)
$p = 1000$	61.8 (37.164)	67.6 (16.926)	67.6 (11.963)	67.6 (8.493)
$p = 2000$	61.8 (0.041)	67.6 (0.018)	67.6 (0.013)	70.6 (0.009)

Table 2: Averaged reconstruction error rates with standard deviations in parentheses over 100 simulations for the proposed and classical factor analysis methods according to different values of p and M on the irregular graph, given two-factor and three-factor models.

Models	Dimension	Proposed				Classical
		$M = 10$	$M = 20$	$M = 30$	$M = 40$	
Two-factor	$p = 10$	0.023 (0.003)	0.020 (0.003)	0.019 (0.002)	0.019 (0.002)	0.705 (0.019)
	$p = 20$	0.027 (0.003)	0.024 (0.003)	0.023 (0.003)	0.022 (0.002)	0.705 (0.019)
	$p = 30$	0.028 (0.004)	0.025 (0.003)	0.024 (0.003)	0.023 (0.003)	0.705 (0.019)
	$p = 40$	0.029 (0.004)	0.025 (0.003)	0.024 (0.003)	0.024 (0.003)	0.705 (0.019)
Three-factor	$p = 10$	0.008 (0.001)	0.006 (0.001)	0.006 (0.001)	0.005 (0.001)	0.881 (0.009)
	$p = 20$	0.010 (0.001)	0.008 (0.001)	0.007 (0.001)	0.007 (0.001)	0.881 (0.008)
	$p = 30$	0.010 (0.001)	0.008 (0.001)	0.008 (0.001)	0.007 (0.001)	0.881 (0.008)
	$p = 40$	0.011 (0.001)	0.009 (0.001)	0.008 (0.001)	0.008 (0.001)	0.881 (0.008)

Table 3: Average RMSE of the rescaled estimated factors, with standard deviations in parentheses, over 100 simulations for p and M , under the three-factor model on the irregular graph, when the underlying graph structure is incorrectly assumed using random graph models.

Graph model	Dimension	$M = 10$	$M = 20$	$M = 30$	$M = 40$
True	$p = 10$	942.0 (149.7)	901.3 (149.4)	890.3 (147.6)	884.5 (146.9)
	$p = 20$	891.9 (136.9)	906.8 (139.8)	916.9 (142.7)	921.2 (144.1)
	$p = 30$	906.0 (136.7)	931.4 (140.5)	945.0 (144.2)	950.8 (145.8)
	$p = 40$	918.6 (137.7)	947.6 (141.7)	962.6 (145.7)	969.0 (147.6)
Erdős–Rényi	$p = 10$	1225.5 (197.1)	1157.2 (183.0)	1126.2 (180.0)	1112.0 (174.4)
	$p = 20$	1049.2 (167.3)	1048.1 (163.6)	1042.0 (162.8)	1039.7 (161.0)
	$p = 30$	1005.3 (158.9)	1021.5 (157.5)	1022.4 (157.2)	1023.8 (156.6)
	$p = 40$	988.8 (156.1)	1012.9 (155.5)	1017.1 (156.2)	1019.8 (155.5)
Barabási–Albert	$p = 10$	1267.8 (222.0)	1206.8 (204.3)	1180.4 (198.3)	1169.6 (199.6)
	$p = 20$	1094.5 (190.3)	1096.5 (183.6)	1092.7 (179.1)	1092.7 (180.9)
	$p = 30$	1051.8 (182.2)	1072.3 (178.2)	1075.3 (175.0)	1078.9 (177.4)
	$p = 40$	1035.6 (178.3)	1064.9 (175.7)	1070.9 (173.3)	1075.6 (176.0)
Watts–Strogatz	$p = 10$	1203.2 (207.7)	1141.3 (196.2)	1115.0 (192.2)	1104.4 (186.1)
	$p = 20$	1034.8 (176.4)	1036.1 (174.4)	1029.6 (172.2)	1030.4 (169.4)
	$p = 30$	994.4 (168.6)	1013.4 (168.0)	1013.2 (165.7)	1017.1 (163.3)
	$p = 40$	979.0 (165.2)	1005.5 (165.4)	1008.4 (163.7)	1014.3 (162.1)

- Based on a scree plot, we chose the factor number as 2.

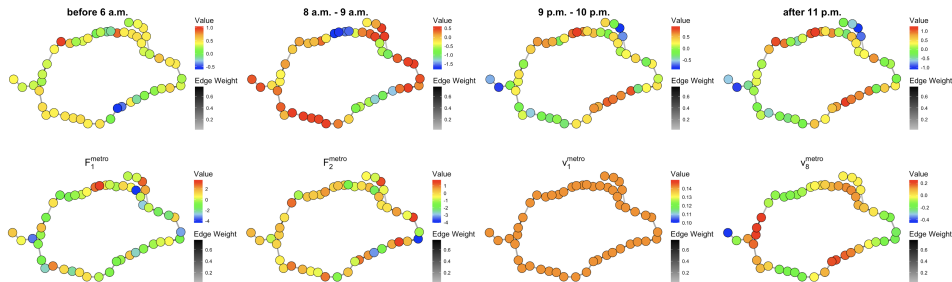


Figure 12: Signals for passenger numbers in selected time intervals and the two factor estimates for the second week of 2019.

- The second factor is associated with signals in several time intervals (before 6 a.m., from 8 a.m. to 9 a.m., from 9 p.m. to 10 p.m., and after 11 p.m.) at the lowest graph frequency, showing opposing directions as reflected by the signs of the loadings.

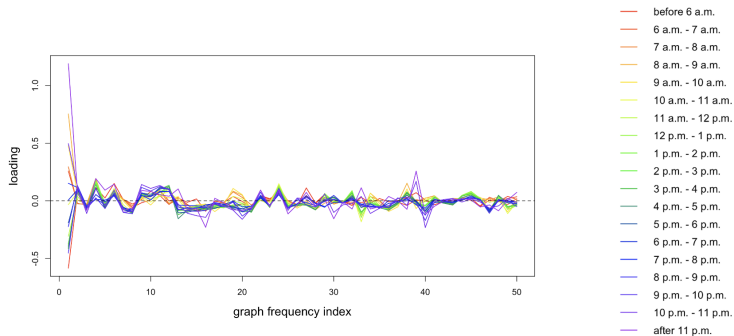


Figure 13: Loadings $b_{i2}^{\text{metro}}(\lambda_{\ell}^{\text{metro}})$ ($i = 1, \dots, 19$, $\ell = 1, \dots, 50$) for the second factor.