# Geum River Network Data Analysis via Weighted PCA


Seeun Park, Kyusoon Kim

Department of Statistics, Seoul National University


July 22, 2021

# Introduction

- River network data provides various measurements of water quality collected at monitoring sites, spread throughout the river network.
- The unique structure of river network interrupts PCA to achieve accurate results.
- Gallacher et al. (2017) proposed the weighted PCA method to solve this problem.
- We applied this method to Geum River network data, and found hidden patterns that were invisible through standard PCA.
- We also suggest the consideration of inhomogeneous covariance structure in river network data would improve the existing method.

# River network data

## Geum River network data

- The dataset shows monthly and yearly (summer) average level of Total Organic Carbon (TOC, $mg/L$) from 2013 to 2020, measured in 127 monitoring sites in Geum River network. For monthly data, seasonality is removed by time series decomposition.

## Unique structure of river network

- Strong spatial and temporal autocorrelation occurs. Since the network is connected by river flow, observed values between sites at subsequent time points are related.
- The relationship between connected sites is direction-dependent, since a downstream site does not affect an upstream site while the opposite is true.
- Standard PCA might lead to an inaccurate result, for example, overemphasizing certain variables in PCA.

# Weighted PCA for river network

- Gallacher et al. (2017) introduces weight matrices to weighted PCA, which reflect the known spatiotemporal structure of river network, in order to adjust autocorrelation among variables.
- For S: spatial weight matrix, T: temporal weight matrix,

$$(S)_{ud} = \begin{cases} \sqrt{\frac{Flow_u}{Flow_d}} & \text{u:upstream, d:downstream are flow-connected} \\ 0 & \text{otherwise} \end{cases}$$

where $Flow_x$ is an indicator of water flow on segment x.

$$(T)_{ij} = \rho^{|i-j|}$$

$T$ is constructed as an AR(1) structure, with $\rho$ being a strength of correlation between time points.

- $S^{-\frac{1}{2}}$ and $T^{-\frac{1}{2}}$ are used to adjust PCA for spatiotemporal correlation.

# Result 1: Time points as variables (T-mode PCA)

**Yearly data**

- The data shows annual summer average from 2013 to 2020.
- PC1 represents the average spatial pattern over all years, and PC2 highlights contrast between early and later years. (2013-2016 and 2017-2020)
- Differences in scores between the standard PCA and weighted PCA are clearer in the higher degree PCs. The weights reflect correlation in the noise structure after removing trend.

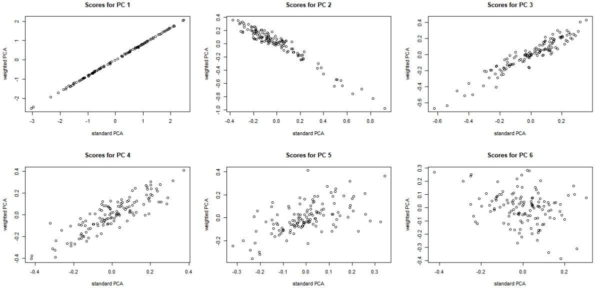



Figure: Comparison of principal component scores (standard PCA vs weighted PCA)

# Result 1: Time points as variables (T-mode PCA)

- The decreases in percentage of variance explained by PC1 show that the weight has removed some of the correlation among variables.

| PCA | PC1 (%) | PC2 (%) | PC3 (%) |
|---|---|---|---|
| Standard | 90.4 | 3.4 | 2.0 |
| Weighted | 81.3 | 6.7 | 4.3 |

## Monthly data

- New patterns were found in the weighted PCA.

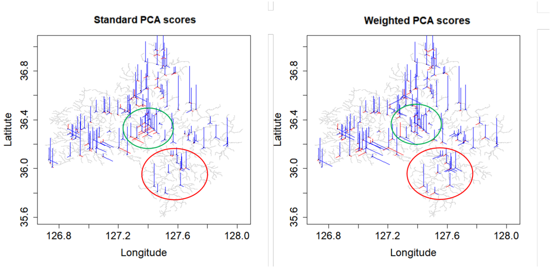
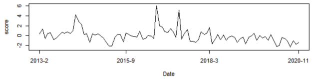



Figure: Glyph plot of scores of standard PCA (left) and weighted PCA (right). The length of line in the direction of 12,4,8 o'clock reflects the magnitude of score of PC1, PC2, PC3 resepectively. Red indicates negative, and blue indicates positive values.

# Result 2: Monitoring sites as variables (S-mode PCA)

**Monthly data**

- PC1 from both the standard and weighted PCA represents temporal pattern from twelve certain sites. The pattern has three peaks on March 2014, August 2016, and April 2017.




- Glyph plot of loadings show which sites contribute a lot to temporal pattern. Weighted PCA distinguishes sites by contribution more clearly.

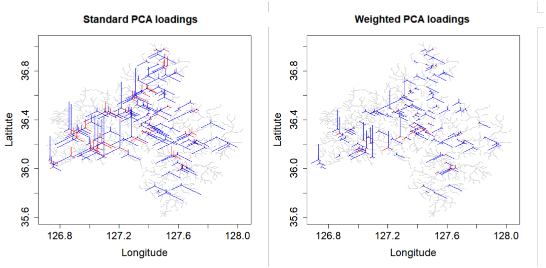



Figure: Glyph plot of loadings of standard PCA (left) and weighted PCA (right).

# Inhomogeneous covariance structure

- Inhomogeneous covariance structures are discovered in the data.
- The structure is visualized with data ellipsoid of monthly data $\epsilon(\bar{y}, S) = \{y : (y - \bar{y})^T S^{-1}(y - \bar{y}) \leq 6\}$, representing score of PCs.
- The weighted PCA should be developed to consider this problem.

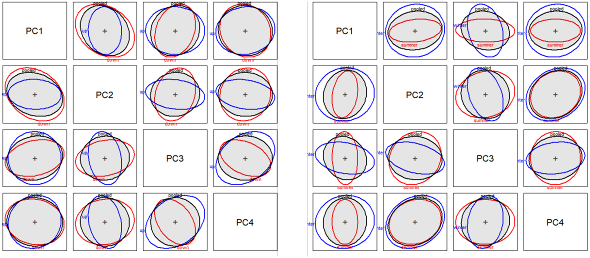



Figure: Downstream vs upstream sites (left) and summer vs winter (right)

## References

1. Gallacher, K., Miller, C., Scott, E. M., Willows, R., Pope, L., and Douglass, J. (2017). Flow-directed PCA for monitoring networks. Environmetrics, 28, e2434.
2. Michael, F., Matthew S. (2020). Visualizing Tests for Equality of Covariance Matrices, The American Statistician, 74(2), 144-155.