Principal Component Analysis for River Network Data

Use of Spatio-temporal Correlation and Heterogeneous Covariance Structure

Kyusoon Kim

Seoul National University

Joint work with
Hee-Seok Oh and Minsu Park

June 23, 2022

Table of Contents

Motivation I : water quality measurements

- Statistical analysis of water quality measurements observed on river networks plays a critical role in understanding spatial and temporal trends of water quality.
- It is vital to reduce the complexity of water quality data since the data are often collected over time at many monitoring sites across the river.

Motivation II : unique structure of the river

- There are **spatial and temporal correlation** between data observed on river networks.
  - Conventional PCA does not take into account the spatial and temporal autocorrelation in river networks, resulting in an inaccurate explanation.
- There exists **spatial heterogeneity** in the flow-connected network.
  - Conventional PCA is a single group method.

Flow-directed PCA

- Gallacher et al. (2017) proposed a new PCA method, called **flow-directed PCA** to adjust the correlations among spatio-temporal data observed in river networks.
- Conventional PCA can be extended to the GMD optimization problem minimizing a $\mathbf{Q},\mathbf{R}$-norm as follows (Allen et al., 2014; Baldwin, 2009).

$$\text{minimize}_{\mathbf{U},\mathbf{D},\mathbf{V}} \left\| \mathbf{X} - \mathbf{U}\mathbf{D}\mathbf{V}^\top \right\|^2_{\mathbf{Q},\mathbf{R}} \text{ subject to}$$
$$\mathbf{U}^\top \mathbf{Q}\mathbf{U} = \mathbf{I}_{(p)}, \mathbf{V}^\top \mathbf{R}\mathbf{V} = \mathbf{I}_{(p)}, \text{diag}(\mathbf{D}) \geq 0,$$

- Here, $\mathbf{Q},\mathbf{R}$-norm is defined as

$$\left\| \mathbf{X} - \mathbf{U}\mathbf{D}\mathbf{V}^\top \right\|^2_{\mathbf{Q},\mathbf{R}} = \sum_{i=1}^{n}\sum_{i'=1}^{n}\sum_{j=1}^{p}\sum_{j'=1}^{p} \mathbf{Q}_{ii'}\mathbf{R}_{jj'} \left(X_{ij} - \boldsymbol{u}_i^\top \mathbf{D}\boldsymbol{v}_j \right) \left(X_{i'j'} - \boldsymbol{u}_{i'}^\top \mathbf{D}\boldsymbol{v}_{j'} \right)$$
$$= \text{tr}\left(\mathbf{Q}\left(\mathbf{X} - \mathbf{U}\mathbf{D}\mathbf{V}^\top \right) \mathbf{R} \left(\mathbf{X} - \mathbf{U}\mathbf{D}\mathbf{V}^\top \right)^\top \right).$$

- Note that it is equivalent to the conventional PCA problem when $\mathbf{Q},\mathbf{R} = \mathbf{I}$, which is identical to the Frobenius norm.

Spatial and temporal weights

- For the flow-directed PCA, $\mathbf{Q} = \left(\mathbf{S}^{-\frac{1}{2}}\right)^{\top} \mathbf{S}^{-\frac{1}{2}}$ and $\mathbf{R} = \left(\mathbf{T}^{-\frac{1}{2}}\right)^{\top} \mathbf{T}^{-\frac{1}{2}}$ by the spatial and temporal weight matrices, $\mathbf{S}$ and $\mathbf{T}$.
- $\mathbf{S}_{s_d,s_u} = \begin{cases} \sqrt{\dfrac{\text{Shreve order}_{s_u}}{\text{Shreve order}_{s_d}}}, & \text{if } s_d \text{ and } s_u \text{ are flow-connected, } s_d \text{ and } s_u \text{ represent downstream site and upstream site respectively.} \\ 0, & \text{otherwise.} \end{cases}$
- $\mathbf{T}_{i,j} = \rho^{|i-j|}$ where ρ is the strength of correlation between observations at consecutive time points under an AR(1) model (Clement et al., 2006; Houseman, 2005).

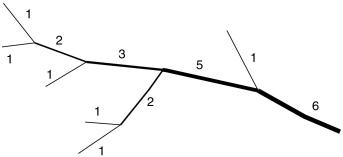

Figure 1: Shreve stream order

Geographically weighted PCA (GWPCA)

- **GWPCA** is one of the techniques accounting for specific spatial heterogeneity (Fotheringham et al., 2002; Harris et al., 2011).
- It is assumed that $\boldsymbol{x}_i | (u_i, v_i) \sim (\boldsymbol{\mu}(u_i, v_i), \boldsymbol{\Sigma}(u_i, v_i))$, where (u_i, v_i) denotes a spatial location of $\boldsymbol{x}_i$.
- The eigendecomposition of GW variance-covariance matrix gives the local loadings for each location.

$$\mathbf{X}^\top \mathbf{W}_i \mathbf{X} = \mathbf{X}^\top \mathbf{W}(u_i, v_i)\mathbf{X} = \mathbf{L}\mathbf{V}\mathbf{L}^\top | (u_i, v_i)$$

  where $\mathbf{W}_i$ is a diagonal matrix of geographic weights based on a distance-decaying kernel weight function (e.g. bi-square function).

Goals

- We propose a new PCA method that can be applied to streamflow data by combining and modifying the flow-directed PCA and the GWPCA.
- The aim of the study is to reduce dimensionality for streamflow data while reflecting the unique structure of the river as follows.
  - adjust for spatio-temporal autocorrelation in river networks.
  - consider spatial heterogeneity and identify heterogeneous covariance group structures.
- We perform a real data analysis for total organic carbon (TOC) from the Geum River network in South Korea to demonstrate the strength and usefulness of the proposed method.

Geum River

- The Geum River basin, located in the midwest of South Korea, was selected for this study.
- There are about 50 streams flowing into the Geum River, and urban sewage flowing from various cities such as Daejeon and Cheongju, and complex climate issues.
- Therefore, it is necessary to examine the water quality data of the Geum River closely.

Geum catchment area

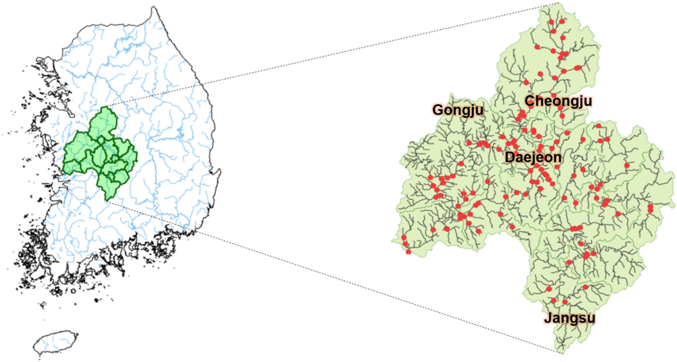

Figure 2: Geum catchment area and monitoring sites

Total organic carbon data

- Total organic carbon (TOC) represents the amount of carbon found in organic compounds and is frequently used as an indicator of water quality.
- TOC level was measured at 127 monitoring sites in the Geum River catchment area from the winter of 2012 to the fall of 2020.
- Annual summer average log TOC was used because higher TOC level poses a greater threat to human and the environment, and the average TOC concentration is generally higher in summer (Kathiresan et al., 2014; LeChvallier et al., 1990; Shen et al., 2013).

Step1: modification of GWPCA for flow-connected networks

- Unlike general geospatial data, they have interrelations that depend heavily on flow connection, flow direction, and stream distance.
- We propose a modified version of the GWPCA to reflect the distinct characteristics of rivers by replacing the weight matrix in GWPCA with a **flow-based weight matrix**.
- To define a flow-based weight matrix, **upstream flow distance** f should be defined in advance.
- The advantage of the modified GWPCA with flow-based weights is that it provides a local structure of the flow-connected network data.

Upstream flow distance

Definition

The **upstream flow distance** between s_1 and s_2, f_{s_1,s_2} is defined as stream distance if s_1 and s_2 are flow-connected and s_1 and s_2 represent downstream and upstream sites, respectively. Otherwise, f_{s_1,s_2} is defined as ∞.

- $f_{A,B} = a + b$, whereas $f_{B,A} = f_{B,C} = \infty$.

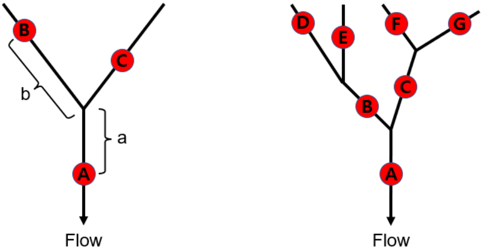

Figure 3: Toy river networks

Flow-based weight

- We use bi-square function.
- A **flow-based weight matrix** for monitoring site s_1, diagonal matrix $\mathbf{W}^*_{s_1}$, is then constructed by

$$\left(\mathbf{W}^*_{s_1}\right)_{s_2,s_2} = \left(1 - \left(f_{s_1,s_2}/r_{s_1,N}\right)^2\right)^2 \mathbb{1}\left(f_{s_1,s_2} \leq r_{s_1,N}\right),$$

  - N: the bandwidth of the adaptive bi-square kernel.
  - $r_{s_1,N}$: the upstream flow distance from s_1 to its $(N-1)$th nearest neighbor.
- In the right panel of Figure 3,
  - $r_{A,3} = f_{A,C}$
  - $r_{B,3} = f_{B,D}$
  - $r_{C,3} = f_{C,G}$

Step2: combination with flow-directed PCA

- We combine the modified GWPCA with the flow-directed PCA to consider both spatio-temporal autocorrelation and spatial heterogeneity.
- Recall that
  - GWPCA uses the eigendecomposition of $\mathbf{X}^\top \mathbf{W}_i \mathbf{X}$ and it is equivalent to minimize $\left\| \mathbf{W}_i^{\frac{1}{2}} \mathbf{X} - \mathbf{U}\mathbf{D}\mathbf{V}^\top \right\|_F^2$.
  - Flow-directed PCA extends the Frobenius norm to the $\mathbf{Q}$,$\mathbf{R}$-norm.
- Thus, we consider the problem of minimizing $\left\| \mathbf{W}_i^{*\frac{1}{2}} \mathbf{X} - \mathbf{U}\mathbf{D}\mathbf{V}^\top \right\|_{\mathbf{Q},\mathbf{R}}^2$ for $\mathbf{Q} = \left(\mathbf{S}^{-\frac{1}{2}}\right)^\top \mathbf{S}^{-\frac{1}{2}}$ and $\mathbf{R} = \left(\mathbf{T}^{-\frac{1}{2}}\right)^\top \mathbf{T}^{-\frac{1}{2}}$ to get a combined result of the two methods.

Step3: identification of heterogeneous covariance structure

- Identifying heterogeneous covariance structure might enable us to perform additional analysis and it is helpful to understand the data better.
- We supply local loadings to Ward's minimum variance method, a hierarchical clustering method, to divide group (Harris et al., 2015; Ward, 1963).
- After partitioning the monitoring sites into several groups, common principal component analysis (CPCA; see Flury, 1987, 1984) is used to complete the analysis.

Flow chart

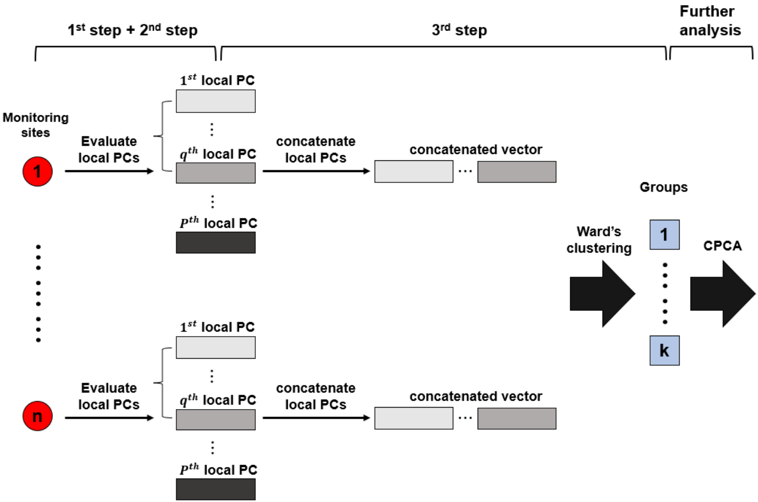

Figure 4: Flow chart outlining the proposed method

Comparison of PCA results

- The proposed method has reduced the variance accounted for by the first PC and increased the variance by the second PC significantly compared to the conventional PCA and the flow-directed PCA.

Table 1: T-mode PCA results.

		PC1(%)	PC2(%)	PC3(%)	Var_2(%)
$TPCA_{uw}$		90.4	3.3	2.0	93.7
$TPCA_S$		87.2	4.8	2.7	92.0
$TPCA_T$		85.0	5.1	3.3	90.1
$TPCA_{ST}$		80.4	7.1	4.6	87.5
Proposed method	1st quantile	67.0	8.6	3.2	83.8
	2nd quantile	75.4	15.2	4.3	89.9
	mean	74.4	14.0	4.3	88.4
	3rd quantile	83.2	16.7	5.5	92.9

Glyph plot

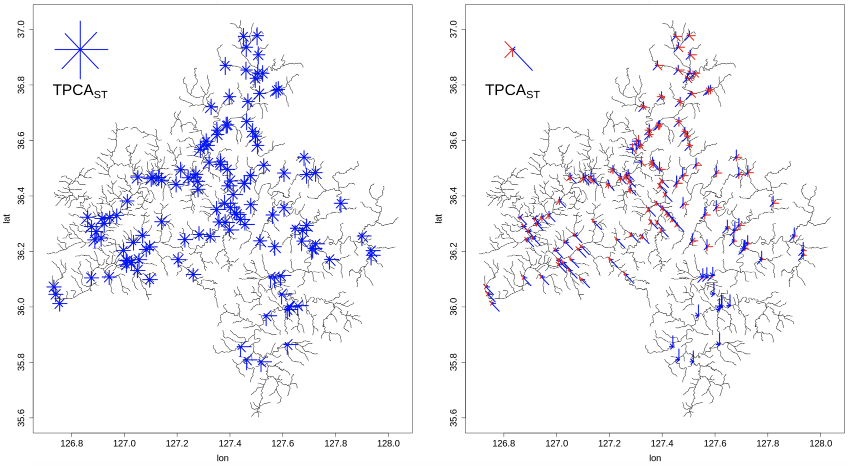

Figure 5: Multivariate glyph plots with local loadings for the first two principal components (left: PC1, right: PC2).

Identification of group structure

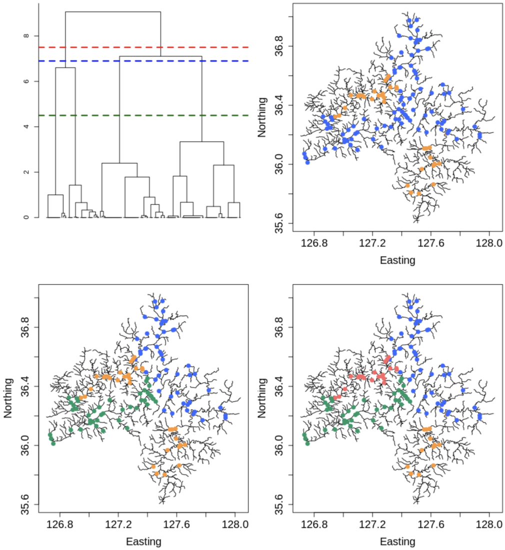

Figure 6: Ward clustering outcome for $k = 2, 3, 4$.

Flury's AIC

- It deals in detail with the case of $k = 2$.
- From the likelihood ratio test and ensemble test (Pepler, 2014), CPC(2) model fits sufficiently well.

Table 2: Flury's AIC and Chi-squared statistics for the Geum River data ($k = 2, p = 8$).

Model	$\chi^2_{partial}$	df	$\chi^2_{partial}/\text{df}$	AIC	No.of.CPCs
Equality	2.72	1	2.72	87.11	8
Proportionality	27.77	7	3.97	86.40	8
CPC	2.90	1	2.90	72.63	8
CPC(6)	0.63	2	0.31	71.73	6
CPC(5)	10.83	3	3.61	75.10	5
CPC(4)	9.58	4	2.39	70.26	4
CPC(3)	13.69	5	2.74	68.69	3
CPC(2)	8.26	6	**1.38**	**64.99**	2
CPC(1)	10.73	7	1.53	68.73	1
Heterogeneity	-	-	-	72.00	0

CPCA results

- CPC1 represents the average spatial pattern over summers.
- CPC2 represents a difference between 2016 and other years.
- There are differences between individual-specific PCs.
  - Unlike in Group2, the component that explained the second-largest proportion of the total variance in Group1, 14.0%, was a group-specific component representing a contrast between 2013 and other years, not CPC2.

Table 3: CPCA results.

CPC	2013	2014	2015	2016	2017	2018	2019	2020	Group1(%)	Group2(%)
CPC1	0.47	0.39	0.47	0.43	0.47	0.44	0.43	0.36	62.5	83.4
CPC2	-0.03	0.11	0.10	0.81	-0.21	-0.32	-0.05	-0.28	10.5	6.5

Hidden patterns

- We can figure out specific spatial patterns that were not visible in the unweighted PCA.

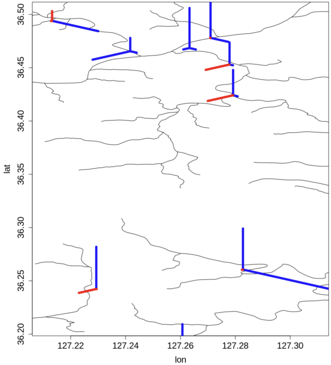

Figure 7: Glyph plots with the first three principal component scores for the unweighted PCA (left), and the CPCA with spatial and temporal adjustment (right).

Concluding remarks

- The proposed method enabled us to effectively reduce the dimensionality of streamflow data considering the unique characteristics of river networks.
- It is possible to figure out interesting spatial features that are not due to the topological structure of the river and temporal correlation among measurements by eliminating correlation between data
- We can identify group structure and spatially varying sources of variation within water quality measurements.
- Weight matrices describing spatial and temporal correlation need to be developed based on a data-adaptive method.

Thank you.

References I

Allen, G. I., Grosenick, L., and Taylor, J. (2014). A generalized least-square matrix decomposition. *Journal of the American Statistical Association*, 109(505):145–159.

Baldwin, M. P. (2009). Spatial weighting and iterative projection methods for eofs. *Journal of Climate*, 22(2):234–243.

Clement, L., Thas, O., Vanrolleghem, P. A., and Ottoy, J. P. (2006). Spatio-temporal statistical models for river monitoring networks. *Water Science & Technology*, 53(1):9–15.

Flury, B. K. (1987). Two generalizations of the common principal component model. *Biometrika*, 74(1):59–69.

Flury, B. N. (1984). Common principal components in k groups. *Journal of the American Statistical Association*, 79:892–898.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley.

Gallacher, K., Miller, C., Scott, E. M., Willows, R., Pope, L., and Douglass, J. (2017). Flow-directed pca for monitoring networks. *Environmetrics*, 28(2):e2434.

References II

Harris, P., Brunsdon, C., and Charlton, M. (2011). Geographically weighted principal components analysis. *International Journal of Geographical Information Science*, 25(10):1717–1736.

Harris, P., Clarke, A., Juggins, S., Brunsdon, C., and Charlton, M. (2015). Enhancements to a geographically weighted principal component analysis in the context of an application to an environmental data set. *Geographical Analysis*, 47(2):146–172.

Houseman, E. A. (2005). A robust regression model for a first-order autoregressive time series with unequal spacing: application to water monitoring. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 54(4):769–780.

Kathiresan, K., Gomathi, V., Anburaj, R., and Saravanakumar, K. (2014). Impact of mangrove vegetation on seasonal carbon burial and other sediment characteristics in the vellar-coleroon estuary, india. *Journal of Forestry Research*, 25(4):787–794.

LeChvallier, M. W., Olson, B. H., and McFeters, G. A. (1990). *Assessing and Controlling Bacterial Regrowth in Distribution Systems*. American Water Works Association.

References III

Pepler, P. T. (2014). *The Identification and Application of Common Principal Components*. PhD thesis, Stellenbosch University.

Shen, J., Wu, X., Zhang, Z., Gong, W., He, T., Xu, X., and Dong, H. (2013). Ti content in huguangyan maar lake sediment as a proxy for monsoon-induced vegetation density in the holocene. *Geophysical Research Letters*, 40(21):5757–5763.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.