A Comparison of Geographically Weighted Principal Components Analysis Methodologies

Narumasa Tsutsumida[1]
Saitama University, Japan

Daisuke Murakami
Institute of Statistical Mathematics, Tokyo, Japan

Takahiro Yoshida
The University of Tokyo, Japan

Tomoki Nakaya
Tohoku University, Sendai, Japan

Binbin Lu
Wuhan University, China

Paul Harris
Rothamsted Research, Harpenden, UK

Alexis Comber
University of Leeds, UK

Abstract

Principal components analysis (PCA) is a useful analytical tool to represent key characteristics of multivariate data, but does not account for spatial effects when applied in geographical situations. A geographically weighted PCA (GWPCA) caters to this issue, specifically in terms of capturing spatial heterogeneity. However, in certain situations, a GWPCA provides outputs that vary discontinuously spatially, which are difficult to interpret and are not associated with the output from a conventional (global) PCA any more. This study underlines a GW non-negative PCA, a geographically weighted version of non-negative PCA, to overcome this issue by constraining loading values non-negatively. Case study results with a complex multivariate spatial dataset demonstrate such benefits, where GW non-negative PCA allows improved interpretations than that found with conventional GWPCA.

2012 ACM Subject Classification Mathematics of computing → Multivariate statistics

Keywords and phrases Spatial heterogeneity, Geographically weighted, Sparsity, PCA

Digital Object Identifier 10.4230/LIPIcs.COSIT.2022.21

Category Short Paper

Funding This research was funded by the Joint Support Center for Data Science Research at Research Organization of Information and Systems (ROIS-DS-JOINT) under Grant 006RP2018, 004RP2019, 003RP2020, and 005RP2021.

1 Introduction

A principal components analysis (PCA) summarizes multi-dimensional data [5], by reducing the number of dimensions of the dataset. It provides a purely mathematical means of highlighting key sources of variation. Due to its form, spatial effects of spatial autocorrelation and spatial heterogeneity are not considered in transforming the multi-dimensional spatial data. For spatial data, some PCA methods have been developed to consider these spatial

[1] corresponding author

autocorrelation and heterogeneity [2]. Spatial PCA, referred to as sPCA, takes spatial autocorrelation into account to reveal spatial patterns [6]. sPCA yields global principal components (PCs) similar to the conventional PCA, but its scores underline the spatial autocorrelation; thus, for example, the spatial distribution of its first PC score is positively high autocorrelated. Similarly, geographically weighted (GW) PCA takes spatial heterogeneity into account. Analogous to GW regression (GWR) [1], GWPCA assembles local PCAs by applying a moving window weighted kernel and yields spatially varying eigenvalues, loadings, and percentage of total variance captured by each PC [4, 7]. The outputs of GWPCA are extensive but unique in terms of their spatial variations, allowing an investigation into the spatial data structure [2]. However, in some cases, GWPCA gives spatially discontinuous loadings (from positive to negative, for example, see [10]), and this presents problems of interpretation, and espcially in terms of relating to its global, conventional PCA counterpart. This issue arises because GWPCA assembles local PCAs that are independent of each other. To deal with this issue, we consider a GW non-negative PCA (GWnnegPCA) to constrain all loadings non-negatively [11]. This results in that at any local PCA calibration point, the input variables are linearly summed to build new PCs, but where loadings are only zero or positive, providing a more intuitive interpretation for the GWnnegPCA mapped outputs as a whole.

2 Methods

GWPCA assembles a series of local PCAs where each PCA is constructed using nearby, spatially-weighted data according to a moving window kernel. Given a $n \times m$ matrix $\overline{\mathbf{X}}$ which consists of m variables at n observation sites and each variable is re-scaled to zero-mean and unit-variance, GWPCA decomposes the GW variance-covariance matrix of $\overline{\mathbf{X}}$ at the p-th location with coordinates (u_p, v_p), which is defined by $\mathbf{\Sigma}_p = \overline{\mathbf{X}}^T \mathbf{W}_p \overline{\mathbf{X}}$, by

$$\mathbf{L}_p \mathbf{V}_p \mathbf{L}_p^T = \mathbf{\Sigma}_p, \tag{1}$$

where $\mathbf{L}_p$ is a GW loading matrix and $\mathbf{V}_p$ is a diagonal matrix of GW eigenvalues at the p-th location. $\mathbf{W}_p$ is a diagonal matrix of geographic weights that can be generated using a given kernel function. In this case study, we used the bisquare kernel function for the q-th location:

$$\omega_{pq} = \begin{cases} \left(1 - \left(\frac{d_{pq}}{b}\right)^2\right)^2 & \text{if } |d_{pq}| < b, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where the bandwidth size b is any distance for a given number of nearest observations (i.e., 100 or 10% of the total observation), and d_{pq} is the euclidean distance between spatial locations of the p-th and q-th data locations in this work.

The first loading $\mathbf{l}_p^{(1)}$ for the GWPCA at the p-th location is applied so that:

$$\underset{\mathbf{l}_p^{(1)}}{argmax}\ \mathbf{l}_p^{(1)T} \mathbf{\Sigma}_p \mathbf{l}_p^{(1)}, \quad subject\ to\ \|\mathbf{l}_p^{(1)}\|_2 = 1 \tag{3}$$

The subsequent loading maximizes the variance under the constraint that it is orthogonal to the previous component(s) [9]: $\mathbf{l}_p^{(i)T} \mathbf{\Sigma} \mathbf{l}_p^{(i)}$ for all $i \in \{2, ..., m\}$, $subject\ to\ \|\mathbf{l}_p^{(i)}\|_2 = 1$ and $\mathbf{l}_p^{(i-1)} \mathbf{l}_p^{(i)} = 0$.

A robust GWPCA has also been proposed to reduce the effect of anomalous observations on the output [3]. This uses a local covariance matrix by using the robust minimum covariance determinant (MCD) estimator [8].

In this context, we develop the GWnnegPCA which applies the following additional restriction to the equation (3):

$$subject\ to\ \ \|\mathbf{l}_p^{(1)}\|_0\ \leq k,\ \ \mathbf{l}_p^{(1)} \geq \mathbf{0}, \tag{4}$$

where k ($\leq m$) is the number of non-zero variables and $\|\mathbf{l}_p^{(1)}\|_0$ is the cardinality of $\mathbf{l}_p^{(1)}$. This makes all the loadings at any location non-negative. However, as the order of local PC axes determined by the eigenvalues, the order at the p-th location from GW-based PCA may not be the same as that from the conventional PCA due to high spatial heterogeneity. Thus, by using the result of the conventional and non-negative PCA loadings, we rearranged the order of local loadings by GW-based PCAs. The Pearson's correlation matrix between the absolute values of the global and the local loadings at every location was calculated, then the local loadings were reordered according to the correlation coefficient. This modification is expected to make global and local loadings comparative.

The local PC scores at the p-th location, $\mathbf{S}_p$, are represented by:

$$\mathbf{S}_p = \overline{\mathbf{X}}_p\mathbf{L}_p \tag{5}$$

It is noted that to introduce the non-negativity, the cardinality in equation (4) requires a minimum angle between components and thus the orthogonality constraint of the PCs is relaxed [9]. This means that components are quasi-orthogonal amongst the PCs. In this sense, loadings from non-negative PCA and GW non-negative PCA are regarded as quasi-eigenvectors.

Bandwidth is a critical parameter in the GW framework as its size determines the localness of the analysis and whether the given process is indeed non-stationary. The bandwidth of GWPCA is optimized by a leave-one-out cross-validation [4] and in this study, the optimized bandwidth by GWPCA is also used for robust GWPCA and GWnnegPCA to be comparative. The bandwidth size used in this study was 45.3% of the total.

3 Case study

For our case study, we build 21 variables using census statistics of Japan in 2005 (Table 1). These variables describe the urban social structure of Tokyo within the 3,134 *chocho-aza* units (the smallest administrative unit in Japan) of the 23 special wards, Tokyo. All 21 variables were standardized (zero mean with a unit variance) before being input into our non-spatial and GW models.

4 Results

Spatial distribution maps of a loading ($HIGHEDU$) for PC1-3 by GWPCA, robust GWPCA, and GWnegPCA were shown in Figure 1 as an example. The loading values of this variable for conventional PCA were 0.30, 0.19, and -0.37, respectively, and those for non-negative PCA were 0.42, 0.02, and 0.11, respectively. All results from the GWPCA, robust GWPCA, and GWnnegPCA are associated with the conventional and non-negative PCAs, while GW-based PCAs show spatial distributions of loading values. The loading map for GWPC1 represents the higher value surrounding the center of Tokyo compared to other regions. The map for GWPC2 represents slightly lower negative values in the east and the north of the regions, while positive values are found in the central part. The negative values are also found with a strip shape within the central part. Furthermore, the map for GWPC3 shows a clear discontinuous spatial pattern with positive/negative patches. Such loading maps make us confuse how to interpret them, resulting in the difficulty of using GWPCA.

Table 1 Variable descriptions used in this study.

Abbreviation	Variable Descriptions
AGE0014	Num. of persons in the age of 0 – 14 / The total pop.
AGE1529	Num. of persons in the age of 15 – 29 / The total pop.
AGE3044	Num. of persons in the age of 30 – 44 / The total pop.
AGE4564	Num. of persons in the age of 45 – 64 / The total pop.
AGE65	Num. of persons in the age over 65 / The total pop.
UNIVPOP	Num. of university students / The total pop.
HIGHEDU	Num. of university graduates / The total pop.
NUCFAMR	Num. of nuclear families / The total households
MULTIFAM	Num. of extended families / The total households
SINGHR	Num. of single households / The total households
OWNHR	Num. of owned housing households / The total households
SELFEMP	Num. of self employments / The total worker pop.
WHITER	Num. of white-coloured employees / The total worker pop.
BLUER	Num. of blue-coloured employees / The total worker pop.
PRIMR	Num. of the prime sector employees / The total worker pop.
SECR	Num. of the second sector employees / The total worker pop.
TERR	Num. of the third sector employees / The total worker pop.
SHORTCOM	Num. of commuters ($\leq$ 30 min) / The total num. of commuters
LONGCOM	Num. of commuters ($\geq$ 1 hour) / The total num. of commuters
NEWCOMER	Num. of in-migrant pop. ($\geq$ 5 yrs.) in 5 yrs. / Total pop. ($\geq$ 5 yrs.)
WPRATIO	Num. of workers / The total pop.

The loading maps for robust GWPCA show a more continuous spatial pattern than those for GWPCA, but at the middle west part of the region, a positive patch is found in the loading map for robust GWPC2. Such a pattern can be found in the loading map for robust GWPC3 such that a negative patch in regions with positive values.

GWnnegPCA also provides loading maps, and those spatial patterns are simpler. The loading map for GWnnegPC1 represents a positive value in large parts of the area. The map for GWnnegPC2 shows zero value in almost all areas, corresponding with the result of non-negative PCA. The map for GWnnegPC3 represents positive values in the middle and western regions while zero in others. These loading values for GWnnegPCA tell in which areas this $HIGHEDU$ variable contributes to each PC.

5 Discussion

This study demonstrates that GWPCA can give quite discontinuous spatial patterns for its localized loadings. This characteristic can be challenging to interpret spatially compared to the corresponding conventional global PCA outputs. This is especially pertinent as GW methods are often employed to (hopefully) reveal smoothly changing spatial drifts of model parameters, not discontinuous ones. The discontinuous issue would come from the flexibly rotated axes of local PCs at each observation point. It is applicable to fix the first axis direction to be all positive in GWPCA by applying a similar modification as the *fix_sign* option in the *princomp* R function. However, such an approach may not work straightforwardly for the subsequent components for GW-based PCA because the reminders of data variance after applying the equation (3) are not constant over space: that is, the

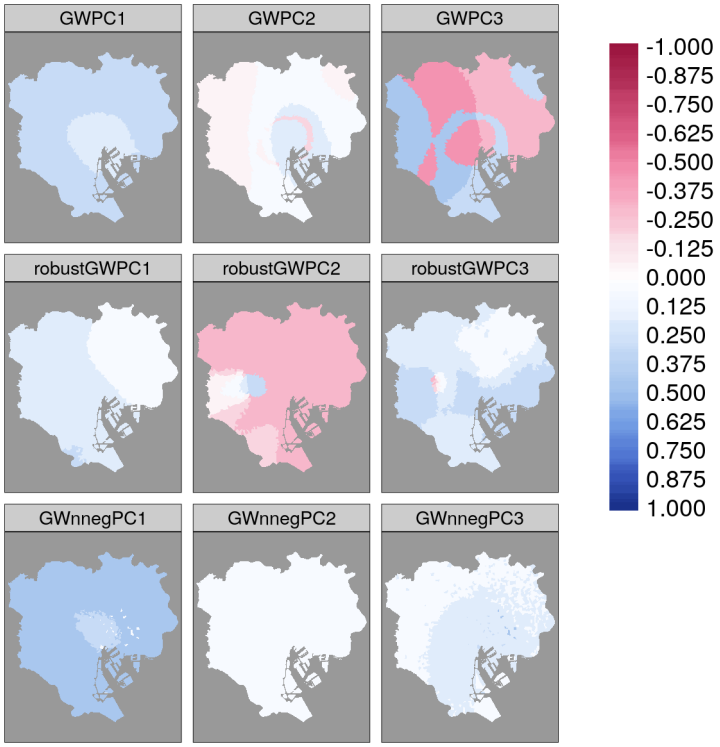

Figure 1 Loading maps of the HIGHEDU variable for the first, the second, and the third principal component by geographically weighted principal components analysis, robust geographically weighted principal components analysis, and geographically weighted non-negative principal components analysis as an example.

explained variation in the data for the GWPC1 varies spatially. Thus, the second component axis (and subsequent component axes as well) at each observation point cannot be determined in a particular manner, and such rotations at the p-th location cannot be coordinated with each other. GWnnegPCA has overcome this issue significantly. Non-negativity of local PCAs fixes the flexible rotation of PC axes of GWPCA, and thus the loading maps represent spatially varying patterns. Figure 1 showed a clear difference of spatially varying patterns of loading maps of an example variable. GWPCA has been used in many case studies to obtain the largest absolute value of loading at every location as the winning variable [4], and is valuable to see the spatial heterogeneity in the input data structure. It is however challenging to investigate the spatial pattern of each loading due to the discontinuous problem as seen in this study, even applying the robust way. In this context, GWnnegPCA has the potential to show the continuous spatial variation smoothly that would contribute to further interpretations of the result.

Similar to the discussion on the validity of installing the non-negativity for PCA, GWnnegPCA also inherits the technical issue of heavy computation and relaxing orthogonality [9]. In addition, we found noisy patterns in loading maps of GWnnegPCA, which may come from the sparsity and outliers in input data. Further investigations will be expected to be the stability of the result.

6 Conclusions

We demonstrated in the case study that geographically weighted principal components analysis (GWPCA) and robust GWPCA yield discontinuous spatial patterns for its localized loadings and introduced geographically weighted non-negative principal components analysis (GWnnegPCA) to overcome this issue. GWnnegPCA would be a reasonable choice to show spatially varying loading values in multivariable spatial data so that the degree of variable(s) contribution to a principle component varies spatially. This allows us to interpret data locally to understand regional characteristics in the spatial data as we expect GW approach. We will work on interpreting regional changes in loading patterns and handling noises on loading maps for further developments.

References

1 Chris Brunsdon, A Stewart Fotheringham, and Martin E Charlton. Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geogr. Anal.*, 28(4):281–298, September 1996. doi:10.1111/j.1538-4632.1996.tb00936.x.

2 Urška Demšar, Paul Harris, Chris Brunsdon, A. Stewart Fotheringham, and Sean McLoone. Principal Component Analysis on Spatial Data: An Overview. *Ann. Assoc. Am. Geogr.*, 103(1):106–128, January 2013. doi:10.1080/00045608.2012.689236.

3 Isabella Gollini, Binbin Lu, Martin Charlton, Christopher Brunsdon, and Paul Harris. GWmodel : An R Package for Exploring Spatial Heterogeneity Using Geographically Weighted Models. *J. Stat. Softw.*, 63(17):85–101, April 2015. doi:10.18637/jss.v063.i17.

4 Paul Harris, Chris Brunsdon, and Martin Charlton. Geographically weighted principal components analysis. *Int. J. Geogr. Inf. Sci.*, 25(10):1717–1736, October 2011. doi:10.1080/13658816.2011.554838.

5 Ian Jolliffe. Principal Component Analysis. In *Wiley StatsRef Stat. Ref. Online*. John Wiley & Sons, Ltd, Chichester, UK, September 2014. doi:10.1002/9781118445112.stat06472.

6 T. Jombart, S. Devillard, A. B. Dufour, and D. Pontier. Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity (Edinb).*, 101(1):92–103, 2008. doi:10.1038/hdy.2008.34.

7 Christopher D. Lloyd. Analysing population characteristics using geographically weighted principal components analysis: A case study of Northern Ireland in 2001. *Comput. Environ. Urban Syst.*, 34(5):389–399, 2010. doi:10.1016/j.compenvurbsys.2010.02.005.

8 Peter Rousseeuw. Multivariate Estimation with High Breakdown Point, 1985.

9 Christian D. Sigg and Joachim M. Buhmann. Expectation-maximization for sparse and non-negative PCA. In *Proc. 25th Int. Conf. Mach. Learn. - ICML '08*, pages 960–967, New York, New York, USA, 2008. ACM Press. doi:10.1145/1390156.1390277.

10 Narumasa Tsutsumida, Paul Harris, and Alexis Comber. The Application of a Geographically Weighted Principal Component Analysis for Exploring Twenty-three Years of Goat Population Change across Mongolia. *Ann. Am. Assoc. Geogr.*, 107(5):1060–1074, 2017. doi:10.1080/24694452.2017.1309968.

11 Narumasa Tsutsumida, Daisuke Murakami, Takahiro Yoshida, Tomoki Nakaya, Binbin Lu, and Paul Harris. Geographically Weighted Non-negative Principal Components Analysis for Exploring Spatial Variation in Multidimensional Composite Index. *GeoComputation 2019*, September 2019. doi:10.17608/k6.auckland.9850826.v1.