Spatial heterogeneity in principal component analysis: a study of deprivation index on Italian provinces

Eterogeneità spaziale nell'analisi delle componenti principali: uno studio sull'indice di deprivazione sulle province italiane

Paolo Postiglione[1], M. Simona Andreano[2], Roberto Benedetti[3], Alfredo Cartone

Abstract Principal Component Analysis (PCA) is a tool often used for the construction of composite indicators even at the local level ([18]). In general, when we are dealing with spatial data, the method of PCA, in its classical version, is not appropriate for the synthesis of simple indicators. The objective of this paper is to introduce a method to take into account the spatial heterogeneity in PCA, extending the contribution introduced by [19]. The proposed method will be implemented for the definition of a deprivation index on Italian provinces.

Abstract *L'analisi delle componenti principali (ACP) è uno strumento spesso utilizzato per la costruzione di indicatori compositi anche a livello locale ([18]). In generale, quando stiamo lavorando su dati spaziali, l'ACP, nella sua versione classica, non è appropriata per la sintesi di indicatori semplici. L'obiettivo di questo lavoro è di introdurre un metodo che considera l'eterogeneità spaziale nell'ACP, estendendo l'idea di [19]. Il metodo proposto sarà implementato per la definizione di un indice di deprivazione nelle province italiane.*

Key words: Simulated annealing, GWPCA, composite indicators, spatial effects.

1. Introduction

Principal Component Analysis (PCA, [12]) is a statistical method largely adopted in empirical applications. PCA returns a set of independent variables of correlated variables by decomposing the eigen-structure of the variance-covariance matrix ([13]). Typical output of PCA are vectors of loadings corresponding to eigenvectors and new sets of coordinates corresponding to components. PCA is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that

[1] University of Chieti-Pescara, Department of Economic Studies, email: postigli@unich.it

[2] Universitas Mercatorum, email: s.andreano@unimercatorum.it

[3] University of Chieti-Pescara, Department of Economic Studies, email: benedett@unich.it

[4] University of Chieti-Pescara, Department of Economic Studies, email: alfredo.cartone@unich.it

still contains most of the information in the large set. It is also used to the aim of exploring the data.

PCA is often applied on geographically distributed data ([3]). Spatial data present particular characteristics that should be considered when applying statistical technique: spatial dependence and spatial heterogeneity that are the two inherent characteristics of spatial data. Spatial dependence can be defined as "the propensity for nearby locations to influence each other and to possess similar attributes" ([8]). On the other hand, as evidenced by [1], there are two distinct forms of spatial heterogeneity: structural instability and heteroscedasticity. Structural instability concerns the presence of varying structural parameters over space. Heteroscedasticity leads to different error variances of the spatial units.

In this paper, our main aim is to consider the problem of spatial heterogeneity when using PCA for geographically distributed data. In particular, we will address the idea that coefficient estimates can vary across space, leading to spatial structural instability. As argued by [19], the analysis and the assessment of heterogeneity for geographically distributed data is one of the main challenges for the spatial analysts. Empirical models that do not take into account for structural heterogeneities may show serious misspecification problems ([20]).

PCA can also be employed to define composite indicators ([17]). The use of composite indicators is common in practical analyses because we often meet multidimensionality in the real world ([16]). The loadings of PCA may be used as weights in the building of the composite indicators. Some Authors (see, for example, [4]) criticize this use of PCA, because the weights from PCA are defined through a statistical technique and may not reflect the relevance of the single variable for the underlying phenomenon. However, weights from PCA may be less "subjective" because these are not assigned by the researcher and are data-driven, differently from the case of "normative" weights. See [2]) for a discussion about the methods for deriving composite indicators.

Spatial heterogeneity has been considered in PCA through the approach denoted as geographically weighted principal components analysis (GWPCA) ([5]). This method allows for differences in the loadings and scores structure due to spatial instability. The output of GWPCA is represented by estimates of the covariance matrix and sets of components for each locality ([10]). In this way, distinct composite indicators are defined differently for each locality. It is clear that the interpretation of such a list of different composite indicators at local level is not entirely straightforward.

In this paper, we propose to use a modified version of simulated annealing (SA) introduced by [19] to identify zones of local stationarity in the eigenvalues and in the corresponding eigenvectors defined by PCA. The presence of the heterogeneity is a criterion to divide the sample of observations (i.e. regions) into smaller homogeneous groups. Therefore, in our case, we are able to define a composite indicator for each partition identified by SA algorithm.

The use of PCA for deriving composite indicators of deprivation has been extensively explored ([18]). Deprivation may be defined "as a state of observable and demonstrable disadvantage relative to the local community or the wider society or nation to which an individual, family or group belongs" ([21]). Its measurement

considers several dimensions from both the social and economic sphere to assess the presence of disparities. In this paper, our aim is to define a composite indicator of deprivation for group of Italian provinces.

The layout of the paper is the following. Section 2 is devoted to briefly summarize the methodological contribution of the paper. In particular, we review the main characteristics of PCA and how SA can be applied to identify zone of local stationarity for eigenvalues and eigenvectors. Section 3 contains the description of our data set and shows the results of the composite indicator for Italian provinces. Finally, section 4 concludes.

2. The methodology

PCA is based on the analysis of a matrix $\mathbf{X}_{nm}$ where $i = 1, \dots, n$ denotes the statistical units and $j = 1, \dots, m$ the variables, respectively. The central idea of PCA is the representation of units in q-dimensional subspaces (with $q < m$) retaining the maximum of statistical information. The reduction of data dimensionality allows us easier interpretative analysis. A primary result in PCA is ([13]):

$$\mathbf{A}\mathbf{\Lambda}\mathbf{A}^t = \mathbf{\Sigma} \tag{1}$$

where $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues, $\mathbf{A}$ is the corresponding matrix of loadings (i.e., the eigenvectors), and $\mathbf{\Sigma}$ is the variance/covariance matrix. The eigenvalues in $\mathbf{\Lambda}$ represent the variance of the principal component, $\mathbf{Y}_j$ defined as:

$$\mathbf{Y}_j = \mathbf{X}\mathbf{a}_j \tag{2}$$

where $\mathbf{a}_j$ is the j-th column of the loading matrix $\mathbf{A}$ of $\mathbf{\Sigma}$ and represents the contribution of each variable in $\mathbf{X}$ to the j-th principal component $\mathbf{Y}_j$.

In practice, the component scores related to components $q + 1$ to m represent the Euclidean distances alongside the axes of the corresponding orthogonal vectors to a q-dimensional linear subspace. The first q loadings are chosen so that this subspace contains the highest proportion of the total variance of the data points. In essence, PCA seeks a linear combination of variables such that the maximum variance is extracted from the variables. The first q components are described by:

$$\mathbf{Y} = \mathbf{X}\mathbf{A}_q \tag{3}$$

where $\mathbf{Y}$ is the score matrix, $\mathbf{A}_q$ is the loading matrix with the only q columns of $\mathbf{A}$. [13] demonstrates that the best (least squares) rank q approximation to $\mathbf{X}$ is $\mathbf{X}\mathbf{A}_q\mathbf{A}_q^t$ and the residual matrix $\mathbf{S}$ can be defined as:

$$\mathbf{S} = \mathbf{X} - \mathbf{X}\mathbf{A}_q\mathbf{A}_q^t = \mathbf{X}\mathbf{A}_{-q}\mathbf{A}_{-q}^t \tag{4}$$

where $\mathbf{A}_{-q}$ is the loading matrix with the first q columns removed. In the case of application of PCA to spatially distributed data, the underlying implicit hypothesis is that the variance and covariance structure of the process is constant throughout the geographical area under investigation. This assumption is obviously not realistic

([10]). Therefore, it is necessary to relax this hypothesis, to consider in some way the spatial effects in the definition of the principal components.

A first appropriate technique for PCA for spatial data is represented by Geographically Weighted Principal Component Analysis, (GWPCA, [5], [9]). The equation (1) can be generalized to the case of GWPCA as ([10]):

$$\mathbf{A}(u_i, v_i)\mathbf{\Lambda}(u_i, v_i)\mathbf{A}(u_i, v_i)^t = \mathbf{\Sigma}(u_i, v_i) \tag{5}$$

where $\mathbf{\Lambda}(u_i, v_i)$ is the diagonal matrix of local eigenvalues, $\mathbf{A}(u_i, v_i)$ is the corresponding matrix of local eigenvectors, $\mathbf{\Sigma}(u_i, v_i)$ is the local variance-covariance matrix, and (u_i, v_i) are the coordinates of spatial unit i.

The output of GWPCA consists in different loadings and component scores defined for each spatial unit. In practice, GWPCA defines completely different index for each spatial unit as function of distinct loadings. This produces remarkable difficulties in the interpretation of the results.

To simplify the reading of the phenomena, in this paper we propose to apply simulated annealing (SA) algorithm to PCA to identify groups of spatial units that are supposed to share the same eigenvectors (i.e., the same composite indicators). This approach was introduced by [19] and improved by [20] for the analysis of economic growth. The main idea of this framework is that the appropriate treatment of spatial heterogeneity is substantially equivalent to partition an area in groups of geographical zones not necessarily conterminous that have similar component scores. Following this methodology, the output is not represented by different loadings for each spatial unit as in the case of GWPCA, but distinct loadings for every groups of regions identified by SA.

SA is a stochastic relaxation algorithm that was originally introduced in statistical mechanics by [15] and [14]. [7] observes that a spatial combinatorial optimization problem might be described through a Markov Random Field (MRF). The probability measure of a MRF using Gibbs distribution is defined through the energy function $U(\mathbf{X}, \mathbf{k})$, that represent in our algorithm the objective function to be minimized, and a control parameter, T (see [6]; [19]). $U(\mathbf{X}, \mathbf{k})$ depends on observed data $\mathbf{X}$, and the label vector $\mathbf{k} = (k_1, k_2, \dots, k_i, \dots, k_n)$, which categorizes the heterogeneous zones, identifying clusters of regions. $U(\mathbf{X}, \mathbf{k})$ is defined by considering two different effects: a measure of the goodness of fit of the model, and a proximity constraint that describes the extent of aggregation of the spatial units. In particular, at the l-th iteration of the procedure, the energy function is defined as:

$$U(\mathbf{X}, \mathbf{k}) = \beta \sum_{i=1}^{n} I_i \; - (1-\beta) \sum_{r=1}^{n} \sum_{s=1}^{n} \mathbf{c}_{rs} \mathbf{1}_{(k(j)_r = k(j)_s)} \tag{6}$$

where the first part in the right-hand-side is the interaction term, with $I_i = \sum_{k=q+1}^{p} s_i^2$, with s_i the entry of the matrix of the residual matrix $\mathbf{S}$ defined by equation (4); while the second one is the penalty term defined through a Potts model (see [19]). Specifically, $\mathbf{c}_{rs}$ is the element (r, s) of a binary contiguity matrix, $\mathbf{1}_{(k(j)_r = k(j)_s)}$ is the indicator function of the event and $k(j)_r = k(j)_s$, and $(1-\beta)$ is a parameter that discourages configurations with not conterminous units. The parameter $(1-\beta)$ is chosen by the researcher and models the importance of the proximity of the spatial units. Note that the two parts of the energy function (6) are

balanced with complementary weight. At the initial value of control parameter T_0, each unit i, is randomly classified as $k_{i,0}$, where $k_{i,0} \in \{1,2,\dots,K\}$ with K is the number of clusters. This step defines the initial configuration S_0. At the $(l+1)$-th iteration, given a current configuration S_l, a different configuration $S_l \neq S_{l+1}$ is randomly chosen, defining a new energy function $U(S_{l+1})$ the is compared with the previous one $U(S_l)$. The old configuration S_l is substituted by the new S_{l+1} in accordance to the probability:

$$Pr_{l,l+1} = max\left\{1, exp\left(-\frac{U(S_{l+1})-U(S_l)}{T_l}\right)\right\} \quad (7)$$

It is worth noting that probability (7) allows to avoid entrapments in local minimum, by defining positive probability for the change of configuration also when the objective function $U(S)$ increases. In essence, more likely patterns (i.e. configurations with lower states of energy) are always accepted, but it is also possible to accept also poorer configurations.

3. Empirical evidence

The proposed methodology is applied to define a deprivation index for Italian provinces. Data set derive from the 15th Population and Housing Census (2011) by Italian National Statistical Institute.

In this paper, a set of ten variables is adopted to build an area-based indicator of material deprivation. The choice of variables has been carried on according to the definition of deprivation index by [17] that suggest choosing a small set of variables able to capture socio-economic deprivation and assist policy makers in a wide set of decisions, for example, public health and tracking inequalities. The variables cover both economic and social domains. Income, educational attainment (proportion of people without high school diploma, School), and employment (Empl) are considered together with social conditions, as the proportion of people living alone (Unip), the percentage of separated, widowed, or divorced people (SVD), and the proportion of single parent families living in each area (Sin_Par). Furthermore, to have a better definition of the deprivation, we include other indicators, and use some of the variables proposed by [11] to assess the level of material deprivation: lack of car possession among resident families (Car), percentage of families living in house of property (Hou), and available surface in residence houses per person (Sqm) are added to assess the level of material deprivation. Moreover, the percentage of foreigners living in the Province (Frg) is considered as an additional variable, particularly to evaluate situation of social exclusion.

PCA is performed on Italian provinces and four components are selected which capture 80% of the variance in the data set. In Table 1, the loadings of the first four components are reported.

The eigenvector corresponding to the first component is characterized by a dominance of the economic variables: employment, income, and house dimension. These are negatively correlated to deprivation. Lack of car possession is positively

linked to deprivation, displaying that owning a car decreases the level of material deprivation. On the other side, social variables tend to be lower in magnitude. In the second component, the picture is substantially different with social variables higher in magnitude, and negatively correlated to the deprivation level. Interestingly, the percentage of foreign people reverses its sign, being negative in the first component and positive in the second.

Table 1: Loadings for first 4 components of global PCA.

	PC 1	PC 2	PC 3	PC 4
Empl	-0.500	0.087	-0.101	0.141
Income	-0.307	0.245	-0.316	-0.062
School	0.110	-0.187	0.083	0.879
Sin_Par	-0.125	-0.602	0.027	-0.230
SVD	-0.161	-0.595	-0.100	-0.193
Unip	-0.294	-0.285	-0.428	0.273
Car	0.386	-0.156	-0.375	0.123
Frg	-0.440	0.238	-0.119	0.105
Hou	-0.153	-0.106	0.686	0.114
Sqm	-0.387	-0.085	0.253	0.024
Eigenvalues	1.84	1.50	1.20	1.01
Proportion of variance	0.34	0.22	0.14	0.10
Cumulative variance	0.34	0.56	0.70	0.80

In standard PCA, homogeneity across space is assumed, and the same set of loadings may be used to derive an indicator of deprivation for the whole Country. Nevertheless, spatial heterogeneity could characterize the structure of the variance-covariance matrix. Therefore, the hypothesis of spatial homogeneity could be relaxed, allowing loadings to change according to different spatial configurations.

To avoid potential drawbacks of spatial heterogeneity in PCA, SA is adopted for identifying clusters of spatial units.

In this paper we identify three different clusters, and this spatial configuration is considered for further analysis. The selected combination produces an improvement in the proportion of explained variance when compared to the standard PCA. Finally, a level of $(1-\beta)$=0.3 is chosen, and 4 components are retained for all configurations.

In Figure 1, the groups are mapped, where white denotes the first cluster, light grey the second group, and dark grey the third regime.

The first group (i.e., white) is mainly composed by provinces in the North-East of the Peninsula and some part of Tuscany. Provinces closer to Alps and the Centre - especially in the Apennine mountains - compose the second groups (i.e., light gray), while the third regime (i.e., dark grey) characterizes mostly the Southern part of the Country and few Provinces of the North (e.g. Turin).

As expected, the three clusters show substantial differences in terms of their indicators structures. The eigenvectors of the first component of the three groups are shown in Table 2. Particularly interesting is the impact of the social variables. While in the first group social variables contribute positively to the level of deprivation, in the other spatial clusters the effects (i.e., Sin_Par, SVD, Unip) on the deprivation is negative. Other significant differences in the loadings structure can be found in the

heterogeneous effect on deprivation of school attainment and the different magnitude of employment in the diverse regimes.

Figure 1: Clusters of spatial units obtained by Simulated Annealing.

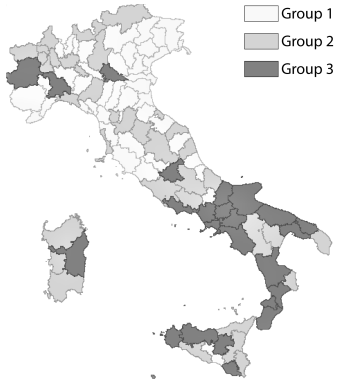

Table 2: Loadings of first component for each group obtained from Simulated Annealing.

	Group 1	Group 2	Group 3
Empl	-0.318	-0.462	-0.360
Income	-0.293	-0.378	-0.007
School	0.283	0.192	0.037
Sin_Par	0.429	-0.245	-0.365
SVD	0.432	-0.300	-0.379
Unip	0.243	-0.349	-0.415
Car	0.380	0.330	0.236
Frg	-0.366	-0.351	-0.280
Hou	-0.081	0.240	-0.320
Sqm	-0.135	-0.211	-0.429
Proportion of Variance	0.40	0.35	0.39

4. Conclusion

In this paper we propose a method for considering spatial heterogeneity in PCA. In fact, when dealing with geographically distributed data, the application of the classical framework of PCA could be misleading and lead to incorrect results. To overcome this drawback, the proposed method extends to PCA, the SA algorithm introduced by [19] for analyzing regional economic growth.

Applying SA to PCA let to highlight different structures of the multivariate phenomenon taking into account the presence of spatial heterogeneity. Results show the differences in the computing the composite indicator in each cluster. Especially the effect of social variables varies from first to second and third groups and a substantial difference of the North provinces with the rest of the Country is highly evident. However, results of SA help policy maker in the interpretation of the global

phenomenon, improving interpretability of the indicator levels while considering different spatial regimes.

References

1. Anselin, L.: Spatial econometrics: Methods and models. Kluwer Academic Publishers, Dordrecht (1988).
2. Decanq, K., Lugo, M.A.: Weights in multidimensional indices of wellbeing: An overview. Econom. Rev. 32: 7-34 (2013).
3. Demšar, U., Harris, P., Brunsdon, C., Fotheringham, A.S., McLoone, S.: Principal component analysis on spatial data: An overview. Ann. Assoc. Am. Geogr. 103: 106-128 (2013).
4. De Muro, P., Mazziotta, M., Pareto, A.: Composite indices of development and poverty: An application to MDGs. Soc. Indic. Res. 104: 1-18 (2011).
5. Fotheringham, A.S., Brunsdon, C., Charlton, M.: Geographically weighted regression - The analysis of spatially varying relationships. Chichester, UK: Wiley (2002).
6. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. 6, 721-741 (1984).
7. Geman, D., Geman, S., Graffigne, C., Dong, P.: Boundary detection by constrained optimization. IEEE Trans. Pattern Anal. Mach. Intell. 12, 609–628 (1990).
8. Goodchild, M.F.: Geographical data modeling. Comput. Geosci. 18: 401–408 (1992).
9. Harris P., Brunsdon C., Charlton. M. (2011). Geographically weighted principal components analysis. Int. J. Geogr. Inf. Sci. 25: 1717–36.
10. Harris, P., Clarke, A., Juggins, S., Brunsdon, C., Charlton M.: Enhancements to a geographically weighted principal component analysis in the context of an application to an environmental data set. Geogr. Anal. 47: 146-172 (2015).
11. Havard, S., Deguen, S., Bodin, J., Louis, K., Laurent, O., Bard, D.: A small-area index of socioeconomic deprivation to capture health inequalities in France. Soc. Sci. Med. 67: 2007-2016 (2008).
12. Hotelling, H.: Analysis of a complex of statistical variables into principal components. Journal of Educ. Psychol. 24: 417-441 (1933).
13. Jolliffe, I.T.: Principal component analysis. Springer (2002).
14. Kirkpatrik, S., Gelatt, Jr C.D., Vecchi, M.P.: Optimization by simulated annealing. Sci. 220: 671-680 (1983).
15. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equations of state calculations by fast computing machines. J. Chem. Phys. 21, 1087–1092 (1953).
16. OECD: OECD Core set of indicators for environmental performance reviews. Environ. Monogr. 83 (1993).
17. Pampalon, R., Raymond, G.: A deprivation index for health and welfare planning in Quebec. Chronic Dis. Can. 21:104-13 (2000).
18. Pampalon, R., Hamel, D., Gamache, P.: Health inequalities, deprivation, immigration and aboriginality in Canada: A geographic perspective. Can. J. Public Health 101:470-4 (2010).
19. Postiglione, P., Andreano, M.S., Benedetti, R.: Using constrained optimization for the identification of convergence clubs. Comput. Econ. 42: 151-174. (2013).
20. Postiglione, P., Andreano, M.S., Benedetti, R.: Spatial clusters in EU productivity growth. Growth Chang. 48: 40-60 (2017).
21. Townsend, P.: Deprivation. J. So. Policy 16: 125-146 (1987).