The Factorial Minimum Spanning Tree as a Reference for a Synthetic Index of Complex Phenomena

Sergio Scippacercola
Dipartimento di Informatica e Sistemistica
Università degli Studi di Napoli "Federico II" – Napoli - E-mail: SS@UNINA.IT

Abstract: *A method is herewith proposed which after analyzing the data matrix in principal components, searches the subspace representing the initial configuration of the inter-point distances by eliminating the "background noise" present in the data considered. In that subspace a factorial minimum spanning tree is built which becomes a reference structure for the design of a synthetic index of the phenomenon analyzed. The initial configuration is compared with the final one by means of adequate "quality indicators". The validity of the method is confirmed by results achieved in various applications to real data.*

Keywords: Principal Components Analysis, Minimum Spanning Tree, Factorial Score, Statistical Synthetic Index.

1. Introduction

Complex phenomena and concepts such as the economic situation, the IQ, health, the standard of living, political ideas, etc. are difficult to measure statistically both as concerns the choice of the method to be adopted and the variables to be applied. For building the synthetic indices of complex phenomena various statistical methods have been proposed in the literature (Rizzi, 1988; Fraire, 1989) but only factorial methods may be considered suitable to synthetize multivariate complex phenomena since they reduce the "background noise" to the minimum and highlight the significant part of the information provided in the data analysed. It is hereby pointed out that to grade statistical multivariate units no method is currently at disposal allowing a total order (Scippacercola, 1997). Unlike other proposals, this paper is aimed at combining both the properties of factorial subspaces and those of the Minimum Spanning Tree Algorithm (Cormen, et. al., 2001; Zhu Mei-Jie, et. al., 2005), to obtain a synthetic tree representing the phenomenon on the basis of the inter-distances between the statistical units to be taken into account. In particular, for ordering the *n* statistical units, the maximum path of the Minimum Spanning Tree (*MST*) for the images of the statistical units in a multi-factorial subspace is taken as a reference. In the second paragraph it is briefly hinted how to deduce the coordinates of the statistical units in a multi-factorial subspace through a Principal Component Analysis. These coordinates are the basis for the design of a $FMST_q$ $(q \leq p)$*(Factorial MST)*. The subsequent linearization of the $FMST_q$ determines the synthetic index whose quality is assessed by three distinct measurements as proposed (Par. 3). An application to real data is then proposed (Par. 4) followed by our conclusions (Par.5) with future prospects for the method illustrated and computational observations.

2. Factor scoring for the design of a synthetic index of a complex phenomenon

Let $X_{n,p}$ be a matrix of data, $\Im$ the set of *n* statistical units and $\boldsymbol{x}_j$ $(j=1,2, ...,p)$ the quantitative variables. Let $\boldsymbol{Z}=\{ z_{ij} \}$ $(i=1,2, ..., n; j=1,2, ..., p)$ be the standardized matrix of $\boldsymbol{X}$. The target of this paper is the design of an index $\boldsymbol{I}_i$ $(i=1,2, ...,n)$ of the multivariate complex phenomenon considered. The purpose of the Principal Component Analysis (Jollife, 2002) is to reduce the dimensions of the phenomenon by analysing a lower number of variables (the Principal Components) vis-a-vis the number of starting variables. The *factor score* of a statistical unit is given by the coordinate the latter assumes on a specific axis. The *k-th* principal component is defined by a linear combination $\boldsymbol{Y_k} = \boldsymbol{Za_k}$ for standardized variables. The factor score for the *i-th* unit is given by:

$$y_{ik} = a_{k1}z_{i1} + a_{k1}z_{i1} + ... + a_{kp}z_{ip} \quad (i=1,\dots,n\ ;\ k=1,\dots,p) \qquad (1).$$

In the $\Re_n$ space of the units, factor scores in (1) represent the coordinates on the first factor axis. The score of a factor is given by the sum of the contributions provided by the different variables. Factor scores may be treated as new variables to be inserted in Cartesian planes identified by factor pairs. The Principal Component Analysis is also used as a method for summing up initial data by means of one or more synthetic indicators.

3. A synthetic index obtained by means of a Factorial *MST*

From a geometric point of view, each row of the matrix $\boldsymbol{Z}$ is matched to a vector point in the *p*-dimensional subspace. To reduce the $\boldsymbol{Z}$ dimensions and the "background noise" present in the data, subspaces of $\boldsymbol{Z}$ images in $\Re_n$ and in $\Re_p$ are first identified by means of a Principal Component Analysis. Inter-point distances between images of the *n* statistical units in each factorial plane are used for the building *of a* $FMST_q$[1] in the *q-th* same factorial plane. In the $FMST_q$, among others, an *M* path of maximum length is searched by means of a suitable algorithm. The *M* maximum path and the similarity relations between adjacent statistical unit on *M*, are the starting point for the determination of the synthetic index values. Let α and ω be the terminal knots of *M*, let ν be one of the vertices of $\Im$ ($\nu \in \Im$). The $g(\alpha,\nu)$ distance, via MST_q, from α to a generic vertex ν is the sum of the lengths of the *m* consecutive segments *d'* starting from α to reach ν through the MST_q:

$$g(\alpha,\nu) \quad = \quad \sum_{i=1}^{m} d'_{i,i+1} \qquad (2).$$

The statistical units which are on the side branches of *M* are projected on *M* (Scippacercola, 1997). In this way the statistical units on the side branches of the maximum path are brought onto the same *M*. The values of all the graduations obtained for each statistical unit may be intended as a synthetic index:

$$I_i = g(\alpha,\nu) \qquad (3).$$

In this way, two statistical units which are adjacent in the $\Re_p$ and in the factorial plane, obtain I_i values which are almost equal. On the contrary two statistical units which are distant in $\Re_p$, obtain very different results. The method and the relative algorithm are mainly applied to the design of synthetic values and in particular of statistical synthetic indices of complex real phenomena. To assess the quality of the index $\boldsymbol{I_i}$ and to choose the best representation in terms of inter-point distances in the $\boldsymbol{X}$ *matrix*, some special indices are here proposed. Let $\boldsymbol{G_{n,1}}$ be the vector containing the configuration in the MST_q of the *n* points positioned in the same order as $\boldsymbol{X}$. It is possible to assess the similarity of the two configurations by means of the following indices:

1) the *RV* coefficient (Robert, Escoufier, 1976)

$$I_E = \frac{tr[(\boldsymbol{XX'})(\boldsymbol{GG'})]}{[tr(\boldsymbol{XX'})^2\, tr(\boldsymbol{GG'})^2]^{1/2}} \quad (0 \le I_E \le 1)$$

. The RV coefficient measures the similarity between the relative positions of the *n* points in the subspace generated by the columns of $\boldsymbol{X}$ and $\boldsymbol{G}$.

2) the procrustean index (Mardia et al., 1979) adequately standardized:

[1] Let $\boldsymbol{D} = \{d_{ij}\}$ ($\boldsymbol{D} \subseteq \Re_1$) be the matrix of the interpoint distances in a factorial plane, the research of the Minimum Spanning Tree (Kruskal, 1956; Prim, 1957; Gower, Ross, 1969) on the set $\boldsymbol{D}$ of distances generates a subset $\boldsymbol{D'} \subseteq \boldsymbol{D}$ ($\boldsymbol{D'} = \{d'_{ij}\}$) meeting the requirements of ultrametric axioms thus allowing the design of the MST_q.

$I_p = \dfrac{2\,tr\,(\boldsymbol{\Gamma}^{1/2})}{[tr(\boldsymbol{XX'})^2 + tr(\boldsymbol{GG'})^2]^{1/2}}$ $(0 \le I_p \le 1)$, where $\boldsymbol{\Gamma}$ is the matrix of the proper values of (***X'GG'X***). Unlike the coefficient *RV*, the procrustean index measures the similarity between two configurations in case of rotational translation and reflection of the configuration.

3) the congruence coefficient (Borg, Lingoes, 1987):

$I_c = \dfrac{\sum_{i=1,k} d_{iX}\, d_{iG}}{[\sum_{i=1,k} d_{iX}^2\, d_{iG}^2]^{1/2}}$ $(0 \le I_c \le 1)$, where d_{iX} is the i-th distance of ***X*** and d_{iG} is the i-th distance of ***G***. The coefficient measures the similarity for each pair of points in the configuration in terms of interpoint distances. If the indices I_E, I_p, I_c are close to zero the similarity of the configurations is low and the synthetic index is scarcely representative. Differently if the three indices I_E, I_p, I_c are close to one their similarity is high. The three indices mentioned above compare two configurations of different dimensions also. The congruence coefficient, however, is especially relevant for types of analysis where the inter-distances and not the projections, are considered relevant for the identification of a synthetic index. These indices provide information on the similarity and order of the configurations and so give a measurement of the quality of the transformation carried out on computing the synthetic index.

4. Application of the methodology

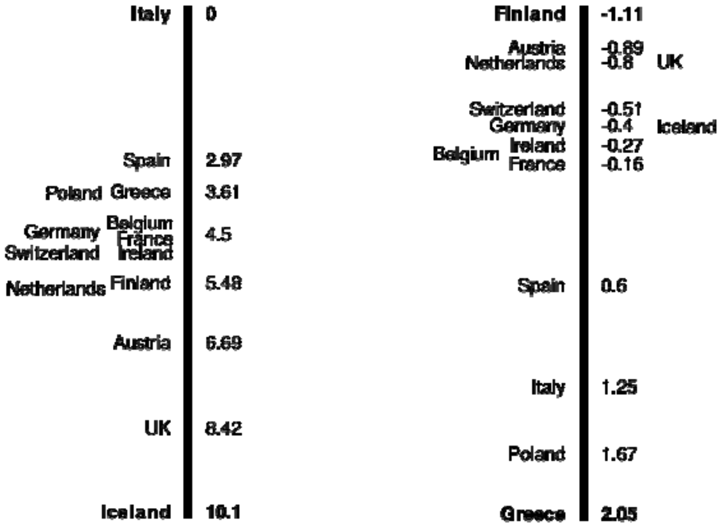
The method here proposed has been applied to a set of multidimensional data to identify a synthetic index for “the standard of living and social protection” in fourteen European Countries in 2006 (Eurostat, 2008) by employing five distinct variables (% of household expenditure for education, health, food and non-alcoholic drinks, for clothing, entertainment and culture). Note in Tab. 1 that the $FMST_5$ (Fig. 1 on the left) maintains 78% only (Index $\boldsymbol{I_c}$ =0.78) of the inter-distances in the initial subspace (axes from 1 to 5) as “being rich in background noise”. The $FMST_1$ on the first axis (Fig. 1 in the middle), though with a variance of 33,5%, maintains the distances between points at 86% and so well represents the synthetic index searched (Table in the Fig. 1 on the right). By also employing more axes, the representation would be scarcely significant as the other indices ($\boldsymbol{I_e}$, $\boldsymbol{I_p}$) reach values less relevant than the ones for the first axis. The synthetic index thus designed must be read in terms of distance between points. Germany and Iceland having an index of 0,70 and 0,71 respectively are to be considered very similar as concerns the standards of living and social protection. Differently Finland is markedly different from Greece as concerns the aspect under examination.

5. Conclusions

The Minimum Spanning Tree (MST) Method is one of the best-known and important Algorithms applied in multivariate statistical analysis. In this paper a new and original method for the computation of a MST in a factorial plane is proposed. This method ensures that the information about “*the distances between points of the original cloud are maintained also after having removed the background noise present in the data*”. The Factorial Minimum Spanning Tree is an Algorithm independent on the Principal Component Analysis. The Principal Component Analysis is introduced only to the purpose of identifying which subspace is fit for a Factorial MST to become a reference tree for the design of a synthetic index of the complex phenomenon under examination. The initial configuration is compared with any other possible configuration by means of adequate similarity measurements which become “quality indicators of the synthetic index”. The Algorithm appears more robust and significant than other indices directly designed by applying the proportions method or the Principal Components one. Its validity is confirmed by the results yielded by various applications to real data. The synthetic index illustrated can be easily computed and is applicable to any multidimensional phenomenon to be synthetized for submittal to corporate decision-makers.

SUBSPACES	VARIANCE EXPLAINED	I_e	I_p	I_c
I-II-III-IV-V	100,0%	0.20	0.22	**0.78**
I	33,5%	0.66	0.43	**0.86**
II	30,3%	0.60	0.41	**0.78**
III	16,1%	0.32	0.30	**0.70**
I-II	63,8%	0.54	0.45	**0.85**
I-II-III	79,9%	0.36	0.42	**0.52**

Tab. 1 – Values of the quality indices for the interpoint distances of Z and the FMST$_q$ (The third factorial axis, even if with proper value lower than one, is here given only for reference)



Country	**Synthetic index I_i**
Finland	0,00000
Austria	0,21381
UK	0,30418
Netherlands	0,31200
Switzerland	0,60212
Germany	0,70402
Iceland	0,71091
Ireland	0,83358
Belgium	0,88332
France	0,94441
Spain	1,71449
Italy	2,36286
Poland	2,78099
Greece	3,16282

Fig. 1 - A FMST$_5$ with a "background noise" (chart on the left). The FMST$_1$ (chart in the middle) from which the synthetic index (chart on the right) is obtained for the standard of living and social protection in fourteen European Countries in year 2006.

Bibliography

Borg L., Lingoes J. (1987), *Multidimensional Similarity Structure Analysis*, Springer-Verlag, 58-64.

Cormen T. H., Leiserson C. L., Rivest R. L., Stein R. (2001) *Introduction to Algorithms*, Second Edition, MIT Press and McGraw-Hill.

Eurostat (2008), *Living conditions and social protection*, website http://epp.eurostat.ec.europa. .eu /portal/.

Fraire M (1989), Problemi e metodologie statistiche di misurazione di fenomeni complessi tramite indicatori e indici sintetici, *Statistica*, n. 2, pp. 245-263.

Gower J.C., J.S. Ross (1969), Minimum Spanning Trees and Single Linkage Cluster Analysis, *Appl. Stat.*,18, pp.54-64.

Jollife I. T (2002), *Principal Components Analysis*, Springer Series in Statistics.

Kruskal J.B. (1956), On the shortest spanning subtree of a graph and the travelling salesman problem, *Pro. of the American Math. Soc.*, 7, pp. 48-50.

Mardia K. V., Kent J. T., Bibby J. M. (1979), *Multivariate Analysis*, Ac. Press, London.

Prim R.C. (1957), Shortest Connection network and some generalizations, *Bell yst. Tech. Journal*, 36, pp. 1389-1401.

Robert P., Escoufier Y. (1976), A unifying tool for linear multivariate statistical methods, the RV coefficient, *Applied Statistics*, 25.

Rizzi A. (1988), Un metodo di graduazione di più unità statistiche, *Rivista di Statistica applicata*, vol. 21, n. 1, pp. 49-64.

Scippacercola S. (1997), Inter-point Distances in the Multivariate Data Ordering, *Statistica Applicata*, 10, 1,73-83.

Zhu Mei-Jie, Gui-Wu Hu, Qi-Lun Zheng, Hong Peng (2005), Multiple sequence alignment using minimum spanning tree, Proc. International Conference on Machine Learning and Cybernetics, 6, 3352-2256.