

Measuring Nonlinearity in Data Analysis (*)

Misurare la Nonlinearità nell'Analisi dei Dati

Marica Manisera

Università degli Studi di Brescia, C.da S. Chiara 50, e-mail: manisera@eco.unibs.it

Riassunto: Le indagini sociali che raccolgono microdati tramite questionari portano di solito ad analisi con numerose variabili qualitative che non possono essere considerate metriche o lineari e che, pertanto, consigliano il ricorso a tecniche nonlineari. Nel presente contributo si propone un semplice indice normalizzato utile per misurare il grado di nonlinearità (da intendere come allontanamento dalla linearità) presente nelle variabili. Valori elevati dell'indice segnalano una forte nonlinearità nei dati e, quindi, la necessità di trattare i dati con tecniche nonlineari che ne tengano conto. Valori bassi dell'indice, invece, rivelano situazioni in cui l'impiego di tecniche lineari, più semplici e parsimoniose, conduce a risultati analoghi a quelli ottenibili con tecniche nonlineari. Viene anche presentata un'applicazione con dati reali riguardanti la *job satisfaction* di un campione di lavoratori operanti nel settore dei servizi sociali in Italia.

Keywords: nonlinearity, categorical variables, quantifications

1. Introduction

Most of the multivariate data analysis techniques can be described as *linear*, because they refer to linear transformations of the measured variables, or they look for linear structure in the data, or they replace one high-dimensional set of coordinates by another which has a linear relation with it (Krzanowski & Marriott, 1994, p. 195). Although linear techniques allow one to solve various multivariate problems, a number of gaps remain, and attention has to be paid to nonlinear techniques. They enable one to deal with the nonlinearity structure that may exist in the data. However, when nonlinearity in the data is negligible, a linear approach could be preferable, because it is usually simpler and more parsimonious. For this reason, it is important to understand if there are nonlinearities worth taking into account.

The present work proposes an index of the nonlinearity in the data, in order to understand when a nonlinear approach is more appropriate than a linear one. Once specified what we mean by nonlinearity, the *NL* index will be introduced and computed with reference to a multivariate real dataset.

2. Nonlinearity

Social and behavioural science research often results in data that are non-numerical, with measurements recorded on scales having an uncertain unit of measurement, uncertain zero point, unknown relationships and distances among the different categories (especially when these scales include categories such as “*not applicable*”,

(*) This paper is financially supported by MIUR funds PRIN 2004, prot. 2004130924-004.

“neutral”, or “I don’t know”). Collecting data often involves the use of questionnaires with several items referring to the different aspects of the concept being measured. Responses indicate the degree of agreement with each statement, with higher scores reflecting greater agreement. Ordinal variables typically stand out, and the persons are described in a limited number of categories. Attention has to be paid to the treatment of ordinal variables, because we cannot assume *a priori* that the data are numerical; this kind of variables leads to nonlinear data, in the sense specified, among others, by Gifi (1990) and Candel (2001): the distances between the categories are not equal, although it is possible to code them metrically.

This study considers nonlinearity of variables in terms of distance from linearity, that is defined by two properties: (1) the ordering of categories, and (2) the equal distance between them. Therefore, we define nonlinearity as distance from linearity by considering the situations in which one or both aspects of linearity fail. Because the focus is here on ordinal variables, the variable’s nonlinearity increases the more its categories are not equally spaced.

In the literature, several methods have been proposed to transform categorical variables into quantitative variables (among others: Torgerson, 1958; Herzog, 1974; Van Rijckevorsel, 1987; Gifi, 1990). They usually assign numeric values called *quantifications* to the original categories of the variables. When dealing with ordinal variables, it is sensible to compute *ordinal* quantifications, respectful of the original ordering of the categories but not necessarily equally spaced. In fact, the fixed *a priori* order of the categories should not be taken to imply that the differences between numerical labels of the categories should be respected.

In order to formalize the problem, the following notation is introduced. A categorical variable, defined on n objects, has k categories, usually given by the first k integers $r = 1, 2, \dots, k$ (k is the number of distinct values h_i , $i = 1, \dots, n$, in the vector $\mathbf{h}$ containing the values of the variable observed on the n objects). Variable $\mathbf{h}$ defines a binary indicator $n \times k$ matrix $\mathbf{G}$, whose element G_{ir} is 1 if $h_i = r$ and 0 if $h_i \neq r$.

In the transformation process aimed at quantifying the categorical variable, each category obtains a *quantification*. The $k \times 1$ vector collecting the quantifications of the k original categories is denoted by $\mathbf{y}$, then the transformed variable $\mathbf{q} = \phi(\mathbf{h})$ can be written as $\mathbf{G}\mathbf{y}$. Without losing generality, we consider standardized transformed variables: $\mathbf{1}'\mathbf{q} = 0$ and $\mathbf{q}'\mathbf{q} = \mathbf{y}'\mathbf{D}\mathbf{y} = n$, where $\mathbf{1}$ is the $n \times 1$ vector of 1’s, and $\mathbf{D} = \mathbf{G}'\mathbf{G}$ is the $k \times k$ diagonal matrix containing the marginal frequencies n_r of the categories of the variable, i.e., the column sums of $\mathbf{G}$.

Information in the original categorical data is retained in the quantifications, depending upon the form of the function ϕ chosen to transform the variables. Because we are dealing with ordinal variable, we are interested in monotonic non-decreasing transformations that preserve grouping and ordering information, with reference to the original categories. The resulting nonlinear quantifications will be compared with those coming from a linear transformation, that maintains not only grouping and ordering, but also interval information. We can assume as a linear transformation the standardization of the original category values.

3. The *NL* index

The proposed index of nonlinearity can be constructed starting from the definition of the following distance index (see, for example, Zani 2000 for other distance indices):

$$\delta = n^{-1}(\mathbf{y}_{NL} - \mathbf{y}_L)' \mathbf{D}(\mathbf{y}_{NL} - \mathbf{y}_L) \quad (1)$$

where $\mathbf{y}_{NL}$ and $\mathbf{y}_L$ are the quantifications obtained by the k categories of the variable when it is transformed according to the nonlinear and the linear transformations, respectively. δ_{NL} is the *average squared Euclidean distance* between $\mathbf{y}_{NL}$ and $\mathbf{y}_L$, weighted by the marginal frequencies. It can be called *weighted average squared Euclidean distance* (for example, Gifi, 1990, p. 261 uses the same distance index but in the context of Multidimensional Scaling). It can be shown (Manisera, 2005) that

$$\delta = n^{-1}(\mathbf{y}_{NL} - \mathbf{y}_L)' \mathbf{D}(\mathbf{y}_{NL} - \mathbf{y}_L) = 2(1 - \rho) \quad (2)$$

with $\rho = \rho(\mathbf{y}_{NL}, \mathbf{y}_L)$ being the Pearson linear correlation coefficient between $\mathbf{y}_{NL}$ and $\mathbf{y}_L$. Note that $0 \leq \rho \leq +1$, because both the linear and the ordinal transformations are non-decreasing. Since $\max(\delta) = 2$, we define the *normalized index of nonlinearity* as

$$NL = \delta / 2, \quad \text{with } 0 \leq NL \leq 1. \quad (3)$$

NL attains its minimum when $\mathbf{y}_{NL} = \mathbf{y}_L$: when it occurs, the linear and the nonlinear transformations (versus the original categories) overlap, there are no nonlinearities, then $\rho = 1$, $\delta = 0$ and $NL = 0$. On the contrary, when $\rho = 0$ and $\delta = 2$, NL reaches its maximum. Larger values of NL characterize transformations associated with “stronger” nonlinearities. Because ordinal quantifications are ordered by definition, values of NL close to 1 indicate that the original categories are not equally spaced and the linear restriction stretches too much the evidence.

In a multivariate analysis, it is possible to compute one NL_j index, $j = 1, \dots, m$, for each of the m variables in the dataset and also a *mean nonlinearity* index $\overline{NL}$, averaging over the m variables. Obviously, $0 \leq \overline{NL} \leq 1$. $\overline{NL}$ measures the nonlinearity under a global point of view, that is considering all m variables at the same time. Examining $\overline{NL}$ together with the m NL_j indices gives a complete picture of the nonlinearities that characterize the analysed variables.

4. An application

With reference to a real dataset concerning the assessment of job satisfaction of 2063 workers in the Italian social service sector (Borzaga & Musella, 2003), we computed the NL indices (%) for the 13 satisfaction variables (each with $k=7$ categories) in the dataset. We obtained the ordinal quantifications by *optimal scaling* in nonlinear *Principal Component Analysis* (PCA): monotonic non-decreasing transformations led to numeric quantifications maintaining ordering information of the original categories but not restricted to be equally spaced (Gifi, 1990).

Figure 1 shows that most variables are associated with small values of NL_j . Some variables have ordinal quantifications that naturally satisfy the equal distance between the quantified categories (see, for example, Figure 2a). The largest $NL_j\%$ are associated with variables *Certainty* (22.6%) and *Colleagues* (12.1%), that show the most nonlinear transformations (Figure 2b-c).

Figure 1: NL_j % indices for the 13 variables of job satisfaction

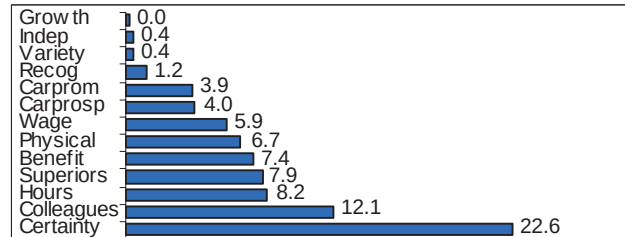
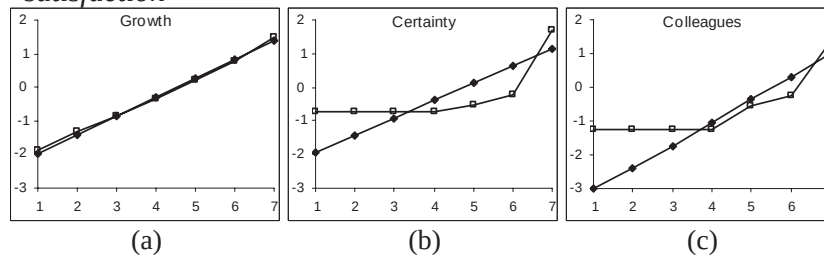


Figure 2: Nonlinear ($\square$) and linear ($\blacklozenge$) transformations for 3 variables of job satisfaction



$\overline{NL}$ equals 6.2%: this means that, except for few variables, respondents considered the categories of the variables as equally spaced. This is reflected in no relevant differences between the solutions obtained from a nonlinear PCA and a linear PCA, with respect to the global solution. However, the usefulness of nonlinear PCA remains valid: we think that performing such analysis and computing the nonlinearity indices are the basis for using the linear approach correctly: a linear PCA is more justified and useful when the nonlinear approach and the proposed indices indicate that a linear analysis is feasible.

References

- Borzaga C. & Musella M. (Eds.) (2003) *Produttività ed efficienza nelle organizzazioni nonprofit*, Edizioni 31, Trento.
- Candel M. J. J. M. (2001) Recovering the metric structure in ordinal data: linear versus nonlinear Principal Component Analysis, *Quality & Quantity*, 35, 91-105.
- Herzel A. (1974) Un criterio di quantificazione. Aspetti statistici, *Metron*, 32, 3-54.
- Gifi A. (1990) *Nonlinear multivariate analysis*, Wiley, Chichester.
- Krzanowski W. J. & Marriott F. H. C. (1994) *Multivariate Analysis*, E. Arnold, London.
- Manisera M. (2005) *Measuring job satisfaction by means of Nonlinear Principal Component Analysis*, unpublished PhD thesis.
- Torgerson W. S. (1958) *Theory and methods of scaling*, Wiley, New York.
- Van Rijckevorsel J. (1987) *The application of fuzzy coding and horseshoes in multiple correspondence analysis*, DSWO Press, Leiden.
- Zani S. (2000) *Analisi dei dati statistici - I and II*, Giuffrè, Milano.