Proceedings e report

114

SIS 2017
Statistics and Data Science: new challenges, new generations

28–30 June 2017
Florence (Italy)

Proceedings of the Conference of the Italian Statistical Society

edited by
Alessandra Petrucci
Rosanna Verde

FIRENZE UNIVERSITY PRESS
2017

SIS 2017. Statistics and Data Science: new challenges, new generations : 28-30 June 2017 Florence (Italy) : proceedings of the Conference of the Italian Statistical Society / edited by Alessandra Petrucci, Rosanna Verde. – Firenze : Firenze University Press, 2017.
(Proceedings e report ; 114)

http://digital.casalini.it/9788864535210

ISBN 978-88-6453-521-0 (online)

Università degli Studi di Firenze
Firenze University Press
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

SOCIETÀ ITALIANA DI STATISTICA
Sede: Salita de' Crescenzi 26 - 00186 Roma
Tel +39-06-6869845 - Fax +39-06-68806742
email: sis@caspur.it web:http://www.sis-statistica.it

La Società Italiana di Statistica (SIS), fondata nel 1939, è una società scientifica eretta ad Ente morale ed inclusa tra gli Enti di particolare rilevanza scientifica. La SIS promuove lo sviluppo delle scienze statistiche e la loro applicazione in campo economico, sociale, sanitario, demografico, produttivo ed in molti altri settori di ricerca.

SIS2017 Committees

Scientific Program Committee:
Rosanna Verde (chair), Università della Campania "Luigi Vanvitelli"
Maria Felice Arezzo, Sapienza Università di Roma
Antonino Mazzeo, Università di Napoli Federico II
Emanuele Baldacci, Eurostat
Pierpaolo Brutti, Sapienza Università di Roma
Marcello Chiodi, Università di Palermo
Corrado Crocetta, Università di Foggia
Giovanni De Luca, Università di Napoli Parthenope
Viviana Egidi, Sapienza Università di Roma
Giulio Ghellini, Università degli Studi di Siena
Ippoliti Luigi, Università di Chieti-Pescara "G. D'Annunzio"
Matteo Mazziotta, ISTAT
Lucia Paci, Università Cattolica del Sacro Cuore
Alessandra Petrucci, Università degli Studi di Firenze
Filomena Racioppi, Sapienza Università di Roma
Laura M. Sangalli, Politecnico di Milano
Bruno Scarpa, Università degli Studi di Padova
Cinzia Viroli, Università di Bologna

Local Organizing Committee:
Alessandra Petrucci (chair), Università degli Studi di Firenze
Gianni Betti, Università degli Studi di Siena
Fabrizio Cipollini, Università degli Studi di Firenze
Emanuela Dreassi, Università degli Studi di Firenze
Caterina Giusti, Università di Pisa
Leonardo Grilli, Università degli Studi di Firenze
Alessandra Mattei, Università degli Studi di Firenze
Elena Pirani, Università degli Studi di Firenze
Emilia Rocco, Università degli Studi di Firenze
Maria Cecilia Verri, Università degli Studi di Firenze

Supported by:
Università degli Studi di Firenze
Università di Pisa
Università degli Studi di Siena
ISTAT
Regione Toscana
Comune di Firenze
BITBANG srl

Index

Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online), CC BY 4.0, 2017 Firenze University Press

Exploratory factor analysis of ordinal variables: a copula approach

Analisi fattoriale esplorativa di variabili ordinali: un approccio via copula

Marta Nai Ruscone

Abstract Exploratory factor analysis attempts to identify the underlying factors that explain the pattern of correlations within a set of observed variables. The analysis is almost always performed with Pearson's correlations even when the data are ordinal, but this is not appropriate since they are not quantitative data. The use of Likert scales is increasingly common in the field of social research, so it is necessary to determine which methodology is the most suitable for analysing the data obtained as non quantitative measures. In this context, also by means of simulation studies, we aim to illustrate the advantages of using Spearman's grade correlation coefficient on a transformation operated by the copula function in order to perform exploratory factor analysis of ordinal variables. Moreover, by using the copula, we consider the general dependence structure, providing a more robust reproduction of the measurement model.
Abstract *L'analisi fattoriale esplorativa vuole identificare i fattori latenti che spiegano un insieme di variabili osservate. L'analisi quasi sempre utilizza la correlazione di Pearson, anche quando i dati sono di natura ordinale, ma questo non é appropriato in quanto questi dati non sono quantitativi. L'uso di scale Likert é sempre piú comune nel campo della ricerca sociale, risulta quindi necessario determinare quale metodo risulta essere piú idoneo per l'analisi di tali dati tenendo presente che spesso vengono analizzati utilizzando tecniche idonee solo per misure quantitative. In questo contesto, e mediante studi di simulazione, si illustrano i vantaggi nell'utilizzo dello Spearman grade correltion ottenuto mediante l'utilizzo dalla funzione copula anziché della correlazione di Pearson. Con l'utilizzo della copula, si considera cosí la struttra di dipendenza generale, fornendo cosí una misurazione piú accurata*

Key words: Factor analysis, copula, ordinal variables, Likert scales, correlation

---

Marta Nai Ruscone
School of Economics and Management - LIUC - University Cattaneo, C.so Matteotti 22 - 21053 Castellanza (VA), Italy, e-mail: mnairuscone@liuc.it

Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online), CC BY 4.0, 2017 Firenze University Press

1 Introduction

Exploratory factor analysis is a widely used statistical technique in the social sciences where the main interest lies in measuring the unobserved construct, such as emotions, attitudes, beliefs and behaviors. The main idea behind the analysis is that the latent variables (also named factors) account for the dependencies among the observed variables (also named items or indicators) in the sense that if the factors are held fixed, the observed variables would be independent. In exploratory factor analysis the goal is the following: for a given set of observed variables $x_1, ..., x_p$ one wants to find a set of latent factors $\xi_1, ..., \xi_k$, fewer in number than the observed variables $(k < p)$, that contain essentially the same information. In its classical formulation [1], it concerns a set of continuous variables measured on a set of independent units. The data usually encountered in social sciences are of categorical nature (ordinal or nominal). The Likert Rating Scale [10], [11] is a simple procedure for generating measurement instruments which is widely used by social scientists to measure a variety of latent constructs, and meticulous statistical procedures have therefore been developed to design and validate these scales [3], [15]. However, most of these ignore the ordinal nature of observed responses and assume the presence of continuous observed variables measured at interval level. Evidence shows that, under relatively common circumstances, classical factor analysis (FA) yields inaccurate results characterizing the internal structure of the scale or selecting the most informative items within each factor [4], [7].

In the present work Spearman's grade correlation coefficient on a transformation operated by the copula function is employed, in order to take into account the ordinal nature of the data. The copula is a helpful tool for handling multivariate continuous distributions with given univariate marginals [14]. It describes the dependence structure existing across pairwise marginal random variables. In this way we can consider bivariate distributions with dependence structures, different from the linear one, that characterises the multivariate normal distribution.

So taking into account that the use of measurement instruments which require categorical responses from subjects is increasingly common in social research, and this implies the use of ordinal scales, the present work aims to point out a correct definition of dependence measure for ordinal variables rather than the Pearson correlation coefficient correctly applied to quantitative variables. Moreover, the use of several copulae with specific tail dependence allow us to obtain an index that weights the ordinal variables categories in several ways. In so doing we can address and recognize the ordinal nature of observed variables and estimate that weight directly from the data.

2 The copula function

The copula function is the key ingredient for handling multivariate continuous distributions with given univariate marginals. We will discuss this issue briefly below, for further details and proofs, see for instance [14], [8] and [2]. It describes the dependence structure existing across pairwise marginal random variables. In this way we can consider bivariate distributions with dependence structures different from the linear one that characterises the multivariate normal distribution.

A bivariate copula $C : I^2 \to I$, with $I^2 = [0,1] \times [0,1]$ and $I = [0,1]$, is the cumulative bivariate distribution function of a random variable (U_1, U_2) with uniform marginal random variables in [0,1]

$$C(u_1, u_2; \theta) = P(U_1 \leq u_1, U_2 \leq u_2; \theta), \quad 0 \leq u_1 \leq 1 \quad 0 \leq u_2 \leq 1 \tag{1}$$

where θ is a parameter measuring the dependence between U_1 and U_2.

The following theorem by Sklar [14] explains the use of the copula in the characterization of a joint distribution. Let (X_1, X_2) be a bivariate random variable with marginal cdfs $F_{X_1}(x_1)$ and $F_{X_2}(x_2)$ and joint cdf $F_{X_1,X_2}(x_1, x_2; \theta)$, then there is always a **copula function** $C(\cdot,\cdot;\theta)$ with $C : I^2 \to I$ such that

$$F_{X_1,X_2}(x_1, x_2; \theta) = C\big(F_{X_1}(x_1), F_{X_2}(x_2); \theta\big), \quad x_1, x_2 \in \mathbb{R}. \tag{2}$$

Conversely, if $C(\cdot,\cdot;\theta)$ is a copula function and $F_{X_1}(x_1)$ and $F_{X_2}(x_2)$ are marginal cdfs, then $F_{X_1,X_2}(x_1, x_2; \theta)$ is a joint cdf.
If $F_{X_1}(x_1)$ and $F_{X_2}(x_2)$ are **continuous** functions then the copula $C(\cdot,\cdot;\theta)$ is **unique**. Moreover, if $F_{X_1}(x_1)$ and $F_{X_2}(x_2)$ are continuous the copula can be found by the inverse of (2):

$$C(u_1, u_2) = F_{X_1,X_2}(F_{X_1}^{-1}(u_1), F_{X_2}^{-1}(u_2)), \tag{3}$$

with $u_1 = F_{X_1}(x_1)$ and $u_2 = F_{X_2}(x_2)$. This theorem states that each joint distribution can be expressed in term of two separate but related issues, the marginal distributions and the dependence structures between them. The **dependence structure** is explained by the copula function $C(\cdot,\cdot;\theta)$. Moreover the (2) provides a general mechanism to construct new multivariate models in a straightforward manner. By changing the copula function we can construct new bivariate distributions with different dependence structures, with the association parameter indicating the strength of the dependence, also different from the linear one that characterizes the multivariate normal distribution.

Each copula is related to the most important measures of dependence: the Pearson correlation coefficient, the Spearman grade correlation coefficient and tail dependence parameters. The Spearman grade correlation coefficient (see [14] pp. 169-170 for the definition of the grade correlation coefficient for continuous random variables) measure the association between two variables and can be expressed as a function of the copula. More precisely, if two random variables are continuous and have copula C with parameter θ, then the Spearman grade correlation is

$$\rho_s(C) = 12 \int_{I^2} C_\theta(u_1, u_2) du_1 du_2 - 3 = \frac{Cov(U_1, U_2)}{\sqrt{Var(U_1)}\sqrt{Var(U_2)}}. \tag{4}$$

For continuous random variables this is invariant with respect to the two marginal distributions, i.e. it can be expressed as a function of its copula. This property is also known as 'scale invariance'. Note that not all measures of association satisfy this property, e.g. Pearson's linear correlation coefficient [6]. Among all copulas $C : I^2 \to I$ such that for every $u, v \in I$, three especially noteworthy ones are $W(u,v) = max(u+v-1,0)$, $\Pi(u,v) = uv$, and $M(u,v) = \min(u,v)$. These copulae correspond to perfect negative association ($\rho_S(C) = -1$), independence ($\rho_S(C) = 0$), and perfect positive association ($\rho_S(C) = +1$) between the two random variables, respectively. For all $(u,v) \in I^2$ it holds that $W(u,v) \leq \Pi(u,v) \leq M(u,v)$.
The tail dependence relationship can be measured by means of the upper and lower tail dependence parameters

$$\lambda_u = \lim_{u \to 1^-} P[X_2 > F_2^{-1}(u) | X_1 > F_1^{-1}(u)] = \lim_{u \to 1^-} \frac{C(u,u)}{u}, \tag{5}$$

$$\lambda_l = \lim_{u \to 0^+} P[X_2 \leq F_2^{-1}(u) | X_1 \leq F_1^{-1}(u)] = \lim_{u \to 0^+} \frac{1-2u+C(u,u)}{1-u}. \tag{6}$$

If $\lambda_u \in (0,1]$ or $\lambda_l \in (0,1]$, the random variables X_1 and X_2 present upper or lower tail dependence. If $\lambda_u = 0$ or $\lambda_l = 0$, there is no upper or lower tail dependence. These parameters measures the dependence in the tails of the joint distribution, i.e. high/low values of one variable are associated with high/low values of the other one. They represent the probability that one variable is extreme given that the other is extreme. The Spearman grade correlation coefficient and both tail dependence parameters are directly associated with the parameters of some copula family [14].

3 Our proposal

Theory and methodology for exploratory factor analysis have been well developed for continuous variables, but in practice observed or measured variables are often ordinal.

Observations on an ordinal variable are assumed to have logical ordering categories. This logical ordering is typical when data are collected from questionnaires. A good example is the Likert Scale that is frequently used in survey research: $1 = Strongly\ disagree$, $2 = Disagree$, $3 = Neutral$, $4 = Agree$, and $5 = Strongly\ agree$. Although a question is designed to measure a theoretical concept, the observed responses are only a discrete realization of a small number of categories and distances between categories are unknown. Following [13], [9] and others, it is assumed that there is a continuous variable x_i* underlying the ordinal variable x_i, $i = 1, ..., p$. This continuous variable x_i* represents the attitude underlying the order responses to x_i and it is assumed to have a range from $-\infty$ to $+\infty$.

The underlying variable x_i* is unobservable. Only the ordinal variable x_i is observed. For an ordinal variable x_i with m_i categories, the connection between the ordinal variable x_i and the underlying variable x_i* is:

$$x_i \Leftrightarrow \tau^i_{i-1} < x_i* < \tau^i_i, \ \ i = 1,2,...,m_i \tag{7}$$

where

$$-\infty = \tau^i_0 < \tau^i_1 < \tau^i_2 < ... < \tau^i_{m_i-1} < \tau^i_{m_i} = +\infty \tag{8}$$

are threshold parameters. For variable x_i with m_i categories, there are $m_i - 1$ strictly increasing threshold parameters $\tau^i_1 < \tau^i_2 < ... < \tau^i_{m_i-1}$.

Let x_i and x_j be the two ordinal variables with m_i and m_j categories respectively. We define now Spearman's grade correlation via copula. We consider a copula C_θ associated with each pair (X_i*, X_j*) underlying the pair (X_i, X_j) in the set of ordinal items $X_1, X_2, ..., X_i$, we thus assume that each pair (X_i, X_j) corresponds to a bivariate discrete random variable obtained by a discretisation of a bivariate continuous latent variable $U_i = F(X_i*), U_j = F(X_j*)$ with support on the unit interval.

Let $A_{ij} = [u_{i-1}, u_i] \times [v_{j-1}, u_j]$, $i = 1,2,...,m_i$ $j = 1,2,...,m_j$, be the rectangles defining the discretisation. Let $p_{11},...,p_{m_im_j}$ be the joint probabilities of the ordinal variables corresponding to the rectangles $A_{11},...,A_{m_im_j}$. Let $V_{C_\theta}(A_{11},...,A_{m_im_j})$ be the volumes of the rectangles under the copula C_θ, then

$$V_{C_\theta}(A_{11},...,A_{m_im_j}) = p_{11},...,p_{m_im_j} \tag{9}$$

There exists a unique element in the family of copula for which (9) holds true. We apply this to each pair (X_i, X_j) $i \neq j$ in the set of the items. θ can be estimated via maximum likelihood [5] [12]. The multivariate normality assumption pertaining to the underlying variables, assumed by polychoric correlation and Pearson correlation, is relaxed. To apply the index one needs only to specify the dependence structure of the variables by means of a copula family.

In this way the construct validity is analysed according to ordinal data obtained from Likert scales using the most suitable method. The factor results show a better fit to the theoretical model when the factorization is carried out using the Spearman's grade correlation via copula rather than Pearson correlation. Our focus here has been to identify the type of correlation that yields a factor solution more in keeping with the original measurement model, as we believe this to have great importances in terms of drawing correct substantive conclusions. When we conduct a FA our results can be summarized as follow:

- regardless of the number of dimensions and items with skewness, Pearson correlations are lower than Spearman's grade correlations. The results are more significant when all items are asymmetric.
- The model obtained is more consistent with the original measurement model when we factorize using the Spearman's grade correlation. This result does not depend on the number of dimensions and asymmetric items.

To summarize the factor results obtained when we use Spearman's grade correlation better reproduce the measurement model present in the data, regardless of the number of factors.

References

1. Anderson, T.W.: An introduction to multivariate statistical analysis. Wiley, New York (2003)
2. Cherubini, U., Luciano, E., Vecchiato, W.: Copula methods in finance. John Wiley & Sons (2004)
3. DeVellis, R.F.: Scale development, theory and applications. Sage, Newbury Park (1991)
4. DiStefano, C.: The impactof categorization with confirmatory Factor Analysis. Structural Equation Modeling: A Multidisciplinary Journal **9**, 327–346 (2002)
5. Ekstrom, J.: Contributions to the Theory of measures of association for ordinal variables. igital Comprehensive Summaries of Uppsala Dissertation from the Faculty of Social Sciences, Uppsala (2003)
6. Embrechts, P., McNeil, A., Straumann, D.: Correlation and dependence in risk management: properties and pitfalls. Risk management: value at risk and beyond, 176–223 (2002)
7. Holgado-Tello, F.P., Chacón-Moscoso, S., Barbero-Garcia, I., Vila-Abad E.: Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. Quality and Quatity **44**, 153–166 (2010)
8. Joe, H.: Multivariate models and multivariate dependence concepts. CRC Press (1997)
9. Jöreskog, K. G.: New developments in LISREL: analysis of ordinal vriables usisng polychoric correlations and weighted least squares. Quality and Quatity **24(4)**, 387–404 (1990)
10. Likert, R.: A technique for the measurement of attitudes. Achives of Psychology, 44–45 (1932)
11. Likert, R., Sydney, R., Murphy, G.: A simple and reliable method of scoring Thurstone attitudes scales. The journal of Social Psychology, 228–238 (1934)
12. Martinson, E.O., Hamdan, M.A.: Maximum likelihood and some other asymptotically efficient estimators of correlation in two way contingency tables. J. Stat. Comput. Simul. **1**, 45-54 (1971)
13. Muthén, B. O.: Full maximum likelihood analysis of structural equation models withpolitomous variables. Psychometrika **9(1)**, 91–97 (1984)
14. Nelsen, R.B.: An introduction to copulas. Springer Science & Business Media (2013)
15. Spector, P.E.: Summating rating scale construction: an introduction. Sage, Newbury Park (1992)