Metodologia delle scienze umane

Giovanni Di Franco
Alberto Marradi

Factor analysis and principal component analysis

FrancoAngeli

Informazioni per il lettore

Questo file PDF è una versione gratuita di sole 20 pagine ed è leggibile con

La versione completa dell'e-book (a pagamento) è leggibile con Adobe Digital Editions. Per tutte le informazioni sulle condizioni dei nostri e-book (con quali dispositivi leggerli e quali funzioni sono consentite) consulta cliccando qui le nostre F.A.Q.

Metodologia delle scienze umane

Collana della Sezione di Metodologia dell'Associazione Italiana di Sociologia

La collana è un punto d'arrivo e allo stesso tempo un punto di partenza delle riflessioni sul metodo entro l'ampio ventaglio delle scienze umane.
Come punto d'arrivo di una tradizione complessa e ricca di solidi sedimenti, la collana intende collocarsi sul versante dell'alta divulgazione e raggiungere non solo gli studenti e i docenti universitari, ma anche il pubblico crescente delle professioni interessate alle varie forme di trattamento delle informazioni.
Come punto di partenza, essa non mancherà di presentare in modo problematico quei settori della tradizione metodologica teoricamente incerti, o fondati su presupposti discutibili, o soggetti ad abusi applicativi; né trascurerà di suggerire nuove direzioni e orientamenti.
Il piano della collana prevede ora una cinquantina di volumi, programmati su un arco di tempo di circa dieci anni e affidati a studiosi di sociologia, psicologia, statistica, storiografia, economia e altre discipline: una enciclopedia per il consolidamento e lo sviluppo delle scienze umane.

1120. *Metodologia delle scienze umane*

1. Gianni Losito, *L'analisi del contenuto nella ricerca sociale*
2. Luca Ricolfi, *Tre variabili. Un'introduzione all'analisi multivariata*
3. Alberto Marradi, *L'analisi monovariata*
4. Roberto Biorcio, *L'analisi dei gruppi*
5. Oscar Itzcovich, *L'uso del calcolatore in storiografia*
6. Giuseppe A. Micheli, Piero Manfredi, *Correlazione e regressione*
7. Francesca Zajczyk, *Fonti per le statistiche sociali*
8. Giampietro Gobo, *Le risposte e il loro contesto. Processi cognitivi e comunicativi nelle interviste standardizzate*
9. Paolo Montesperelli, *L'intervista ermeneutica*
10. Roberto Fideli, *La comparazione*
11. Antonio M. Chiesi, *L'analisi dei reticoli*
12. Cinzia Meraviglia, *Le reti neurali nella ricerca sociale*
13. Elisabetta Ruspini, *La ricerca longitudinale*
14. Juan Ignacio Piovani, *Alle origini della statistica moderna. La scuola inglese di fine Ottocento*
15. Giovanni Di Franco, *Corrispondenze multiple e altre tecniche multivariate per variabili categoriali*
16. Ivana Acocella, *Il focus group: teoria e tecnica*
17. Erika Cellini, *L'osservazione nelle scienze umane*
18. Paolo Parra Saiani, *Gli indicatori sociali*
19. Maria C. Pitrone, *Sondaggi e interviste. Lo studio dell'opinione pubblica nella ricerca sociale*
20. Giovanni Delli Zotti, *Tecniche grafiche di analisi e rappresentazione dei dati*
21. Federico Podestà, *Tecniche di analisi per la ricerca comparata trans-nazionale*
22. Fabrizio Martire, *La regressione logistica e i modelli log-lineari nella ricerca sociale*
23. Giovanni Di Franco, Alberto Marradi, *Factor analysis and principal component analysis*

Questo volume è stato accettato nella collana in seguito
al giudizio positivo conforme di due *referees* anonimi,
di cui uno straniero.
Per conto del Comitato Scientifico della collana
hanno seguito la redazione del volume:

Antonio De Lillo (†)
Franco Rositi

Giovanni Di Franco
Alberto Marradi

Factor analysis and principal component analysis

Metodologia delle scienze umane / 23

FrancoAngeli

Authors' note

Chapters 1, 2, and sections 6.1, 6.3, 7.1, 7.3, 8.1, 8.2.1 and 8.3 have been written by Alberto Marradi. Chapters 3, 4 and 5 and sections 6.2, 7.2 and 8.2.2 have been written by Giovanni Di Franco.

English translation by Maureen Galvin with the supervision by Alberto Marradi.

Grafica della copertina: Elena Pellegrini

Index

Presentazione editoriale

La collana *Metodologia delle scienze umane* è stata sempre finora edita in lingua italiana. Questa è la prima volta in cui si è deciso di editare in lingua inglese uno suoi volumi. Non so se resterà l'ultima; è possibile, soltanto possibile, che ne seguano altre. Sebbene a me piacciano le eccezioni, non è per amore di eccezioni che si è fatta questa scelta linguistica. Né per ossequio astratto o interessato alla recente imponente pressione generalizzata su tutta la produzione accademica italiana (da parte di una nuova casta di valutatori) a "sprovincializzarsi" mediante il privilegio a riviste scritte in lingua inglese e comunque alla redazione in lingua inglese dei propri testi (questa pressione diviene particolarmente strana quando vuol sostenersi su ragioni ideali, e non sulla semplice obbedienza all'egemonia culturale-politica-economica del *basic english*; basterà ricordare che mai in Italia siamo stati così cosmopoliti come in quella breve stagione fra fine '800 e inizio '900 nella quale positivismo settentrionale e idealismo centro-meridionale dialogarono intensamente con le principali culture occidentali: ascoltando e ascoltati, e scrivendo rigorosamente in italiano, e normalmente in un buon chiaro italiano). Quale dunque il motivo di questa nostra novità editoriale?

Non sempre le giustificazioni sono uno sviluppo lineare dei motivi che hanno condotto a decidere. Nel presente caso il motivo originario, occorre riconoscerlo, è spurio rispetto a una giustificazione che possa considerarsi generale e quale poi è in noi effettivamente maturata. Il "motivo originario" è consistito nella volontà, mia in particolare, di porre rimedio a un piccolo in-

cidente. Data la marginale rilevanza della vicenda, qui accennerò soltanto al fatto che il buon libro di Di Franco e Marradi, proposto inizialmente alla nostra collana, era poi scivolato verso un altro editore (Bonanno, 2003): era accaduto che alcuni eccessivi ritardi dei nostri *referees* avevano reso giustamente impazienti i due autori e convinto lo stesso direttore della collana, Alberto Marradi, a dare così a noi tutti un segno di rimprovero. Io me ne dispiacqui molto.

Avevo trovato il testo di Di Franco e Marradi quale testo esemplare delle intenzioni originarie della nostra collana: conteneva una sostanziosa e intelligente ricostruzione storica dell'analisi fattoriale e della analisi in componenti principali, aveva uno stile di estrema precisione e chiarezza, uno svolgimento coerente, una cura estrema dei particolari; ed inoltre, distingueva chiaramente fra una presentazione didattica della tradizione metodologica da una parte e, dall'altra, non solo qualche gustosa polemica verso la *blind factor analysis* e le *factor fishing expeditions* (spedizioni a pesca di fattori), costumi quantofrenici così in voga fra gli utilizzatori sconsiderati di queste sofisticate tecniche statistiche, ma anche una proposta originale di uno dei due autori. Dicendolo proprio in breve, questa proposta consiste nel riconoscere un privilegio di efficacia conoscitiva alla relativa semplicità dell'analisi delle componenti principali, di contro alle manipolazioni ipercomplesse dell'analisi fattoriale. I due autori apprezzano la posizione di Steiger e Schönemann (1978), ripresa in Italia da Gangemi (1982), secondo la quale l'analisi dei fattori può essere descritta con una semplice equazione: analisi dei fattori = analisi in componenti principali + assunti ingiustificati + complicazione dei calcoli + indeterminatezza. Ma – ed è in questo un particolare merito del libro – i lettori potranno giudicare sull'opzione dei due autori non in carenza di informazioni a riguardo di opzioni avverse.

È sulla base della convinzione sul valore di quel testo che io ho insistito per anni, con Alberto Marradi, nell'idea che, mancando del resto ancora nella nostra collana un titolo su acp e af, lui e Di Franco dovessero riscriverlo per noi: una idea assurda per chi ritiene di aver già fatto tutto quello che c'era da fare e che comunque era in grado di fare. Qualche anno fa è spuntata

la proposta di una pubblicazione della traduzione in inglese di quel libro. Non sapevamo però come giustificarla. Si è pensato alla fine che la nostra collana potesse aprirsi a testi già pubblicati altrove in lingua italiana nel caso che: a) fossero particolarmente ben riusciti; b) ovviamente che l'autore o gli autori fossero d'accordo. In questo modo la collana avrebbe assunto anche l'onere di rendere più accessibili fuori del nostro paese la nostra riflessione metodologica, attualmente forse meno riconosciuta di quanto le sarebbe dovuto. Questo indirizzo avrebbe potuto valere anche per qualche volume già da noi stessi pubblicato.

Ho detto agli inizi di queste pagine che a me piacciono le eccezioni. Non so se qualche lettore abbia già avvertito qualche fastidio per una siffatta del tutto irrilevante notizia. Vorrei però almeno alleviare quel tanto di sconsiderato e perfino di pretenzioso che c'è in questa esibizione di eccentricità aggiungendo che non sono mai riuscito a tollerare perfino le eccezioni da me stesso praticate senza cercare di iscriverle in più accettabili giustificazioni. Con una punta di sospetto si potrebbe dire: senza una razionalizzazione. Ma non sempre le razionalizzazioni coprono una cieca volontà: qualche volta scoprono nuove opportunità. Penso che la decisione editoriale che ho appena descritto renda ancora migliore la nostra collana.

Franco Rositi

1. Factor analysis and principal component analysis: nature and functions

With the terms 'principal component analysis' and 'factor analysis' a set[1] of techniques is designated, the common denominator of which is the application of the procedures of matrix algebra (see chapters 3 and 4) to a correlation matrix of cardinal variables[2] so as to sum up the variance — and consequently, the information — of the set of variables within a more limited

[1] Reference is made to a set as there is only one way of carrying out a principal component analysis, while the techniques that go under the name of factor analysis are many (see section. 5.2). In this context, the French school of *Analyse des données* calls 'factor analysis' any technique that reconstructs the dimensional aspect of a semantic space, including principal component analysis and even correspondence analysis, operating on categorial variables (Amaturo 1989; Di Franco 2006).

[2] In reality, not unfrequently the correlation coefficients in the matrix are not calculated from cardinal variables, but from variables generously called "interval variables". The degree of legitimacy with which the correlation coefficients are calculated naturally depends on the degree in which the variables in question approach one of the two conditions imposed by Stevens (1946) in order to speak of 'interval scales': the equality of the intervals between each pair of points adjacent on the scale. Sometimes correlation coefficients are calculated between dichotomic variables and even between categorial variables. However, there are serious biases in the correlation coefficients between dichotomies if their distribution is uneven (Gangemi 1977; Marradi 1997). The idea of applying techniques that presuppose cardinal variables to categorial variables the categories of which are not even ordinate, testifies the aberrations which can induce "quantofrenia" (see Rummel 1967; Bellucci 1984; De Mucci 1984). On one principal component analysis conducted by Capecchi (1967) using manycategorial variables (i.e. far from cardinality) see note 6.

number of vectors, called components or factors according to the technique applied.

In most methodology handbooks, principal component analysis and factor analysis are considered multivariate techniques like the others, and therefore – due to their complexity – are usually presented after multiple and partial correlation, path analysis, discriminant analysis and before canonical correlation[3].

From a technical point of view, this collocation is correct: students cannot be aware of the procedures for extracting components and factors if they have not already mastered the concepts of multiple correlation and partial correlation.

From a methodological point of view, on the other hand, there is a radical difference between the techniques of the so-called "general linear model" (regression and correlation, variance and covariance analysis, discriminant analysis and so on) and the techniques we are analyzing in this volume. In the first case, a dependent variable is identified and the linear combination (in the case of cardinal variables) or the combinations of categories (in the case of categorial variables) are sought that best reproduce the variance of the dependent (or that best predicts category frequencies, if we are talking about categorials). The degree of association that has been found is often interpreted as proof of a causal influence – i. e. an empirical tie – between variables of a different kind: for example between socio-demographical properties and attitudes.

In the pca and in the fa[4], on the contrary, linear combinations are sought between the variables of a set which best reproduce the variance of the same set: the degree of statistical association between pairs of variables of the set is interpreted as proof of

[3] See e. g. Blalock (1960), Morrison (1967), Cooley e Lohnes (1971), Van de Geer (1971), Maxwell (1977), Perrone (1977), Bouroche e Saporta (1980), Sadocchi (1980), Bailey (1982), Memoli e Saporiti (1985), Rizzi (1985), Bohrnstedt e Knoke (1994), Fabbris (1997), Bolasco (1999), Di Franco (1997; 2001; 2011) and so on.

[4] From now on, when necessary, we shall use the acronym pca in place of the expression 'principal component analysis' and the acronym fa in place of the expression 'factor analysis'.

the existence (or inexistence) of a significant semantic overlap between (some of) these variables and eventually as the authorisation (or lack thereof) for considering[5] (some of) such variables as indicators of the same general concept. For this reason, as underlined by Eysenck (1953), Capecchi (1968) and Ricolfi (1982) amongst others, one should not submit to the same pca or fa variables between which a causal relation occurs[6]: therefore "attitudinal variables should not be mixed with behavior variables, cognitive variables with valutative variables" (Galtung 1967, 263).

[5] A century ago the ideators of fa, who had an ontological conception of the relation of indication, would have said a "proof of the fact that (some of) these variables are indicators of that given general concept". It should be pointed out that – when the variables submitted to pca or to fa derive, as frequently occurs, from a Likert scale – often the associations between them depend in the main from the *agreement bias*, i. e. from the tendency on the part of the interviewees to be in agreement with all, or nearly all, of the statements submitted to them. The *agreement bias* is an endemic phenomenon in social research, albeit in the literature on pca and fa there is a tendency to ignore it. On the contrary, it is often recognized a more general spurious effect of questions' wording that Ricolfi has called "linguistic disturbance" (1987, 119; see also Blalock 1961/1967, 169). Naturally, "if the variances of the variables are sharply inflated by errors, the weight of the components can turn out to be spuriously high, given that the first component will have extracted the maximum of the total variance (true plus error) and the subsequent components the maximum of the remaining variance" (Maxwell 1977/1981, 93-4).

[6] In effect, Capecchi does not respect this principle at all when he submits to the same pca a set of 26 variables comprising the education of the parent and of the interviewee, the profession of the interviewee and that of his/her father and grandfather, age and the family's political inclinations, religious practice, indicators of intra- e inter-generational mobility, indicators of general culture and opinions on various issues (1967, 430-42). Among other things, it is easy to see how many of these variable are not even ordinal, and consequently how by applying techniques that presuppose a cardinal level nonsensical results are obtained.

De Mucci does the same with ecological data (1984, 18-20) submitting to the same fa the percentage of votes attributed to the various parties, the percentage of inhabitants in the various professions and the percentage of inhabitants belonging to the various categories of age and income.

For the same reason, pca and fa are inadequate for formulating or testing hypotheses or theories[7], at least unless such assertions concern the semantic relations between the variables considered, and/or between them and more general concepts[8].

As mentioned previously, pca and fa sum up the the variance of one set of variables in a small number of vectors. The result (of summing up the variance) is generally seen by statisticians and economists as the function of these techniques, and of pca in particular[9]. However – as we shall see in the follonig section – for the inventors of the techniques themselves, as for many psychologists, sociologists, politologists, synthetizing the variances aims at a different target; in other words, it is the result but not the function – or at least not the main function – of these techniques.

The goals that have been proposed for fa and pca are many, more or less related. For this reason, every time that it is not considered necessary to make reference to a specific purpose,

[7] Conflicting opinions are occasionally expressed by authors of great technical competence albeit not provided with equal epistemological awareness (Harman 1967, 6, 268; Mulaik 1972, 10) and by others (Henrysson 1960, 56; Rummel 1967). See below on pca or confirmative fa.

[8] In this respect, see fa and confirmative pca at the end of this chapter.

[9] See for instance Henrysson (1960), Harman (1967), Morrison (1967), Maxwell (1968), Jöreskog (1978), Perrone (1977), Bouroche and Saporta (1980), Sadocchi (1980), Bailey (1982), Memoli and Saporiti (1985), Rizzi (1985), Bohrnstedt and Knoke (1994), Fabbris (1997), and Bolasco (1999).

The distinction between pca, which is limited to describing synthetically – and therefore is a "method" – and fa, which starts from assumptions relative to the nature of the variances and their erratic part – and consequently is a "model" – is widely accepted (see Ricolfi 1992, 80). However, as will be shown, from the genetic point of view, the difference was not conceived in these terms. Furthermore, whatever the theoretical constructions of creators and interpreters of these techniques may be, the effective difference is reduced to the fact that in pca the diagonal of the correlation matrix is left intact, while in the various versions of fa it is manipulated in some way (see chapter 5). Furthermore, this difference has very few practical effects, as recognized by many authors (Harris 1962 and 1967; Lawley and Maxwell 1963; Morrison 1967; Kendall 1968; Velicer 1976; Steiger and Schönemann 1978; Kline, 1993, 123, and 1994, 58; Fabbris 1997, 164; Gangemi 1997, 268-9; Di Franco 1997; 2001; 2011; 2013a, 2013b)) and confirmed by an empirical test carried out *ad hoc* and reported in this book (see section. 5.3).

during the course of this book we will refer to the result (synthetizing the variance) as the function of pca (and also of fa).

In this chapter, on the contrary, we propose to classify the purposes in a discursive way, ordering the categories of purposes along a dimension ranging from simple to complex, bearing in mind, however, diachronic considerations and other criteria deemed relevant.

To one or several of these purposes reference will be made, when necessary, during the course of the discussion.

1.1. A classification of the goals of factors analysis and principal component analysis

It can be said that the idea of pca in its summarizing function originated at the close of the 1800s when Galton faced the problem of reducing the number of anthropometric measures relative to criminals, that he considered redundant (see section 2.1). His colleague Edgeworth (1893) resolved the problem calculating three "linear orthogonal functions" of the anthropometric variables (what today we would call 'components'); in so doing he developed the original idea of Galton in the direction of synthesis as the function of (the future) pca[10]. The idea could have been developed in another direction, i. e. towards the selection of some of the anthropometric measures as the most valid indicators of propensity toward crime. The point is that at the end of the 1800s, the concept of indicator was hardly familiar and the concept of validity still had to be envisaged[11]. Nonetheless,

[10] We have already spoken about this function: it need be remembered in this classification, but is so obvious that further comments are not necessary. Moreover, we shall not deal with other functions of a prevalently technical interest also mentioned in the literature, such as that of accounting for the relation between sets of variables (see for instance Holzinger 1940; Guttman 1953; Maxwell 1968; Macrae 1970; Brislin *et al.* 1973; Weisberg 1974; Jöreskog 1978) or that of reducing the range of a correlation matrix, which Thurstone (1947) attributes to factor analysis.

[11] It is well known that a concept analogous to that of indicator is already present in the works of Villermé (1840), Quetelet (1869), Durkheim (1893;