RAGIONANDO DI SVILUPPO LOCALE: UNA LETTURA "NUOVA" DI TEMATICHE "ANTICHE"

a cura di
Francesco Contò
Mariantonietta Fiore

Università degli Studi di Foggia
Dipartimento di Economia

FrancoAngeli
OPEN ACCESS

Copyright © 2020 by FrancoAngeli s.r.l., Milano, Italy. ISBN 9788835104605

RAGIONANDO DI SVILUPPO LOCALE: UNA LETTURA "NUOVA" DI TEMATICHE "ANTICHE"

a cura di
Francesco Contò
Mariantonietta Fiore

FrancoAngeli
OPEN ACCESS

 ISBN 9788835104605

Il lavoro si colloca nell'ambito del progetto SKIN – **S**hort supply chain **K**nowledge and **I**nnovation **N**etwork (www.shortfoodchain.eu) finanziato dall'Unione Europea con il programma Horizon 2020, bando H2020-RUR-2016-2017 (Rural Renaissance – Fostering innovation and business opportunities), Grant Agreement n. 728055. Capofila Università degli Studi di Foggia – Dipartimento di Economia.

La stampa è stata finanziata dall'Università degli Studi di Foggia – Dipartimento di Economia, Delibera Consiglio di Dipartimento del 20 maggio 2020 punto 5bis, nell'ambito del progetto SKIN per la parte riguardante la stampa cartacea dei volumi, mentre l'edizione Open Access è stata finanziata da Tinada s.r.l. – Spin off dell'Università di Foggia.

Supervisione scientifica: prof. Francesco Contò, prof.ssa Mariantonietta Fiore.

La pubblicazione è stata sottoposta a processo di revisione tra pari (peer review).

Hanno curato la collocazione, l'organicità e la revisione dei testi del volume: prof. Francesco Contò, prof.ssa Mariantonietta Fiore.

Coordinamento editoriale, elaborazioni, segreteria: Società Tinada s.r.l. – Spin off dell'Università di Foggia.

Copyright © 2020 by FrancoAngeli s.r.l., Milano, Italy. ISBN 9788835104605

INDICE

PARTE II
PROGETTARE IDEE E FUTURO:
LE OPPORTUNITÀ DELL'UE

PARTE III
VALUTIAMO LO SVILUPPO

PARTE IV
ATTREZZI PER L'ANALISI DELLA REALTÀ

PARTE V
"NUOVE" METODOLOGIE E CHIAVI DI LETTURA
PER LA VALORIZZAZIONE DEL "VECCHIO" STRUMENTO DELLA FILIERA CORTA: UNA CARRELLATA DI EVIDENZE EMPIRICHE

PARTE VI
FOCUS DI APPROFONDIMENTO SU ASPETTI EMERSI DAL PROGETTO SKIN

3. ANALISI DELLE COMPONENTI PRINCIPALI

di *Leonardo Salvatore Alaimo*, Istituto Nazionale di Statistica Istat
e *Carlotta Pacifici*, Sapienza Università di Roma

Lo studio dei fenomeni complessi pone spesso il ricercatore di fronte alla necessità di analizzare molteplici variabili, ognuna delle quali contribuisce in modo diverso a definire il fenomeno di interesse. La gestione di un numero elevato di variabili può rappresentare un problema, specialmente se si punta a una lettura sintetica dei dati di partenza.

L'obiettivo del metodo delle componenti principali – concepito da Pearson nel 1901 e in seguito sviluppato da Hotelling dal 1933 – è quello di individuare, a partire da un set di variabili tra loro correlate, ulteriori nuove variabili, dette componenti, ottenute tramite una combinazione lineare delle variabili originarie. Una volta calcolate, le componenti forniranno la sintesi più fedele dei dati, e per questo entreranno nel modello in numero minore rispetto alle variabili di partenza.

Nei seguenti paragrafi verrà presentata l'analisi delle componenti principali (3.1), che sarà formalizzata analiticamente dal punto di vista della popolazione (3.2). Successivamente, verranno illustrati dei metodi utili alla scelta delle componenti (3.3) e alla loro interpretazione (3.4).

3.1. Introduzione

L'analisi delle componenti principali si configura come una delle tecniche di statistica multivariata usate per la riduzione dimensionale dei fenomeni, che punta ad un trade-off ottimale tra riduzione della complessità e perdita di informazioni rilevanti. Si utilizza esclusivamente per variabili quantitative che siano correlate tra loro. Data la presenza di interdipendenza

tra variabili, le dimensioni effettive attraverso cui il fenomeno verrà spiegato saranno inferiori a quelle osservate. Per mezzo delle componenti principali, sarà quindi possibile sostituire lo spazio $\mathbb{R}^p$ formato dalle p variabili originarie, con uno di dimensioni ridotte, $\mathbb{R}^k$. Il nuovo spazio sarà caratterizzato dalla scelta delle prime k componenti (con $k \ll p$) che permetteranno una spiegazione esaustiva del fenomeno di interesse. Il parametro che permette di definire la potenza informativa di una componente è la sua variabilità, e l'indice statistico più usato per quantificarla è la varianza.

Le componenti principali sono dunque delle variabili ausiliarie, non direttamente osservate, che forniscono una misura complessiva delle variabili di partenza. Le componenti vengono espresse in funzione delle variabili originarie tramite una combinazione lineare di cui si dovrà massimizzare la varianza.

Ogni componente dovrà mettere in luce un aspetto diverso del dataset originario: per questo motivo è necessario che esse siano fra loro in-correlate (ortogonali). Il procedimento che consente di individuare le componenti è di tipo iterativo. Le componenti vengono estratte in modo sequenziale così che la prima spieghi il massimo della varianza comune a tutte le variabili, la seconda, non correlata con la prima, spieghi il massimo della varianza comune residua e così fino alla spiegazione di tutta la varianza. Riassumendo, si punta alla stima di componenti principali tali che:

- siano combinazioni lineari delle variabili di partenza;
- abbiano varianza massima;
- siano in-correlate tra loro.

Prima di presentare nel dettaglio il funzionamento del modello, occorre precisare che si possono adottare due approcci: uno inferenziale, basato sulla risoluzione di un problema di ottimo e su aspetti statistici riguardanti la matrice varianze-covarianze, e l'altro più geometrico, basato sugli indici di distanza. A prescindere dall'approccio scelto, la sostanza del modello rimane la stessa. L'analisi che verrà realizzata in questo lavoro seguirà il primo approccio.

3.2. Formalizzazione del modello

Considerato il vettore delle variabili casuali,

$$\boldsymbol{x} = (X_1, \quad X_2, \quad \ldots \; X_i \quad \ldots, \quad X_p)' \tag{1}$$

con $i = 1, \dots, p$, si cerca una combinazione lineare che vada a formare la prima componente principale L_1 con la massima varianza:

$$L_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p = \boldsymbol{a_1'x} \tag{2}$$

dove $a_1 = (a_{11}, a_{12}, \dots, a_{1p})$' indica il vettore di coefficienti associati alle p variabili casuali. Il vettore dei coefficienti da individuare dovrà massimizzare la varianza totale della combinazione lineare espressa nella formula (2). A tal fine, si definisce dapprima la matrice varianze-covarianze del vettore di variabili casuali $\boldsymbol{x}$:

$$\boldsymbol{\Sigma} = E[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})'] = \begin{pmatrix} Var(X_1) & \dots & Covar(X_1, X_p) \\ \vdots & \ddots & \vdots \\ Covar(X_p, X_1) & \dots & Var(X_p) \end{pmatrix} \tag{3}$$

dove Σ è una matrice (p x p) avente sulla diagonale principale le varianze e nei termini extra-diagonali le covarianze, e $\mu = [E(X_1), \dots, E(X_p)]'$ è il vettore dei valori attesi associato a ciascuna variabile casuale.

Per le proprietà di linearità della varianza, essendo la prima componente principale il risultato di una combinazione lineare, anche la sua varianza sarà una combinazione lineare data dalla matrice nella formula (3) a cui verrà premoltiplicato e post-moltiplicato il vettore di coefficienti a_1.

$$Var(L_1) = Var(\boldsymbol{a_1'x}) = \boldsymbol{a_1'\Sigma a_1} \tag{4}$$

La varianza della componente L_1 sarà proprio la funzione obiettivo da massimizzare. Affinché la soluzione sia unica e finita, è necessario imporre il vincolo di normalizzazione sulla norma del vettore dei coefficienti:

$$\boldsymbol{a_1'a_1} = 1. \tag{5}$$

Il problema di massimo vincolato diventa, quindi:

$$\begin{cases} \max (\boldsymbol{a_1'\Sigma a_1}) \\ sub\ \boldsymbol{a_1'a_1} = 1 \end{cases} \tag{6}$$

da risolvere per mezzo della funzione Lagrangiana. Ponendo le derivate parziali uguali a zero, si ottiene l'equazione caratteristica di $\boldsymbol{\Sigma}$:

$$(\boldsymbol{\Sigma} - \lambda\mathbf{I})\boldsymbol{a_1} = 0. \tag{7}$$

dove **I** è la matrice identità[1] di dimensione (p x p).

Le soluzioni dell'equazione, ottenute ponendo $|\boldsymbol{\Sigma} - \lambda\mathbf{I}| = 0$, sono proprio gli autovalori della matrice $\boldsymbol{\Sigma}$. Dato che la matrice è semi-definita positiva, avrà p autovalori non negativi che risultano essere le soluzioni del sistema lineare in p incognite. Tenendo a mente il problema di massimizzazione, si sceglierà dapprima l'autovalore maggiore (chiamato λ_1) a cui corrisponderà l'autovettore $\boldsymbol{a_1}$.

Moltiplicando nella (7), la soluzione diventa:

$$\boldsymbol{\Sigma a_1} = \lambda_1 \boldsymbol{a_1} \tag{8}$$

Data la condizione di normalizzazione vista in nella formula (5), moltiplicando ambo i membri per $\boldsymbol{a_1'}$, si ottiene:

$$\boldsymbol{a_1'\Sigma a_1} = \lambda_1 \boldsymbol{a_1'a_1} = \lambda_1 = Var(L_1) \tag{9}$$

La varianza della prima componente principale coincide con l'autovalore λ_1, pertanto questa risulta essere massimizzata, dal momento che è stato scelto il massimo autovalore associato a $\boldsymbol{\Sigma}$.

Passando alla determinazione della seconda componente principale, il percorso logico da seguire resta analogo. Tuttavia, il sistema di vincoli viene ampliato, aggiungendo la condizione di in-correlazione tra le prime due componenti principali.

Data $L_2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p = \boldsymbol{a_2}'\boldsymbol{x}$,

$$Cov(L_2, L_1) = Cov(\boldsymbol{a_2'x}, \boldsymbol{a_1'x}) = E[\boldsymbol{a_2'}(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{a_1}] =$$
$$= \boldsymbol{a'}_2 E[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})']\boldsymbol{a_1} = \boldsymbol{a'}_2\boldsymbol{\Sigma a_1} = 0 \tag{10}$$

Il problema di massimo vincolato diventa quindi

$$\begin{cases} max\ (\boldsymbol{a_2}'\boldsymbol{\Sigma a_2}) \\ sub\ \boldsymbol{a_2'a_2} = \mathbf{1} \\ sub\ \boldsymbol{a}'_2\boldsymbol{\Sigma a_1} = \mathbf{0} \end{cases} \tag{11}$$

Risolvendo in modo analogo a quanto visto per la prima componente, si giunge all'equazione caratteristica:

[1] La matrice identità è una matrice quadrata in cui tutti gli elementi sulla diagonale principale sono costituiti dal numero 1 e tutti gli altri dal numero 0.

$$(\boldsymbol{\Sigma} - \lambda\mathbf{I})\boldsymbol{a_2} = 0 \tag{12}$$

dove la soluzione che massimizzerà la varianza della componente L_2 sarà λ_2, il secondo più grande autovalore di $\boldsymbol{\Sigma}$, avente come corrispondente autovettore proprio $\boldsymbol{a_2}$. Il sistema non può avere come soluzione λ_1 poiché altrimenti gli $\boldsymbol{a_1}$ e $\boldsymbol{a_2}$ coinciderebbero, violando così il primo vincolo. Generalizzando, per l'm-esima componente si avrà la seguente combinazione lineare:

$$L_m = a_{m1}X_1 + a_{m2}X_2 + \cdots + a_{mp}X_p = \boldsymbol{a_m}'\boldsymbol{x} \tag{13}$$

la cui varianza sarà massimizzata risolvendo il seguente problema di massimo vincolato:

$$\begin{cases} max(\boldsymbol{a_m}'\boldsymbol{\Sigma a_m}) \\ sub\ \boldsymbol{a'_m a_m} = 1 \\ sub\ \boldsymbol{a'}_m\boldsymbol{\Sigma a_1} = \boldsymbol{a'}_m\boldsymbol{\Sigma a_2} = \boldsymbol{a'}_m\boldsymbol{\Sigma a_3} = \cdots = \boldsymbol{a'}_m\boldsymbol{\Sigma a_p} = 0 \end{cases} \tag{14}$$

avente come soluzione l'autovalore λ_m, rispettiva varianza di L_m.

L'analisi finora condotta permette di elencare alcune proprietà algebriche associate alle componenti principali della popolazione:

Dato che ad ogni generica componente si lega l'autovalore con varianza immediatamente maggiore rispetto alla componente precedente, si possono ordinare le varianze di ciascuna componente:

$$Var(L_1) \geq Var(L_2) \geq \cdots \geq Var(L_p) \tag{15}$$

Come visto in precedenza, la varianza di ogni componente principale è pari all'autovalore corrispondente. Ne segue che la somma delle varianze delle componenti è uguale alla somma delle varianze delle variabili casuali di $\boldsymbol{x}$:

$$\sum_{m=1}^{p} Var(L_m) = \sum_{m=1}^{p} \lambda_m = tr(\boldsymbol{\Sigma}) = \sum_{m=1}^{p} Var(x_m) \tag{16}$$

È, quindi, possibile costruire un indicatore che quantifica la quota di varianza totale spiegata da ciascuna componente:

$$Porzione\ di\ varianza\ spiegata\ da\ \lambda_m = \frac{\lambda_m}{tr(\boldsymbol{\Sigma})} \tag{17}$$

Un problema applicativo dell'ACP riguarda la dipendenza delle componenti principali dall'unità di misura delle variabili di partenza. Infatti, il confronto tra componenti risulta alterato se:

- le variabili originarie non hanno la stessa unità di misura;
- si effettuano cambiamenti di scala sulle variabili originarie;
- la variabilità di una variabile è maggiore delle altre.

Tutto ciò altererà la matrice varianze-covarianze e di conseguenza varieranno anche i risultati dell'ACP. In riferimento all'ultimo punto, considerato che l'obiettivo dell'ACP è quello di riprodurre la variabilità del dataset originario, la variabile che avrà maggior peso nel determinare le componenti, sarà senz'altro quella con maggiore variabilità.

Per ovviare a questi problemi, prima di condurre un'ACP, si suggerisce l'uso di un campione che abbia variabili espresse in percentuale, sotto forma di numeri indici, oppure variabili standardizzate. Nel caso di quest'ultime, la matrice varianze-covarianze diventa la matrice di correlazione campionaria **R**:

$$\mathbf{R} = \begin{pmatrix} 1 & \cdots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{p1} & \cdots & 1 \end{pmatrix} \tag{18}$$

Il problema di massimizzazione vincolata avrà, quindi, come matrice centrale **R** anziché **S**. È utile notare che, nel caso delle variabili standardizzate, $tr(\boldsymbol{R}) = \sum_{m=1}^{p} 1 = p$.

3.3. Scelta delle componenti principali

Una volta definite le componenti principali, l'obiettivo è ridurre lo spazio della matrice dei dati di partenza, prendendo il numero minore di componenti che, allo stesso tempo, ricostruisce quasi integralmente l'informazione iniziale. Come già accennato nel paragrafo introduttivo, si cerca un numero di componenti k, con $k \ll p$, la cui varianza totale non sia eccessivamente inferiore a quella iniziale e sia tale da sintetizzare in modo esaustivo lo spazio p-dimensionale originario.

Esistono diversi metodi di supporto alla decisione sul numero di componenti da selezionare. In questo contesto ne prenderemo in considerazione quattro:

- stabilire una soglia minima sulla quota di varianza spiegata da un gruppo di componenti;
- scree-graph;

- regola di Kaiser;
- test d'ipotesi.

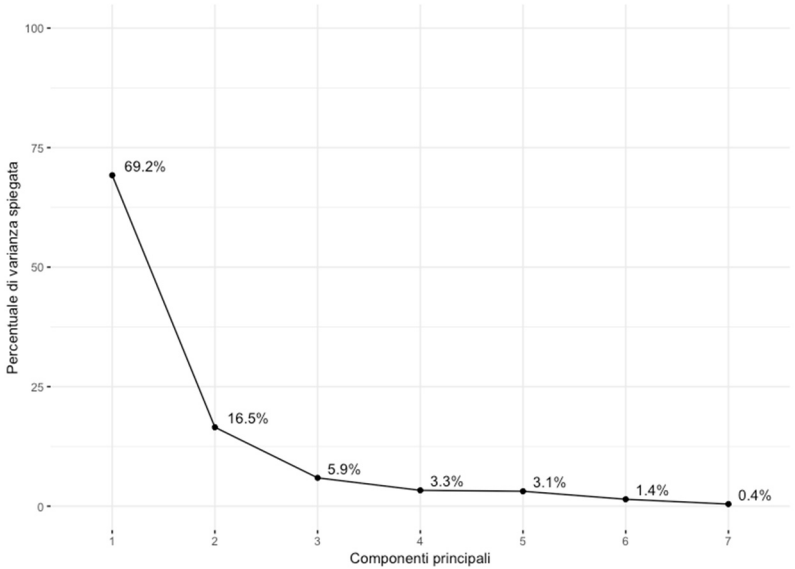
Il primo metodo consiste nello **stabilire una soglia minima Q*** sulla quota di varianza totale che si vorrebbe spiegare tramite le prime *k* componenti principali (spesso $Q^* \approx 80\%$ della varianza totale). Si sceglie il più piccolo numero delle componenti *k* il cui potere informativo sia maggiore o uguale della soglia stabilita.

$$\frac{l_1+l_2+\cdots+l_k}{tr(\mathbf{S})} \geq Q^* \qquad \frac{l_1+l_2+\cdots+l_k}{p} \geq Q^* \qquad (19)$$

per variabili standardizzate

Può risultare utile alla scelta rappresentare graficamente la percentuale di varianza spiegata da ciascuna componente, attraverso un grafico simile a quello riportato in Figura 1.

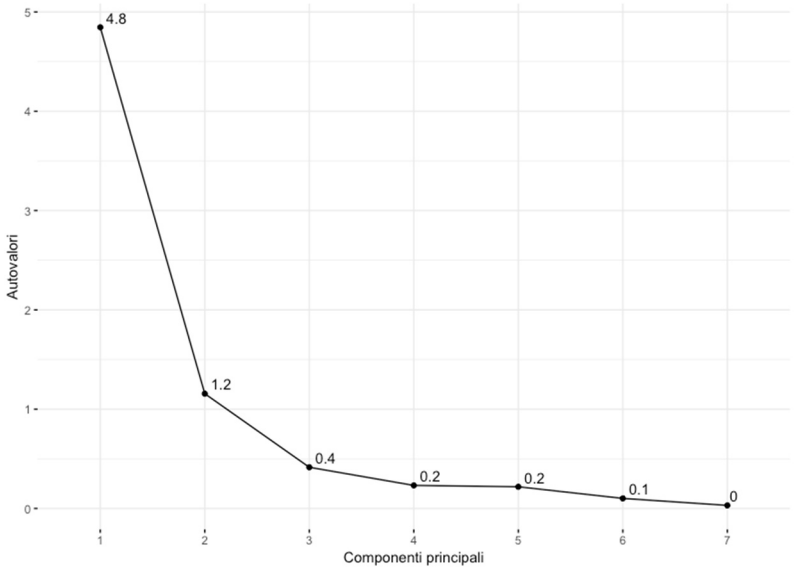
Fig. 1 – Percentuale di varianza spiegata da ciascuna componente



Questo metodo può risultare fuorviante nel caso in cui solo poche variabili influiscono sulla determinazione delle componenti, portando a componenti poco rappresentative del dataset originario oppure troppo specifiche del campione considerato.

Lo **scree-graph** è un grafico che mette in relazione il numero di componenti principali (in ascissa) e gli autovalori corrispondenti (in ordinata). La Figura 2 riporta un esempio del grafico in questione con variabili standardizzate.

Fig. 2 – Esempio di scree-graph



Per scegliere il numero ottimale di componenti principali si utilizza la c.d. *regola del gomito*. Il numero di componenti k da scegliere è quello in corrispondenza del quale l'andamento degli autovalori risulta essere fortemente decrescente a sinistra di k e circa costante a destra. Se a ridosso di k c'è una brusca variazione dell'andamento della spezzata, allora k è una valida soglia e tutte le componenti principali dopo questo valore sono considerate trascurabili. Nell'esempio in figura 2, prenderemo in considerazione le prime due componenti. Tuttavia, può accadere che lo scree-graph presenti un andamento decrescente, ma in modo non così evidente. In questi casi questo metodo risulta essere eccessivamente soggettivo e non attendibile.

La **regola di Kaiser** si applica sia alle variabili non standardizzate che a quelle standardizzate. Nel primo caso, si sceglierà un numero di componenti k i cui autovalori sono maggiori della media, calcolata come: $\frac{1}{p}\sum_{m=1}^{p} l_m$. Nel caso di variabili standardizzate, la media è pari a 1, dunque verranno considerate solamente le componenti che hanno autovalori maggiori di 1, perché

spiegano una quota di varianza maggiore rispetto a quella di una singola variabile. Nell'esempio riportato in Figura 2, si prenderanno soltanto le prime due componenti, perché presentano un autovalore maggiore di 1.

L'ultimo metodo che viene presentato è di tipo **inferenziale** e si fonda sull'ipotesi di distribuzione multinormale della variabile casuale X. Questo metodo consiste nel testare che tutti gli autovalori della popolazione dopo k siano piccoli e uguali tra loro, e quindi trascurabili. Formalizzando il concetto seguendo, avremo la seguente ipotesi nulla:

$$H_{0k}: \lambda_{p-k+1} = \lambda_{p-k+2} = \cdots = \lambda_p \tag{20}$$

Se l'ipotesi nulla è vera, allora sarà possibile effettivamente ridurre la complessità di partenza e tenere solamente le prime k componenti che catturano la struttura sottostante ai dati.

Per testare H_{0k}, è necessario calcolare la statistica test a partire dalla media $\bar{l}$ degli ultimi k autovalori della matrice di varianze-covarianze campionaria: $\bar{l} = \frac{1}{k}\sum_{m=p-k+1}^{p} l_m$.

La statistica test v sarà:

$$v = \left(n - \frac{2p+11}{6}\right)\left(k \ln \bar{l} - \sum_{m=p-k+1}^{p} \ln l_m\right) \tag{21}$$

Sotto l'ipotesi nulla, v assume una distribuzione Chi-Quadrato con w $= \frac{1}{2}(k-1)\,(k+2)$ gradi di libertà; verrà rifiutata se sarà maggiore del valore critico associato ad un Chi-Quadrato con w gradi di libertà, fissato ad un dato livello di significatività.

Per capire quale sia il numero ottimale di componenti da scegliere, si procede testando le componenti a partire dalle ultime, fino ad arrivare al valore k per cui si rifiuta l'ipotesi nulla di uguaglianza tra gli autovalori. Ad esempio, si comincia col testare $H_{02}: \lambda_{p-1} = \lambda_p$, se non la si rifiuta, allo step successivo si testerà $H_{03}: \lambda_{p-2} = \lambda_{p-1} = \lambda_p$; se non si rifiuta neanche questa, allora si ingloberà l'altro autovalore λ_{p-3}nell'ipotesi H_{04}, e così via. Ci si fermerà solamente quando si avrà quel k che implica il rifiuto di H_{0k}. Da ciò deriva che: l_k sarà il primo autovalore di $\boldsymbol{S}$ statisticamente significativo, e quindi L_k è la prima componente da considerare sul totale delle p componenti.

3.4. Interpretazione delle componenti principali

L'interpretazione delle componenti principali dipende molto dal grado di conoscenza che si possiede del fenomeno oggetto di studio. Esistono tuttavia dei metodi oggettivi che sono di supporto all'individuazione del contesto semantico in cui si inserisce ciascuna componente. I metodi principali sono 3:

- interpretazione in base ai pesi delle variabili;
- correlazione tra variabile e componente;
- biplot.

Una volta che il vettore di variabili casuali $\boldsymbol{x}$ si è realizzato in un campione di n unità statistiche, il vettore $\boldsymbol{L_v}$ rappresenta il punteggio della v-esima componente principale lungo tutte le n unità:

$$\boldsymbol{L_v} = \begin{pmatrix} L_{1v} \\ \vdots \\ L_{nv} \end{pmatrix} = a_{v1}\boldsymbol{x}_{.1} + a_{v2}\boldsymbol{x}_{.2} + \cdots + a_{vm}\boldsymbol{x}_{.m} + \cdots + a_{vp}\boldsymbol{x}_{.p}, \qquad (22)$$

dove $\boldsymbol{x_{.m}} = (x_{1m}, \cdots, x_{im}, \cdots, x_{nm})'$.

I coefficienti associati a ciascun vettore colonna di variabili osservate $(\boldsymbol{x_{.1}}, \quad \boldsymbol{x_{.2}}, \quad \cdots, \boldsymbol{x_{.p}})$, rappresentano i pesi assunti da ciascuna variabile nella determinazione delle componenti. Ad esempio, a_{vm} è il peso che il vettore delle realizzazioni della variabile $\boldsymbol{x_{.m}}$ ha nella definizione della v-esima componente.

Per far sì che **i pesi** aiutino nell'interpretazione delle componenti principali, è bene guardare al loro segno e ammontare. Il segno indica il verso della relazione che esiste tra la m-esima variabile e la v-esima componente: se questo è positivo, la relazione è diretta, altrimenti sarà inversa. In aggiunta, valutare l'ammontare dà un'idea generica sull'entità della relazione: maggiore, in valore assoluto, sarà il coefficiente, maggiore è l'influenza della m-esima variabile sulla v-esima componente. Analizzando soltanto i pesi, quindi, si è già in grado di capire quali variabili di partenza caratterizzano in misura maggiore, e in quale direzione, la componente di interesse. Questo metodo presenta dei limiti, poiché il valore dei pesi è espresso in termini assoluti, pertanto, leggendo solo un valore non si riesce a dedurre il grado di importanza di una specifica variabile, finché non lo si confronta con altri pesi.

L'analisi della **correlazione tra variabile e componente** risolve il limite del metodo precedente perché si basa sui coefficienti di correlazione tra variabili e componenti. Nel caso di variabili non standardizzate, la correlazione tra la v-esima componente e la m-esima variabile sarà pari a:

$$r_{L_v,x_m} = \frac{cov(L_v,x_m)}{\sqrt{var(L_v)}\times\sqrt{var(x_m)}} = \frac{l_v a_{vm}}{\sqrt{l_v}\times s_m} = \frac{\sqrt{l_v}}{s_m} a_{vm} \quad (23)$$

dove $s_m = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{im} - \overline{x_m})}$.

Nel caso di variabili standardizzate, la correlazione sarà

$$r_{L_v,z_m} = \sqrt{l_v} a_{vm} \quad (24)$$

dato che, sulla diagonale principale della matrice di correlazione, tutte le varianze, e quindi tutte le deviazioni standard, sono pari a 1.

Siccome l'indice di correlazione varia fra -1 e 1, non è più necessario leggere un valore di concerto con altri per trarre conclusioni sul legame tra variabili e componenti: infatti, se l'indice di correlazione $r_{L_v,z_m/x_m}$ è prossimo in valore assoluto ad 1, si potrà dire con sicurezza (ossia a prescindere dal comportamento delle altre variabili) che la variabile m-esima influenza molto la v-esima componente principale e, il segno indicherà la direzione di questa relazione.

Il **biplot** è un metodo grafico di supporto all'interpretazione simultanea di componenti, variabili e unità. Deriva dalla scomposizione in valori singolari della matrice dei dati standardizzati, procedura che non verrà trattata in questo paragrafo. Gli elementi utili alla costruzione del biplot sono:

- i punteggi standardizzati delle unità per le componenti principali da inserire nel biplot. Nella formula (19) è stato illustrato il punteggio in termini assoluti della componente L_v, per standardizzarlo è necessario dividerlo per la radice quadrata del rispettivo autovalore l_v, che rappresenta la soluzione del problema di massimizzazione:

$$\acute{\boldsymbol{L}}_v = \frac{\boldsymbol{L_v}}{\sqrt{l_v}} = \left(\frac{L_{1v}}{\sqrt{l_v}}, \frac{L_{2v}}{\sqrt{l_v}}, \cdots, \frac{L_{nv}}{\sqrt{l_v}}\right)' \quad (25)$$

  Questi punteggi sono definiti *punti-unità* e sono rappresentati nel grafico tramite dei puntini.
- I coefficienti di correlazione tra componenti principali e variabili, detti *punti-variabile*. Le correlazioni sono rappresentate tramite una freccia uscente dall'origine degli assi. La direzione della freccia indicherà il segno della relazione con le componenti.
- La porzione di varianza di ciascuna variabile spiegata dalle componenti principali. La lunghezza delle frecce associate ad ogni variabile

dipende proprio da questo parametro: più lunga sarà la freccia, maggiore sarà la quota di varianza spiegata dalle componenti inserite nel biplot.

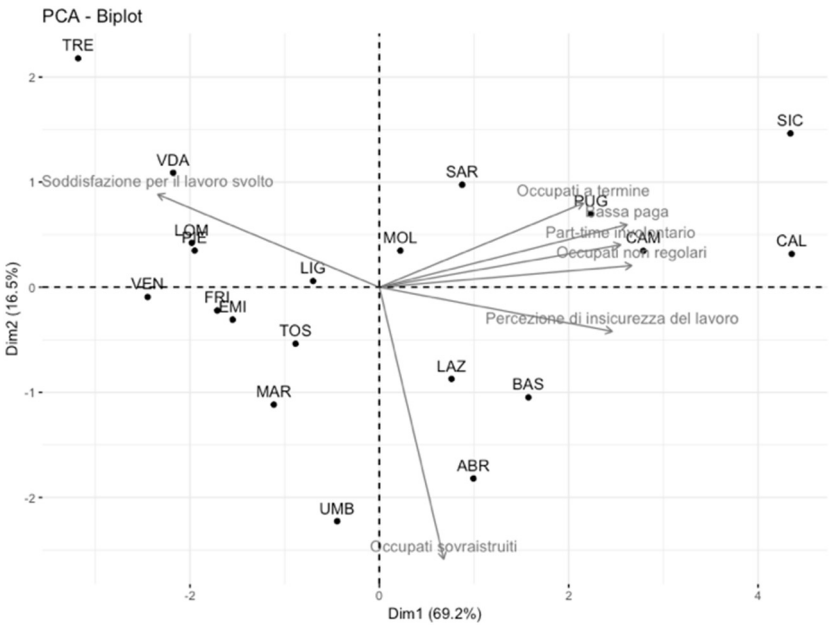
Per semplicità si considera, solitamente, una rappresentazione bidimensionale comprendente le sole prime due componenti: in tal caso, le coordinate dei *punti-unità* saranno date dai punteggi delle unità per le prime due componenti, mentre le coordinate dei *punti-variabile* rappresentano le correlazioni tra una specifica variabile e le due componenti. Di conseguenza, la quota di varianza della variabile catturata dall'ACP sarà calcolata solamente sulle prime due componenti. I confronti che possono essere fatti tramite l'analisi del biplot permettono di interpretare i *punti-unità* e i *punti-variabile* tra loro ma anche rispetto alle componenti principali.

Si procede dapprima con l'interpretazione dei *punti-unità*. Se più unità risultano vicine tra loro nel biplot, queste avranno una composizione simile sia in termini di componenti principali, che in termini di variabili di partenza. Nel caso dei *punti-variabile*, si può analizzare sia la correlazione tra variabili e componenti principali, ma anche la correlazione delle variabili tra loro. Iniziando col primo concetto, si può affermare che, se la freccia associata ad una variabile è parallela all'asse indicante una componente, la correlazione tra questi due elementi è forte. Se, al contrario, la freccia è perpendicolare all'asse di una componente, la correlazione sarà prossima a zero. Il segno della correlazione sarà dato dalla direzione della freccia.

Passando al secondo concetto, due variabili sono tanto più positivamente correlate quanto più le loro frecce formano un angolo piccolo; al contrario, se le frecce formano un angolo prossimo a quello piatto, le due variabili sono correlate negativamente. Se, invece, le frecce formano un angolo retto, le variabili possono ritenersi in-correlate. Infine, è interessante guardare alla relazione simultanea tra *punti-unità* e i *punti-variabile.* Se un'unità si posiziona oltre la freccia associata ad una variabile, si può affermare che l'unità presenta valori della variabile in questione superiori alla media. Se, al contrario, un'unità è prossima all'origine degli assi, assumerà valori delle variabili originarie rispettivamente prossimi alla media.

La Figura 3 riporta un esempio di biplot, che descrive la relazione tra le 20 regioni italiane (*punti-unità*) e 7 variabili (*punti-variabile*) che mirano a sintetizzare la qualità del lavoro.

Fig. 3 – Biplot



Per quanto concerne l'interpretazione dei *punti-unità*, è possibile individuare almeno tre gruppi che racchiudono le regioni italiane: il primo è quello che comprende la maggior parte delle regioni del Sud (Sardegna, Puglia, Campania, Sicilia e Calabria); il secondo ingloba le regioni del Nord e del Centro-nord (Valle d'Aosta, Lombardia, Piemonte, Veneto, Friuli–Venezia Giulia, Emilia Romagna, Liguria e Toscana); infine l'ultimo gruppo, meno numeroso, è formato dalle regioni Lazio, Basilicata ed Abruzzo. Le regioni appartenenti a ciascuna di queste tre macro-aree sono caratterizzate dalla stessa composizione in termini di componenti principali e di variabili originarie. Riguardo i *punti-variabile*, la prima componente è fortemente e positivamente correlata con le variabili che rappresentano una condizione contrattuale del lavoratore particolarmente precaria (occupati a termine, bassa paga, part-time involontario, occupati non regolari, percezione di insicurezza del lavoro); le suddette variabili sono anche positivamente correlate fra loro perché l'angolo che formano a due a due è molto piccolo. Al contrario, è possibile notare come la percezione di insicurezza del lavoro e la soddisfazione per il lavoro svolto siano negativamente correlate perché formano pressoché un angolo piatto. La seconda componente principale, invece, è correlata negativamente con la variabile riferita agli occupati sovraistruiti. Infine,

quanto alla relazione tra *punti-unità* e *punti-variabile*, si può affermare che la Sicilia e la Calabria sono caratterizzate da una presenza di lavoro precario maggiore in media rispetto a tutte le altre regioni, e che, invece, i valori delle variabili originarie assunti dalle regioni Liguria e Molise sono pressoché nella media.

Riferimenti bibliografici

Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). New York: Springer Serires in Statistics.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate Analysis.* Padstow, Cornwall: Academic Press, pp. 213-254.

Rencher, A. C., & Christensen, W. F. (2012). *Methods of Multivariate Analysis* (Terza edizione ed.). Hoboken, New Jersey: Wiley, pp. 405-433 & 565-580.

Vitali, O. (1993). *Statistica per le scienze applicate.* (Vol. II). Bari: Cacucci Editore, pp. 203-209 & 501-535.

Zani, S., & Cerioli, A. (2007). *Analisi dei dati e data mining per le decisioni aziendali.* Milano: Giuffrè, pp. 215-264 & 267-300.

Il presente manuale è una raccolta collettanea di contributi che presentano una nuova chiave di lettura dello sviluppo locale, declinando in prospettiva scientifico-divulgativa "nuove" teorie e tematiche classiche che hanno costituito da sempre lo schema di ciò che normalmente viene identificato sviluppo locale. Vi è, quindi, una multiformità di tematiche che potrebbero, a primo impatto, apparire eterogenee e distanti dagli obiettivi di sviluppo: come il tema della salute e della sanità. Il Covid-19 ha confermato, però, che immaginare oggi uno sviluppo locale a prescindere dalle tematiche sanitarie di un territorio rappresenta sicuramente un'assurdità. Ma nelle vecchie teorie dello sviluppo locale questa tematica non è riscontrabile. Questa declinazione dello sviluppo locale è inserita all'interno della nuova vision della cosiddetta "Economia di Francesco" che è il chiaro riferimento all'evoluzione dell'economia civile di A. Genovesi arricchita dalle suggestioni francescane e benedettine compendiate nella "Laudato Sii" di San Francesco e nella "Regola" di San Benedetto. L'occasione di questo manuale è data dalla presentazione degli atti del progetto SKIN (Short supply chain Knowledge and Innovation Network, finanziato nell'ambito del programma Horizon 2020) che ha avuto come Lead Partner l'Università di Foggia (Dipartimento di Economia) e si è appena concluso dando alla luce un'importante rete tematica europea sulla filiera corta.

Francesco Contò è professore ordinario presso l'Università di Foggia, Dipartimento di Economia. È stato Direttore del Dipartimanto di Economia e coordinatore di corsi di dottorato. Dal 1977 gli sono state affidate importanti posizioni scientifico-accademiche e professionali. È autore di oltre 270 pubblicazioni nazionali e internazionali. Ha ricoperto incarichi di ricerca e insegnamento presso università e centri di ricerca qualificati e attualmente è direttore scientifico di alcuni importanti laboratori di ricerca regionali. È coordinatore di numerosi progetti di ricerca europei, nazionali e regionali.

Mariantonietta Fiore è professore associato presso l'Università di Foggia, Dipartimento di Economia. È membro della Scuola di dottorato internazionale della Warsaw University of Life Sciences e del Board of Directors dell'International Food and Agribusiness Management Association. È Fellow dell'EuroMed Academy of Business. È stata vicecoordinatrice scientifica del progetto SKIN ed esperto junior del Ministero dell'Ambiente: attualmente è responsabile o membro di progetti scientifici, gruppi di ricerca e comitati editoriali internazionali e nazional. Ha ricevuto oltre dieci premi scientifici.

FrancoAngeli
La passione per le conoscenze